ELSEVIER

# Atom-based 3D-chiral quadratic indices. Part 2: Prediction of the corticosteroid-binding globulin binding affinity of the 31 benchmark steroids data set

Juan A. Castillo-Garit,[a,b,*] Yovani Marrero-Ponce[b,c] and Francisco Torrens[c]

[a]Applied Chemistry Research Center, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba
[b]Department of Pharmacy, Faculty of Chemistry-Pharmacy and Department of Drug Design,
Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba
[c]Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E-46100 Burjassot, València, Spain

**Abstract**—A quantitative structure–activity relationship (QSAR) study to predict the relative affinities of the steroid 'benchmark' data set to the corticosteroid-binding globulin (CBG) is described. It is shown that the 3D-chiral quadratic indices closely correlate with the measured CBG affinity values for the 31 steroids. The calculated descriptors were correlated with biological data through multiple linear regressions. Two statistically significant models were obtained when non-stochastic ($R = 0.924$ and $s = 0.46$) as well as stochastic ($R = 0.929$ and $s = 0.46$) 3D-chiral quadratic indices were used. A leave-one-out (LOO) approach to model validation is used here; the best results obtained in the cross-validation procedure with non-stochastic ($q^2 = 0.781$) and stochastic ($q^2 = 0.735$) 3D-chiral quadratic indices are better or similar to most of the 3D-QSAR approaches reported so far. These results support the idea that the 3D-chiral quadratic indices may be helpful in prediction of the corticosteroid-binding affinity for new compounds.
© 2005 Elsevier Ltd. All rights reserved.

## 1. Introduction

Corticosteroids are essential for the regulation of human physiology. Glucocorticoids are involved in the regulation of protein, carbohydrate, and lipid metabolism. Mineralocorticoids control salt and water metabolism by maintaining proper electrolyte balance. Prednisone and prednisolone are potent antirheumatic and antiallergenic agents. Cortisone is effective in the treatment of rheumatoid arthritis.[1] The corticosteroid drug budesonide is the preferred treatment for Crohn's disease.[2] Baud et al. have found that the corticosteroid drug betamethasone has fewer side effects when used for preventing many pregnant women from delivery complications of premature birth.[3] Corticosteroid-receptor activation is crucial in determining steroid-mediated effects.

In recent years, modeling of the relative affinities of steroids to the corticosteroid-binding globulin (CBG) has become a very important research topic. The steroid structures examined in the original Comparative Molecular Field Analysis (CoMFA) paper[4] have served as a benchmark for numerous subsequent quantitative structure activity relationship (QSAR) studies.[5–20] In this significant research effort, there is need to find models of the relationship between the structure of steroids and the corticosteroid-binding globulin. Such models can assist researchers in understanding the structural basis of binding as well as providing a basis for development of new compounds.

Our research group has recently developed simple non-stochastic and stochastic atom- and bond-based molecular descriptors based on algebraic theory. They have been defined by analogy with the quadratic, linear, and bilinear mathematical maps.[21–26] Applications included the prediction of several physical, physicochemical, chemical, and pharmacokinetic properties of organic compounds.[21–26] In addition, these indices have been extended by considering three-dimensional features of small/medium-sized molecules, based on

the trigonometric 3D-chirality correction factor approach.[27] Promising results have been obtained modeling steroids binding to CBG using 3D-chiral linear indices.[28] The present report is written with two objectives in mind; first, to investigate the potential use of the 3D-chiral quadratic indices as a modeling tool for obtaining significant correlation with the corticosteroid-binding affinity; second, to compare our results with those of other methods previously used for examining this system.

## 2. Theoretical scaffold

In earlier publications, we outline outstanding features concerned with the theory of 2D atom-based TOMO-COMD-CARDD molecular descriptors (MDs). This method codifies molecular structures by means of mathematical quadratic, linear, and bilinear transformations.[21–26] In order to calculate these algebraic maps for a molecule, the atom-based molecular vector, $\bar{x}$ (vector representation), and $k$th 'non-stochastic and stochastic graph-theoretic electronic-density matrices,' $\mathbf{M}^k$ and $\mathbf{S}^k$ (matrix representations), correspondingly, are constructed.[21,22,28–47] Such atom-adjacency relationships and chemical-information codification are applied to generate a series of atom-based TOMOCOMD-CARDD MDs, namely atomic, group, and atom-type as well as total quadratic indices, to be used in drug design and chemoinformatics studies.

Therefore, the structure of this section will be as follows: (1) a background in atom-based molecular vector, as well as non-stochastic and stochastic graph-theoretic electronic-density matrices will be described in the next two subsections, and (2) an outline of the mathematical definition of quadratic maps and a definition of our procedures will be developed in the third subsection.

### 2.1. Chemical information and atom-based molecular vector

The atom-based molecular vector ($\bar{x}$), used to represent small-to-medium-sized organic chemicals, has been explained elsewhere in some detail.[21,22,28–44] The components ($x$) of $\bar{x}$ are numeric values, which represent a certain standard atomic property (atomic label). Therefore, these weights correspond to different atom properties for organic molecules. Thus, a molecule having $5, 10, 15, \ldots, n$ atomic nuclei can be represented by means of vectors with $5, 10, 15, \ldots, n$ components, belonging to the spaces $\Re^5, \Re^{10}, \Re^{15}, \ldots, \Re^n$, respectively, where $n$ is the dimension of the real set ($\Re^n$). Therefore, $\bar{x}$ is the $n$-dimensional property vector of the atoms (atomic nuclei) in a molecule.

This approach allows us to encode organic molecules such as 3-mercapto-pyridine-4-carbaldehyde through the molecular vector $\bar{x} = [x_{N1}, x_{C2}, x_{C3}, x_{C4}, x_{C5}, x_{C6}, x_{C7}, x_{O8}, x_{S9}]$. This vector belongs to the product space $\Re^9$. However, diverse kinds of atomic weights ($x$) can be used for codifying information related to each atomic nucleus in the molecule. These atomic labels are

**Table 1.** Values of the atomic weights used for TOMOCOMD-CARDD MDs[48–50]

| ID | Atomic mass | VdW[a] volume (Å$^3$) | Polarizability (Å$^3$) | Pauling electronegativity |
|----|-------------|----------------------|------------------------|---------------------------|
| H | 1.01 | 6.709 | 0.667 | 2.2 |
| B | 10.81 | 17.875 | 3.030 | 2.04 |
| C | 12.01 | 22.449 | 1.760 | 2.55 |
| N | 14.01 | 15.599 | 1.100 | 3.04 |
| O | 16.00 | 11.494 | 0.802 | 3.44 |
| F | 19.00 | 9.203 | 0.557 | 3.98 |
| Al | 26.98 | 36.511 | 6.800 | 1.61 |
| Si | 28.09 | 31.976 | 5.380 | 1.9 |
| P | 30.97 | 26.522 | 3.630 | 2.19 |
| S | 32.07 | 24.429 | 2.900 | 2.58 |
| Cl | 35.45 | 23.228 | 2.180 | 3.16 |
| Fe | 55.85 | 41.052 | 8.400 | 1.83 |
| Co | 58.93 | 35.041 | 7.500 | 1.88 |
| Ni | 58.69 | 17.157 | 6.800 | 1.91 |
| Cu | 63.55 | 11.494 | 6.100 | 1.9 |
| Zn | 65.39 | 38.351 | 7.100 | 1.65 |
| Br | 79.90 | 31.059 | 3.050 | 2.96 |
| Sn | 118.71 | 45.830 | 7.700 | 1.96 |
| I | 126.90 | 38.792 | 5.350 | 2.66 |

[a] VdW: van der Waals.

chemically meaningful numbers such as the atomic masses, the atomic polarizabilities, and so on. In the present report, we characterized each atomic nucleus with the following parameters (weighting scheme): the atomic masses ($M$), the van der Waals volumes ($V$), the atomic polarizabilities ($P$), atomic electronegativities in Mulliken scale ($K$), and atomic electronegativities in Pauling scale ($G$). The values of these atomic labels are shown in Table 1.[48–50]

### 2.2. Background in non-stochastic and stochastic graph-theoretic electronic-density matrices

In molecular topology, molecular structure is expressed, generally, by the hydrogen-suppressed graph. Therefore, a molecule is represented by a graph. Informally, a graph $G$ is a collection of vertices (points) and edges (lines or bonds) connecting these vertices.[51–53] In more formal terms, a simple graph $G$ is defined as an ordered pair $[V(G), E(G)]$, which consists of a nonempty set of vertices $V(G)$ and a set $E(G)$ of unordered pairs of elements of $V(G)$, called edges.[51–53] In this particular case, we are not dealing with a simple graph but with a so-called pseudograph ($G$). Informally, a pseudograph is a graph with multiple edges or loops between the same vertices or the same vertex. Formally, a pseudograph is a set $V$ of vertices along a set $E$ of edges and a function $f$ from $E$ to $\{\{u,v\} \mid u,v \text{ in } V\}$ (The function $f$ shows which pair of vertices is connected by which edge). An edge is a loop if $f(e) = \{u\}$ for some vertex $u$ in V.[21,22,54]

In earlier reports we have introduced new molecular matrices that describe changes along time in the electronic distribution throughout the molecular backbone. The $n \times n$ $k$th non-stochastic graph-theoretic electronic-density matrix of the molecular pseudograph ($G$), $\mathbf{M}^k$, is a symmetric and square matrix, where $n$ is the number of atoms (atomic nuclei) in the molecule.[21,22,28–44] The

coefficients $^{k}m_{ij}$ are the elements of the $k$th power of $\mathbf{M}(G)$ and are defined as follows:

$$m_{ij} = P_{ij} \text{ if } i \neq j \text{ and } \exists e_k \in E(G) \qquad (1)$$
$$= L_{ii} \text{ if } i = j$$
$$= 0 \text{ otherwise },$$

where $E(G)$ represents the set of edges of $G$. $P_{ij}$ is the number of edges (bonds) between vertices (atomic nuclei) $v_i$ and $v_j$, and $L_{ii}$ is the number of loops in $v_i$.

The elements $m_{ij} = P_{ij}$ of such a matrix represent the number of chemical bonds between an atomic nucleus $i$ and other $j$. The matrix $\mathbf{M}^k$ provides the number of walks of length $k$ that link every pair of vertices $v_i$ and $v_j$. For this reason, each edge in $\mathbf{M}^1$ represents 2 electrons belonging to the covalent bond between atomic nuclei $i$ and $j$; for example, the inputs of $\mathbf{M}^1$ are equal to 1, 2 or 3 when single, double or triple bonds, correspondingly, appear between vertices $v_i$ and $v_j$. On the other hand, molecules containing aromatic rings with more than one canonical structure are represented by a pseudograph; for instance, substituted aromatic compounds such as pyridine, naphthalene, quinoline, and so on, where the presence of pi ($\pi$) electrons is accounted for by means of loops in each atomic nucleus of the aromatic ring. However, aromatic rings having only one canonical structure, such as furan, thiophene, and pyrrol, are represented by a multigraph. In order to illustrate the calculation of these matrices, let us consider the same molecule selected in the previous section. Table 2 depicts the molecular structure of this compound and its labeled molecular pseudograph. The zero ($k = 0$), first ($k = 1$), second ($k = 2$), and third ($k = 3$) powers of the non-stochastic graph-theoretic electronic-density matrix are also given in this table.

As it can be seen, $\mathbf{M}^k$ are graph-theoretic electronic-structure models, like an 'extended Hückel theory (EHT) model.' The $\mathbf{M}^1$ matrix considers all valence-bond electrons ($\sigma$- and $\pi$-networks) in one step and its power ($k = 0, 1, 2, 3...$) can be considered as interacting-electron chemical-network models in the $k$ step. The complete model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical-chemical bonding ideas.[55]

The present approach is based on a simple model for the intramolecular movement of all outer-shell electrons. Let us consider a hypothetical situation in which a set of atoms is free in space at an arbitrary initial time ($t_0$). At this time, the electrons are distributed around the atomic nuclei. 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 (step-by-step) to other atoms at different discrete time periods $t_k(k = 0, 1, 2, 3,...)$ throughout the chemical-bonding network. A similar approach was introduced by González-Díaz et al. for a simple graph.[56,57]

On the other hand, the $k$th stochastic graph-theoretic electronic-density matrix of G, $\mathbf{S}^k$, can be directly obtained from $\mathbf{M}^k$. Here, $\mathbf{S}^k = [^{k}\mathbf{s}_{ij}]$ is a square matrix of order $n$ ($n$ = number of atomic nuclei) and the elements $^{k}\mathbf{s}_{ij}$ are defined as follows:

$$^{k}s_{ij} = \frac{^{k}m_{ij}}{^{k}SUM_i} = \frac{^{k}m_{ij}}{^{k}\delta_i}, \qquad (2)$$

where $^{k}m_{ij}$ are the elements of the $k$th power of M, and the SUM of the $i$th row of $\mathbf{M}^k$ are named the $k$-order vertex degree of atom $i$,$^{k}\delta_i$. It should be remarked that the matrix $\mathbf{S}^k$ in Eq. 2 has the property that the sum of the elements in each row is 1. An $n \times n$ matrix with nonnegative entries having this property is called a 'stochastic matrix;'[30] it was introduced in QSAR by Carbó-Dorca[58] and, more recently, it was used by González-Díaz et al. for a simple graph[59–62] and by Marrero-Ponce et al.[29,31] The $k$th $s_{ij}$ elements are the transition probabilities of the electrons moving from atom $i$ to $j$ at discrete time periods $t_k$. It should also be pointed out that $k$th element $s_{ij}$ takes into consideration the molecular topology in $k$ step throughout the chemical-bonding ($\sigma$- and $\pi$-) network. In this sense, the $^{2}s_{ij}$ values can distinguish between hybrid states of atoms in bonds. For instance, the self-return probability of second order ($^{2}s_{ii}$) [i.e., the probability with which an electron returns to the original atom at $t_2$] varies regularly according to the different hybrid states of atom $i$ in the molecule, for example, an electron will have a higher probability of returning to the sp C atom than to the sp$^2$ (or sp$^3$) C atom in $t_2$ [$p(C_{sp}) > p(C_{sp^2}) > p(C_{sp\ arom^2}) > p(C_{sp^3})$] (see Table 2 for more details). This is a logical result if the electronegativity scale of these hybrid states is taken into account. The same analysis was reported for a simple graph for stochastic matrix.[56]

### 2.3. Calculation of quadratic indices for atoms, groups, atom types, and the whole molecule
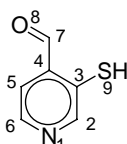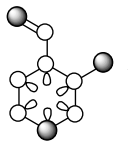
If a molecule consists of $n$ atoms (vector of $\mathfrak{R}^n$), then the $k$th total (whole) quadratic indices are calculated as quadratic forms on $\mathfrak{R}^n$ in a canonical basis set. Specifically, the $k$th non-stochastic and stochastic atom-based quadratic indices for a molecule, $q_k(\bar{x})$ and $^{s}q_k(\bar{x})$, respectively, are computed from these $k$th non-stochastic and stochastic graph-theoretic electronic-density matrices, $\mathbf{M}^k$ and $\mathbf{S}^k$, as shown in Eqs. 3 and 4:[21,22,28–47]

$$q_k(\bar{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}m_{ij}x_ix_j = [X]^t\mathbf{M}^k[X] \qquad (3)$$

$$^{s}q_k(\bar{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}s_{ij}x_ix_j = [X]^t\mathbf{S}^k[X] \qquad (4)$$

where $n$ is the number of atoms of the molecule and $x_1,...,x_n$ are the coordinates or components of the 'molecular vector' ($\bar{x}$) in a canonical ('natural') basis set of $\mathfrak{R}^n$. In this basis set, the coordinates of any vector $\bar{x}$, namely $x_1,...,x_n$, coincide with the components of this vector.[63,64] Therefore, those coordinates can be considered as weights (atomic labels) of the vertices in the molecular pseudograph. The coefficients $^{k}m_{ij}$ and $^{k}s_{ij}$

**Table 2.** (A) Chemical structure of 3-mercapto-pyridine-4-carbaldehyde and its labeled molecular pseudograph, G. (B) and (C) The zero ($k = 0$), first ($k = 1$), second ($k = 2$), and third ($k = 3$) powers of the non-stochastic and stochastic graph-theoretic electronic-density matrices of G, respectively

(A) Molecular structure

Molecular pseudograph (H atom-suppressed pseudograph)[a]



Vertices { ● Heteroatoms / ○ C atoms }

(B) $k$th non-stochastic graph-theoretical electronic-density matrices, $\mathbf{M}^k$ ($k = 1–3$)

Zero order ($k = 0$)

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

First order ($k = 1$)

$$
\begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0\\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1\\
0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Second order ($k = 2$)

$$
\begin{bmatrix}
3 & 2 & 1 & 0 & 1 & 2 & 0 & 0 & 0\\
2 & 3 & 2 & 1 & 0 & 1 & 0 & 0 & 1\\
1 & 2 & 4 & 2 & 1 & 0 & 1 & 0 & 1\\
0 & 1 & 2 & 4 & 2 & 1 & 1 & 2 & 1\\
1 & 0 & 1 & 2 & 3 & 2 & 1 & 0 & 0\\
2 & 1 & 0 & 1 & 2 & 3 & 0 & 0 & 0\\
0 & 0 & 1 & 1 & 1 & 0 & 5 & 0 & 0\\
0 & 0 & 0 & 2 & 0 & 0 & 0 & 4 & 0\\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

Third order ($k = 3$)

$$
\begin{bmatrix}
7 & 6 & 3 & 2 & 3 & 6 & 0 & 0 & 1\\
6 & 7 & 7 & 3 & 2 & 3 & 1 & 0 & 2\\
3 & 7 & 9 & 8 & 3 & 2 & 2 & 2 & 4\\
2 & 3 & 8 & 9 & 7 & 3 & 8 & 2 & 2\\
3 & 2 & 3 & 7 & 7 & 6 & 2 & 2 & 1\\
6 & 3 & 2 & 3 & 6 & 7 & 1 & 0 & 0\\
0 & 1 & 2 & 8 & 2 & 1 & 1 & 10 & 1\\
0 & 0 & 2 & 2 & 2 & 0 & 10 & 0 & 0\\
1 & 2 & 4 & 2 & 1 & 0 & 1 & 0 & 1
\end{bmatrix}
$$

(C) stochastic graph-theoretic electronic-density matrices, $\mathbf{S}^k$ ($k = 0–3$)[b,c]

First order ($k = 1$)

$$
\begin{bmatrix}
0.3333 & 0.3333 & 0 & 0 & 0 & 0.3333 & 0 & 0 & 0\\
0.3333 & 0.3333 & 0.3333 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0.25 & 0.25 & 0.25 & 0 & 0 & 0 & 0 & 0.25\\
0 & 0 & 0.25 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0\\
0 & 0 & 0 & 0.3333 & 0.3333 & 0.3333 & 0 & 0 & 0\\
0.3333 & 0 & 0 & 0 & 0.3333 & 0.3333 & 0 & 0 & 0\\
0 & 0 & 0 & 0.3333 & 0 & 0 & 0 & 0.6666 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Second order ($k = 2$)

$$
\begin{bmatrix}
0.3333 & 0.2222 & 0.1111 & 0 & 0.1111 & 0.2222 & 0 & 0 & 0\\
0.2 & 0.3 & 0.2 & 0.1 & 0 & 0.1 & 0 & 0 & 0.1\\
0.0833 & 0.166 & 0.3333 & 0.1666 & 0.0833 & 0 & 0.0833 & 0 & 0.0833\\
0 & 0.0714 & 0.1429 & 0.2857 & 0.1429 & 0.0714 & 0.0714 & 0.1429 & 0.0714\\
0.1 & 0 & 0.1 & 0.2 & 0.3 & 0.2 & 0.1 & 0 & 0\\
0.2222 & 0.1111 & 0 & 0.1111 & 0.2222 & 0.3333 & 0 & 0 & 0\\
0 & 0 & 0.125 & 0.125 & 0.125 & 0 & 0.625 & 0 & 0\\
0 & 0 & 0 & 0.3333 & 0 & 0 & 0 & 0.6666 & 0\\
0 & 0.25 & 0.25 & 0.25 & 0 & 0 & 0 & 0 & 0.25
\end{bmatrix}
$$

Third order ($k = 3$)

$$
\begin{bmatrix}
0.25 & 0.2142 & 0.1071 & 0.0714 & 0.1071 & 0.2143 & 0 & 0 & 0.0357\\
0.1935 & 0.2258 & 0.2258 & 0.0967 & 0.0645 & 0.0967 & 0.0323 & 0 & 0.0645\\
0.075 & 0.175 & 0.225 & 0.2 & 0.075 & 0.05 & 0.05 & 0.05 & 0.1\\
0.0455 & 0.0682 & 0.1818 & 0.2045 & 0.1591 & 0.0682 & 0.1818 & 0.0455 & 0.0455\\
0.0909 & 0.0606 & 0.0909 & 0.2121 & 0.2121 & 0.1818 & 0.0606 & 0.0606 & 0.0303\\
0.2143 & 0.1071 & 0.0714 & 0.1971 & 0.2143 & 0.25 & 0.0357 & 0 & 0\\
0 & 0.0385 & 0.0769 & 0.3076 & 0.0769 & 0.0385 & 0.0385 & 0.3846 & 0.0385\\
0 & 0 & 0.125 & 0.125 & 0.125 & 0 & 0.625 & 0 & 0\\
0.0833 & 0.1666 & 0.333 & 0.1666 & 0.0833 & 0 & 0.0833 & 0 & 0.0833
\end{bmatrix}
$$

[a] Each edge in $\mathbf{M}^1$ represents two electrons belonging to the covalent bond between atoms (vertices) $v_i$ and $v_j$; for example, the inputs of $\mathbf{M}^1$ are equal to 1, 2 or 3 when single, double or triple bonds, correspondingly, appear between vertices $v_i$ and $v_j$. The presence of pi ($\pi$) electrons in aromatic systems such as benzene is accounted for by means of loops in each atom of the aromatic ring. Therefore, the $\mathbf{M}^1$ matrix considers all valence-bond electrons ($\sigma$- and $\pi$-networks) in one step, and their powers ($k = 0, 1, 2, 3...$) can be considered as an interacting-electron chemical-network model in the $k$ step.
[b] The zero power ($k = 0$) of the stochastic graph-theoretical electronic-density matrix, $\mathbf{S}^0$, coincides with the non-stochastic matrix one ($\mathbf{M}^0 = \mathbf{S}^0$).
[c] The values of the elements of $k$th matrices $\mathbf{S}^k$ ($^k s_{ij}$) have been rounded.

are the elements of the $k$th power of the matrices $\mathbf{M}(G)$ and $\mathbf{S}(G)$, correspondingly, in the molecular pseudograph.

The defined Eqs. 3 and 4 for $\boldsymbol{q}_k(\bar{x})$ and $^s\boldsymbol{q}_k(\bar{x})$ may also be written in matrix form, where [X] is a column vector (an $n \times 1$ matrix) of the coordinates of $\bar{x}$ in the canonical basis of $\Re^n$ and $[\mathbf{X}]^t$ (a $1 \times n$ matrix) is the transpose of [X]. Here, $\mathbf{M}^k$ and $\mathbf{S}^k$ denote the matrices of quadratic maps with respect to the natural basis set.

In addition to total quadratic indices, computed for the whole molecule, a local-fragment (atomic and atom type as well as group) formalism can be developed. These descriptors are termed local non-stochastic and stochastic quadratic indices, $\boldsymbol{q}_{kL}(x)$ and $^s\boldsymbol{q}_{kL}(x)$, correspondingly.[21,22,28–47] The definition of these descriptors is as follows:

$$q_{kL}(\bar{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k m_{ijL} x_i x_j = [\mathbf{X}]^t \mathbf{M}_L^k [\mathbf{X}] \qquad (5)$$

$$^s q_{kL}(\bar{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k s_{ijL} x_i x_j = [\mathbf{X}]^t \mathbf{S}_L^k [\mathbf{X}] \qquad (6)$$

where $n$ is the number of atoms (atomic nuclei) in the fragment of interest and ${}^k m_{ijL}$ [${}^k s_{ijL}$] is the $k$th element of the row '$i$' and column '$j$' of the local matrix $\mathbf{M}_L^k$ [$\mathbf{S}_L^k$]. This matrix is extracted from the $\mathbf{M}^k$ [$\mathbf{S}^k$] matrix and contains information referring to the vertices (atomic nuclei) of the specific molecular fragments and also of the molecular environment in $k$ steps. The matrix $\mathbf{M}_L^k$ [$\mathbf{S}_L^k$] with elements ${}^k m_{ijL}$ [${}^k s_{ijL}$] is defined as follows:

$$
\begin{aligned}
{}^k m_{ijL}[{}^k \mathbf{s}_{ijL}] = {}^k m_{ij}[{}^k \mathbf{s}_{ijL}] \ & \text{if both } v_i \text{ and } v_j \text{ are atoms} \\
& \text{contained within the molecular fragment} \\
= 1/2\, {}^k m_{ij}[{}^k \mathbf{s}_{ijL}] \ & \text{if either } v_i \text{ or } v_j \text{ is an atom} \\
& \text{contained within the molecular fragment} \\
= 0 \ & \text{otherwise} \qquad (7)
\end{aligned}
$$

These local analogues can also be expressed in matrix form for each molecular vector $\bar{x} \in \Re^n$. It should be remarked that the scheme above follows the spirit of the Mulliken population analysis of atomic net charges.[65] It should also be pointed out that for every partitioning of a molecule into $Z$ molecular fragments, there will be $Z$ local molecular fragmental matrices. In this case, if a molecule is partitioned into $Z$ molecular fragments, the matrix $\mathbf{M}^k$ [$\mathbf{S}^k$] can be partitioned into $Z$ local matrices $\mathbf{M}_L^k$ [$\mathbf{S}_L^k$], $L = 1, \dots, Z$, and the $k$th power of matrix $\mathbf{M}$ [$\mathbf{S}$] is exactly the sum of the $k$th power of the local $Z$ matrices. Therefore, the total non-stochastic and stochastic quadratic indices are the sum of the non-stochastic and stochastic quadratic indices, respectively, of the $Z$ molecular fragments:

$$\boldsymbol{q}_k(\bar{x}) = \sum_{L=1}^{Z} q_{kL}(\bar{x}) \qquad (8)$$

$$^s\boldsymbol{q}_k(\bar{x}) = \sum_{L=1}^{Z} {}^s q_{kL}(\bar{x}) \qquad (9)$$

Atomic, atom-type, and group quadratic fingerprints are specific cases of local quadratic indices. First, it should be noted that atomic quadratic indices, $\boldsymbol{q}_k(\bar{x}_i)$ and $^s\boldsymbol{q}_k(\bar{x}_i)$, can be computed for each atom $i$ in the molecule and contain electronic and topological structural information from all other atoms within the structure. The values of atom-level quadratic indices for the common scaffold atoms can be directly used as variables in seeking a QSPR/QSAR model, as long as these atoms are numbered in the same way in all molecules in the database. As it can be seen, the $k$th total quadratic indices (both non-stochastic and stochastic) are calculated by summing the atomic quadratic indices of all atoms in the molecule.

Moreover, the atom-type quadratic indices can also be calculated as local MDs. In the same way, as atom-type E-state values[66] and other local descriptors[59,67,68] for all data sets (including those with a common skeletal core as well as those with very diverse structures), these novel local MDs provide much useful information. That is, this approach provides the basis for application to a wider range of problems to which the atomic quadratic indices formalism is directed without the need for superposition. For this reason the present method represents a significant advantage over traditional QSAR methods. The atom-type quadratic descriptors are calculated by adding the $k$th atomic quadratic indices for all atoms of the same type in the molecule. This atom-type index allows a group additive-type scheme in which an index appears for each atom type in the molecule.

In the atom-type quadratic indices formalism, each atom in the molecule is classified into an atom type (fragment), such as –F, –OH, =O, –CH$_3$, and so on.[66] That is to say, each atom in the molecule is categorized according to a valence-state classification scheme including the number of attached H atoms.[66] The atom-type descriptors combine three important aspects of structural information: (1) collective electron and topologic accessibility to the atoms of the same type (for each structural feature: atom or hybrid group such as –Cl, =O, –CH$_2$–, etc.), (2) presence/absence of the atom type (structural features), and (3) count of the atoms in the atom-type sets.

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

## 2.4. 3D-chiral extended non-stochastic and stochastic quadratic indices for atoms, atom types, and group as well as for whole molecule

The total and local, non-stochastic, and stochastic quadratic indices, as defined above, cannot codify any information about 3D molecular structure. In order to solve this problem, we introduced a trigonometric 3D-chirality correction factor in the molecular vector $\bar{x}$.[28,42] In this sense, a chirality molecular vector is obtained ($^*\bar{x}$), where the components of $\bar{x}$ (for instance, Mulliken electronegativity ($x_A$) of the atom A) are substituted by the following term $\{x_A + \sin[(\omega_A + 4\varDelta)\pi/2)]\}$.

The trigonometric 3D-chirality correction factor uses a dummy variable, $\omega_A$, and an integer parameter, $\varDelta$:[28,42]

$\omega_A = 1$ and $\varDelta$ is an odd number when $A$ has $R$(rectus), $E$(entgegen), or $a$(axial)notation according to the Cahn–Ingold–Prelog(IUPAC)rules

$= 0$ and $\varDelta$ is an even number, if $A$ does not have 3D specific enviroment,

$= -1$ and $\varDelta$ is an odd number when $A$ has $S$(sinister), $Z$(zusammen), or e(equatorial)notation according to Cahn–Ingold–Prelog rules. (10)

Thus, this 3D-chirality factor $\sin[(\omega_A + 4\varDelta)\pi/2]$ takes different values in order to codify specific stereochemical information such as chirality, $Z/E$ isomerism, and so on. This factor takes, therefore, values in the following order $1 > 0 > -1$ for atoms that have specific 3D environments. The chemical idea here is not that the attraction of electrons by an atom depends on their chirality, due to the fact that experience shows that chirality does not change the electronegativities of the atoms in the molecule in an isotropic environment in an observable way.[69] This correction, as pointed out by González-Díaz et al.,[70,71] has principally a mathematical meaning and must not be source of any misunderstanding.

A severe limitation of the Golbraikh, Bonchev, and Tropsha (GBT)[72] approach is the existence of different chirality corrections and we had great difficulty in selecting one of these. In this sense, González-Díaz et al.[73] introduced an exponential chirality factor [(exp ($\omega_A\varDelta$)], which eliminated indetermination in the selection of chirality and 3D scales for stochastic topological indices. Unfortunately, this exponential factor does not solve the problem in GBT-like approaches.[72]

Therefore, the present trigonometric 3D-chiral correction factor is invariant with respect to the selection of other chirality scales for all kinds of such chiral topological indices (GBT-like ones). Table 3 depicts the values of the trigonometric 3D-Chirality correction factor for all allowed values of $\omega_A$ and $\varDelta$ (GBT-like chirality scale and other alternative chirality scales). In this table, it is clearly shown that the trigonometric 3D-chirality factor is invariant with respect to the selection of all possible real scales. That is to say, the factor gets over the values 1, 0, and $-1$ for R, non-chiral, and S atoms, correspondingly. As outlined above the demonstration of invariance for this factor with respect to other 3D features such as $a/e$ substitutions and $Z/E$ or $\pi$-isomerism is straightforward to realize by homology. Henceforth, we do not need to answer the question regarding the best value for chirality correction, at least for linear scales.[72–74]

A very interesting point is that the present 3D-chiral descriptor reduces to simple (2D) non-stochastic and stochastic quadratic indices, for molecules without specific 3D characteristics because $\sin[(0 + 4\varDelta)\pi/2] = 0$, being $\varDelta$ zero or any integer (particularly even) number. Therefore, when all the atoms in the molecule are not chiral, the TOMOCOMD-CARDD MDs or any GBT-like chiral topological index do not change upon introduction of this factor. This means that $^*\bar{x} = \bar{x}$ and, thus, $^*q_k(\bar{x}) = q_k(\bar{x})$.

The following descriptors were calculated in this work:

(i) $k$th total 3D-chiral quadratic indices not considering and considering H atoms in the molecular pseudograph ($G$) [$^*q_k(\bar{x})$ and $^*q_k^H(\bar{x})$, respectively].

(ii) $k$th local (atomic group = heteroatoms: S, N, O) 3D-chiral quadratic indices not considering and considering H atoms in the molecular pseudograph ($G$) [$^*q_{kL}(\bar{x}_E)$ and $^*q_{kL}^H(\bar{x}_E)$, correspondingly]. These local descriptors are putative H-bonding acceptors; in addition, charge as well as dipole moment.

(iii) $k$th local (atomic group = H-atoms bonding to heteroatoms: S, N, O) 3D-chiral quadratic indices considering H atoms in the molecular pseudograph ($G$) [$^*q_{kL}^H(\bar{x}_{E-H})$]. These local descriptors are putative H-bonding donors.

The $k$th stochastic total [$^{*s}q_k(\bar{x})$ and $^{*s}q_k^H(\bar{x})$] and local [$^{*s}q_{kL}(\bar{x}_E)$, $^{*s}q_{kL}^H(\bar{x}_E)$, and $^{*s}q_{kL}^H(\bar{x}_E - H)$] 3D-chiral quadratic indices were also computed. All these molecular descriptors were calculated with the in-house TOMOCOMD-CARDD software.[75]

**Table 3.** Values of trigonometric 3D-chirality correction factor $\{\sin[(\omega_A + 4\varDelta)\pi/2)]\}$ within the allowed domain

| $\omega_A$ | $\varDelta$ | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **−7** | **−6** | **−5** | **−4** | **−3** | **−2** | **−1** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| $\omega_{R=}$ 1 | 1 | | 1 | | 1 | | 1 | | 1 | | 1 | | 1 | | 1 |
| $\omega_{\text{non-chiral}=}$ 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | |
| $\omega_{S=}$ −1 | −1 | | −1 | | −1 | | −1 | | −1 | | −1 | | −1 | | −1 |

## 3. Results and discussion

The molecular set used in our study is made up of 31 steroids for which the binding affinity to the corticosteroid-binding globulin was measured. The so-called Cramer's steroid data set is well known to QSAR researchers,[5–20] and the molecular structures have already been depicted in several papers, so they are not shown here. For more details, see, for example, Figure 1 in Ref. 4 or Figure 1 in Ref. 11. Table 4 gathers the entire studied set with the actual binding affinities, taken from Robert et al.[13] The obtained models are given below together with their statistical parameters:

$$CBG = -6.39(\pm0.08) - 6.75(\pm1.04)^* q_{10L}(x_E)$$
$$+ 5.72(\pm1.06)^* q_{11L}(x_E)$$
$$- 0.43(\pm0.15)^* q_{2L}(x_{E-H})$$
$$+ 2.47(\pm0.72)^* q_2(x) - 0.73(\pm0.21)^* q_{13}(x)$$
$$- 1.56(\pm0.57)^* q_2^H(x) \tag{11}$$

$N = 31; \quad R = 0.924; \quad R^2 = 0.855; \quad F(6, 24) = 23.50;$
$s = 0.46; \quad q^2 = 0.781; \quad s_{cv} = 0.51; \quad p < 0.0001$

$$CBG = -6.39(\pm0.08) + 36.47(\pm9.94)^{*s} q_{10}^H(x)$$
$$- 37.15(\pm9.92)^{*s} q_{12}^H(x)$$
$$+ 4.46(\pm1.13)^{*s} q_{0L}(x_E)$$
$$+ 3.34(\pm1.07)^{*s} q_{6L}(x_E)$$
$$- 7.71(\pm1.54)^{*s} q_2^H(x_E)$$
$$- 7.28(\pm1.08)^{*s} q_{14L}(x_{E-H})$$
$$+ 7.64(\pm1.09)^{*s} q_{15L}(x_{E-H}) \tag{12}$$

$N = 31; \quad R = 0.929; \quad R^2 = 0.863; \quad F(7, 23) = 20.75;$
$s = 0.46; \quad q^2 = 0.735; \quad s_{cv} = 0.56 \quad p < 0.0001$

As it can be seen, the non-stochastic model (Eq. 11) explains more than 85% of the variance of the experimental CBG values using six variables to describe the 31 steroids, while the stochastic model (Eq. 12) explains more than 86% of this experimental value using seven variables. Both models showed a small value of standard deviation ($s = 0.46$).

An important aspect of QSAR modeling is the development of a way to validate the model. Good direct

**Table 4.** Results for the steroids data set used for QSAR study

|  |  | Observed CBG affinity (p$K_a$)[a] | Pred value[b] | % E[c] | % $E_{cv}$[d] | Pred value[e] | % E[c] | % $E_{cv}$[d] |
|---|---|---|---|---|---|---|---|---|
| 1 | Aldosterone | −6.279 | −6.259 | 0.324 | 0.367 | −6.339 | −0.955 | −3.237 |
| 2 | Androstanediol | −5.000 | −5.500 | −9.996 | −15.287 | −4.641 | 7.172 | 8.753 |
| 3 | Androstenediol | −5.000 | −5.026 | −0.511 | −0.624 | −5.008 | −0.169 | −0.250 |
| 4 | Androstenedione | −5.763 | −6.257 | −8.581 | −10.473 | −6.615 | −14.791 | −17.111 |
| 5 | Androsterone | −5.613 | −5.286 | 5.829 | 7.106 | −5.526 | 1.550 | 1.906 |
| 6 | Corticosterone | −7.881 | −7.242 | 8.110 | 9.038 | −7.204 | 8.586 | 10.403 |
| 7 | Cortisol | −7.881 | −7.226 | 8.315 | 10.163 | −7.370 | 6.483 | 7.951 |
| 8 | Cortisone | −6.892 | −7.212 | −4.638 | −6.224 | −6.908 | −0.232 | −0.301 |
| 9 | Dehydroepiandrosterone | −5.000 | −4.713 | 5.743 | 7.508 | −5.094 | −1.873 | −2.113 |
| 10 | Deoxycorticosterone | −7.653 | −7.081 | 7.479 | 8.141 | −7.067 | 7.660 | 8.589 |
| 11 | Deoxycortisol | −7.881 | −7.238 | 8.164 | 9.249 | −7.393 | 6.194 | 7.436 |
| 12 | Dihydrotestosterone | −5.919 | −5.570 | 5.904 | 7.330 | −5.477 | 7.461 | 8.390 |
| 13 | Estradiol | −5.000 | −5.096 | −1.922 | −3.054 | −5.454 | −9.070 | −13.408 |
| 14 | Estriol | −5.000 | −5.061 | −1.217 | −2.435 | −4.937 | 1.268 | 1.926 |
| 15 | Estrone | −5.000 | −4.747 | 5.060 | 8.780 | −5.175 | −3.502 | −4.838 |
| 16 | Ethiocholanolone | −5.255 | −5.286 | −0.586 | −0.714 | −5.526 | −5.157 | −6.341 |
| 17 | Pregnenolone | −5.255 | −5.579 | −6.157 | −7.064 | −5.671 | −7.918 | −9.156 |
| 18 | 17-Hydroxyregnenolone | −5.000 | −5.536 | −10.712 | −16.112 | −5.598 | −11.967 | −17.313 |
| 19 | Progesterone | −7.380 | −6.939 | 5.981 | 7.013 | −7.110 | 3.652 | 4.252 |
| 20 | 17-Hydroxyprogesterone | −7.740 | −7.071 | 8.637 | 10.090 | −7.116 | 8.059 | 9.626 |
| 21 | Testosterone | −6.724 | −6.572 | 2.259 | 2.436 | −6.619 | 1.569 | 1.968 |
| 22 | Prednisolone | −7.512 | −7.812 | −3.987 | −5.421 | −7.830 | −4.234 | −5.684 |
| 23 | Cortisolacetate | −7.553 | −7.789 | −3.125 | −3.907 | −7.957 | −5.345 | −12.580 |
| 24 | 4-Pregnene-3,11,20-trione | −6.779 | −7.069 | −4.282 | −5.672 | −6.587 | 2.831 | 3.741 |
| 25 | Epicorticosterone | −7.200 | −7.425 | −3.130 | −3.581 | −7.671 | −6.543 | −8.439 |
| 26 | 19-Nortestosterone | −6.144 | −6.549 | −6.587 | −8.529 | −6.016 | 2.079 | 2.495 |
| 27 | 16α,17α-Dihydroxyprogesterone | −6.247 | −7.232 | −15.760 | −16.963 | −5.715 | 8.513 | 12.495 |
| 28 | 16α-Methylprogesterone | −7.120 | −7.243 | −1.723 | −3.157 | −7.319 | −2.795 | −3.742 |
| 29 | 19-Norprogesterone | −6.817 | −6.949 | −1.931 | −2.317 | −6.936 | −1.748 | −2.024 |
| 30 | 2α-Methylcortisol | −7.688 | −7.756 | −0.882 | −1.095 | −8.122 | −5.646 | −8.039 |
| 31 | 2α-Methyl-9α-fluorocortisol | −5.797 | −5.657 | 2.412 | 5.708 | −5.970 | −2.988 | −17.663 |

[a] Observed CBG affinity values taken from Ref. 13.
[b] Predicted CBG affinity values using Eq. 11.
[c] Percentage of relative error; % $E = 100 \times [Obsd \times Pred/Obsd]$.
[d] Percentage of relative errors in leave-one-out cross-validation procedure; % $E_{cv} = 100 \times [Obsd \times Pred_{LOO-CV}/Obsd]$.
[e] Predicted CBG affinity values using Eq. 12.

statistical criteria to fit the data set are not a guarantee that the model can make accurate predictions for compounds outside the data set. The leave-one-out (LOO) statistic has been used as a means of demonstrating predictive capability. These models showed cross-validation square correlation coefficients of 0.781 and 0.735, respectively. These values of $q^2 (q^2 > 0.5)$ can be considered as a proof of the high predictive ability of the models.[76–79]

Now we are going to discuss the non-coincidence of the molecular descriptors used in Eqs. 11 and 12. To develop this topic, we must first analyze the analogue equations. The new Eqs. 11a and 12a are analogues of Eqs. 11 and 12, correspondingly; these new equations with their statistical parameters are shown below:

$$CBG = -6.39(\pm 0.12) - 0.36(\pm 1.35)^{*s}q_{10L}(x_E)$$
$$- 0.57(\pm 1.41)^{*s}q_{11L}(x_E)$$
$$+ 0.17(\pm 0.24)^{*s}q_{2L}(x_{E-H})$$
$$+ 6.63(\pm 1.77)^{*s}q_2(x)$$
$$- 5.64(\pm 1.62)^{*s}q_{13}(x)$$
$$- 0.88(\pm 0.42)^{*s}q_2^H(x) \quad (11a)$$

$N = 31;\quad R = 0.826;\quad R^2 = 0.628;\quad F(6,24) = 8.583;$
$s = 0.68;\quad p < 0.0001$

$$CBG = -6.39(\pm 0.11) - 0.07(\pm 1.65)^{*}q_{10}^H(x)$$
$$- 0.51(\pm 1.65)^{*}q_{12}^H(x)$$
$$+ 2.41(\pm 0.84)^{*}q_{0L}(x_E)$$
$$+ 1.55(\pm 0.81)^{*}q_{6L}(x_E)$$
$$- 3.56(\pm 1.23)^{*}q_2^H(x_E)$$
$$- 8.56(\pm 1.96)^{*}q_{14L}(x_{E-H})$$
$$+ 8.74(\pm 1.98)^{*}q_{15L}(x_{E-H}) \quad (12a)$$

$N = 31;\quad R = 0.881;\quad R^2 = 0.775;\quad F(7,23) = 11.34;$
$s = 0.58;\quad p < 0.0001$

Even when these new equations are both significant from the statistical point of view each of them presents three non-significant parameters, which are $^{*s}q_{10L}(x_E)$, $^{*s}q_{11L}(x_E)$, and $^{*s}q_{2L}(x_{E-H})$ for model 11a as well as $^{*}q_{10}^H(x)$, $^{*}q_{12}^H(x)$, and $^{*}q_{6L}(x_E)$ for model 12a. In addition, it is remarkable that the statistical parameters of the analogue equations are worse in comparison with the statistical parameters of the original equations (Eqs. 11 and 12). The determination coefficient decreases from 0.855 (Eq. 11) to 0.628 (Eq. 11a) and from 0.863 (Eq. 12) to 0.775 (Eq. 12a); the standard deviation increases the value from 0.46 (Eq. 11) to 0.68 (Eq. 11a) and from 0.46 (Eq. 12) to 0.58 (Eq. 12a). Therefore, we can say that the transformations of the non-stochastic descriptors into stochastic descriptors and vice versa do not improve the parameters of the models.

As we mentioned above, in this study we used the entire data set as training set and LOO cross-validation as a procedure to verify the predictive power of the model. On the other hand, in the development of our study no compound was detected as statistical outlier, but some compounds of this data set have been reported as outliers in many other researches.[4,5,13,16,29,80,81] In addition, it is remarkable that different validation procedures have been used for this data set, but the LOO cross-validation is the most widely used. Specifically, the values of cross-validation square correlation coefficients have been used for the sake of comparability.[13,28] As we previously pointed out, one of the objectives of the present report is to compare with other 3D-QSAR methods used for describing the property under study. For that reason, we only take into account QSAR models that used the entire data set of 31 steroids. The results of these works are summarized in Table 5, where the results were arranged in decreasing value of $q^2$ and the comparison can be more easily carried out. However, it should be remarked that

**Table 5.** Comparison of 3D-chiral quadratic indices prediction for the CBG of the steroid data set with other 3D QSAR approaches

| QSAR method | $N$ | $n$ | Statistical method | $q^2$ | Reference |
|---|---|---|---|---|---|
| Similarity matrices-based molecular descriptors | 31 | 6 | genetic NN | 0.940 | 9 |
| TQSAR | 31 | 6 | MLR after PCA | 0.842 | 13 |
| 3D-chiral linear indices (stochastic) | 31 | 7 | MLR | 0.788 | 28 |
| 3D-chiral quadratic indices (non-stochastic) | 31 | 6 | MLR | 0.781 | Eq. 11 |
| MEDV | 31 | 5 | GA and RLM | 0.777 | 20 |
| TQSI | 31 | 3 | MLR | 0.775 | 14 |
| 3D-chiral linear indices (non-stochastic) | 31 | 6 | MLR | 0.767 | 28 |
| MEDV | 31 | 6 | GA and RLM | 0.765 | 20 |
| 3D-chiral quadratic indices (stochastic) | 31 | 7 | MLR | 0.735 | Eq. 12 |
| Similarity indices | 31 | 1 | PLS | 0.734 | 16 |
| E-state and kappa shape index | 31 | 4 | MLR[a] | 0.730 | 18 |
| MQSM | 31 | 4 | MLR after PLS | 0.727 | 19 |
| E-state and kappa shape index | 31 | 4 | MLR | 0.720 | 18 |
| MQMS | 31 | 3 | MLR and PCA | 0.705 | 14 |
| CoMMA | 31 | 6 | PCR | 0.689 | 81 |
| MEDV | 31 | 4 | GA and RLM | 0.648 | 20 |
| Wagener's | 31 | — | k-NN and FNN | 0.630 | 5 |

$N$, number of steroids; $n$, number of variables; $q^2$, leave-one-out cross-validated coefficient of determination.
[a] One variable has a non-linear relationship.

the present QSAR method, non-stochastic and stochastic 3D-chiral quadratic indices, obtains results comparable to those of other highly predictive QSAR models even when they use more sophisticated statistical methods such as partial least squares, principal component analysis, non-linear neural network techniques, and so on. Many of the model objects of comparison were obtained from different procedures based on quantum mechanics and/or geometric principles as well as molecular mechanical approaches.

## 4. Conclusions

The aims of this work were: First, to demonstrate that 3D-chiral quadratic indices can be successfully applied to predict the corticosteroid-binding globulin binding affinity of the Cramer's steroid data set. In this sense, two models were developed which have good statistical parameters and adequate predictive power evidenced in the regression's statistical parameters and in the values of the leave-one-out press statistics. Second, to compare our result with those of other methods previously reported. From the previous analysis we can conclude that, for this data set, the 3D-chiral quadratic indices exhibit better or similar predictive ability when compared to those of other previously reported 3D-QSAR Methods.

## 5. Experimental

### 5.1. Data set

The molecular set used in this study is a well-known set for QSAR researches; it has already been studied by several authors.[4,5,9,11–18] As previously indicated, the steroid data set has been repeatedly employed by QSAR researches since it was introduced by Cramer et al. in 1988 using the CoMFA method.[4] Other methods used for developing 3D-QSAR models for this data set were Comparative Molecular Similarity Indices Analysis (CoMSIA),[12] Molecular Quantum Similarity Measures (MQSM),[13] Topological Quantum Similarity Indices (TQSI),[14] Comparative Molecular Moment Analysis (CoMMA)[11,37], and Mapping Property (MaP) Distributions of Molecular Surfaces[15] among others.[5,9,16–20]

### 5.2. Computational strategies

For the computation of 3D-chiral quadratic indices we used the TOMOCOMD software.[75] It is an interactive program for molecular design and bioinformatics research, which consists of four subprograms: 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 study, we used the CARDD module for the calculation of non-stochastic and stochastic 3D-chiral quadratic indices considering and not considering H atoms in the molecule.

The main steps for the application of the present method in QSAR/QSPR and drug design can be summarized briefly in the following algorithm:

1. Draw the molecular structure for each molecule of the data set, using the software drawing mode. This procedure is performed by a selection of the active atomic symbol belonging to the different groups in the periodic table of the elements.
2. Use appropriate weights in order to differentiate the molecular atoms. The weights used in this work are those previously proposed for the calculation of the DRAGON descriptors,[48–50] that is, atomic mass ($M$), atomic polarizability ($P$), atomic Mulliken electronegativity ($K$), van der Waals atomic volume ($V$) plus the atomic electronegativity in Pauling scale ($G$).[48] The values of these atomic labels are shown in Table 1.[48–50]
3. Compute the total and local (atomic and atom-type) non-stochastic and stochastic quadratic indices. It can be carried out in the software calculation mode, where one can select the atomic properties and the descriptor family 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 quadratic maps or other molecular descriptor family implemented in this program.
4. Find a QSPR/QSAR equation by using several multivariate analytical techniques, such as multilinear regression analysis, neural networks, linear discrimination analysis (LDA), and so on. Therefore, we can find a quantitative relation between an activity **A** and the quadratic fingerprints having, for instance, the following appearance,

$$\mathbf{A} = a_0^* q_0(x) + a_1^* q_1(x) + a_2^* q_2(x) + \dots + a_k^* q_k(x) + c$$

(13)

where **A** is the measured activity, $^*q_k(x)$ are the $k$th total 3D-chiral quadratic 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 validation technique.

### 5.3. Statistical analysis

Statistical analysis was carried out with the STATISTICA software.[82] The considered tolerance parameter (proportion of variance that is unique to the respective variable) was the default value for minimum acceptable tolerance, which is 0.01. Forward stepwise procedure was selected as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as strategy for model selection. In this connection, we selected the model with a high statistical signification but having as few parameters ($a_k$) as possible. A Multiple Linear Regression (MLR) was carried out to predict the CBG binding affinity of the steroid data set. The quality of models was determined by examining the regression's statistical parameters. Therefore, the

quality of models was determined by examining the determination coefficients (also known as square regression coefficients, $R^2$), Fisher-ratio's $p$ levels [$p(F)$], and standard deviations of the regression ($s$).[65] The leave-one-out press statistics ($q^2$, $s_{cv}$)[66,67] has been used as a means of demonstrating predictive capability.

## Acknowledgments

## References and notes

1. Foye, W. O. *Principles of Medicinal Chemistry*, 3rd ed.; Lea & Febiger: Philadelphia, 1989, pp 433–477.
2. Rampton, D. S. *Br. Med. J.* **1999**, *319*, 1480.
3. Baud, O.; Foix-L'Helias, L.; Kaminski, M.; Audibert, F.; Jarreau, P.; Papiernik, E.; Huon, C.; Lepercq, J.; Dehan, M.; Lacaze-Masmonteil, T. *N. Engl. J. Med.* **1999**, *341*, 1190.
4. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
5. Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Am. Chem. Soc.* **1995**, *117*, 7769.
6. Coats, E. A.. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; Vol. 3, pp 199–213.
7. Kubinyi, H. In *Computer-Assisted Lead Finding and Optimization Current Tools for Medicinal Chemistry*; Waterbeemd, H., van de, Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, 1997; pp 7–28.
8. Crippen, G. M. *J. Med. Chem.* **1997**, *40*, 3161.
9. So, S. S.; Karplus, M. *J. Med. Chem.* **1997**, *40*, 4347.
10. Turner, D. B.; Willet, P.; Ferguson, A. M.; Heritage, T. W. *J. Comput. Aided Mol. Des.* **1996**, *13*, 271.
11. Silverman, B. D. *Quant. Struct.-Act. Relat.* **2000**, *19*, 237.
12. Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994**, *37*, 4130.
13. Robert, D.; Amat, L.; Carbó-Dorca, R. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333.
14. Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465.
15. Stief, N.; Baumann, K. *J. Med. Chem.* **2003**, *46*, 1390.
16. Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. *J. Comput. Chem.* **1997**, *18*, 1344.
17. Chen, H.; Zhou, J.; Xie, G. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243.
18. Maw, H.; Hall, L. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248.
19. Besalu, E.; Girones, X.; Amat, L.; Carbo-Dorca, R. *Acc. Chem. Res.* **2002**, *35*, 289.
20. Liu, S.; Yin, C.; Wang, L. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 749.
21. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010.
22. Marrero-Ponce, Y. *Molecules* **2003**, *8*, 687.
23. Marrero-Ponce, Y.; Torrens, F. *Molecules*. Submitted for publication (see also ECSOC-9, Conference Hall *G*, G-014).
24. Marrero-Ponce, Y.; Torrens, F. *J. Comput. Aided Mol. Des.* Accepted for publication.
25. Marrero-Ponce, Y.; Torrens, F. *J. Phys. Chem. A.* Submitted for publication (see also ECSOC-9, Conference Hall *G*, G-015).
26. Marrero-Ponce, Y.; Torrens, F. *J. Mol. Struct. (THEOCHEM)*. Submitted for publication.
27. Marrero-Ponce, Y.; Díaz, H. G.; Romero, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
28. Marrero-Ponce, Y.; Castillo-Garit, J. A. *J. Comput. Aided Mol. Des.* **2005**, *19*, 369.
29. Marrero-Ponce, Y.; Huesca-Guillen, A.; Ibarra-Velarde, F. *J. Mol. Struct. (THEOCHEM)* **2005**, *717*, 67.
30. Edwards, C. H.; Penney, D. E. *Elementary Linear Algebra*; Prentice-Hall, Englewood Cliffs: New Jersey, USA, 1988.
31. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, M.; Mayón-Pérez, M.; García-Sánchez, R. *Bioorg. Med. Chem.* **2005**, *13*, 1293.
32. Marrero-Ponce, Y.; Iyarreta-Veitía, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Ávila, P. E.; Kirchgatter, K.; Machado, Y. *J. Chem. Inf. Model.* **2005**, *45*, 1082.
33. Marrero-Ponce, Y.; Medina-Marrero, R.; Martinez, Y.; Torrens, F.; Romero-Zaldivar, V.; Castro, E.A. *J. Mol. Model.* DOI doi:10.1007/s00894-005-0024-8.
34. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazábal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sánchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005.
35. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazábal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E. A. *J. Comput. Aided Mol. Des.* **2004**, *18*, 615.
36. Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 2881.
37. Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Montero Pereira, D.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Arán, V. J.; Martínez-Fernández, A. R.; García Sánchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *17*, 3838.
38. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351.
39. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. *Molecules* **2004**, *9*, 1100.
40. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, H. D.; Torrens, F. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186.
41. Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. *Internet Electron. J. Mol. Des.* **2005**, *4*, 124.
42. Montero-Torres, A.; Celeste Vega, A.; Marrero-Ponce, Y.; Rolón, M.; Gómez-Barrio, A.; Escario, J. A.; Arán, V. J.; Martínez-Fernández, A. R.; Meneses-Marcel, A. *Bioorg. Med. Chem.* **2005**, *13*, 6264–6275.
43. Marrero-Ponce, Y.; Meneses-Marcel, A.; Machado-Tugores, Y.; Montero Pereira, D.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Arán, V. J.; Martínez-Fernández, A. R., García Sánchez, R. N.; Montero-Torres, A.; Francisco Torrens. *Curr. Drug Discov. Technol.* Accepted for publication.
44. Marrero-Ponce, Y.; Nodarse, D.; González, H. D.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. *Int. J. Mol. Sci.* **2004**, *5*, 276–293.
45. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Nodarse, D. *Bioorg. Med. Chem.* **2005**, *13*, 3397–3404.

46. Marrero-Ponce, Y.; Medina, R.; Castro, E. A.; de Armas, R.; González, H.; Romero, V.; Torrens, F. *Molecules* **2004**, *9*, 1124–1147.
47. Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 3003–3015.
48. Pauling, L. *The Nature of Chemical Bond*; Cornell University Press: Ithaca (New York), 1939, pp 2–60.
49. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
50. Todeschini, R.; Gramatica, P. *Perspect. Drug Discov. Des.* **1998**, *9–11*, 355–380.
51. Rouvray, D. H. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976; pp 180–181.
52. Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; 2nd ed.; 1992; pp 32–33.
53. Gutman, I.; Polansky, O. E. *Mathematical Concepts in Organic Chemistry*; Springer: Berlin, 1986.
54. Estrada, E.; Patlewicz, G. *Croat. Chim. Acta* **2004**, *77*, 203.
55. Klein, D. J. *Internet Electron. J. Mol. Des.* **2003**, *2*, 814.
56. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model.* **2003**, *9*, 395.
57. González-Díaz, H.; Marrero, Y.; Hernández, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castañedo, N. C.; Cabrera-Pérez, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
58. Carbó-Dorca, R. *Int. J. Quantum Chem.* **2000**, *79*, 163–177.
59. González-Díaz, H.; Olazabal, E.; Castañedo, N.; Hernádez, I.; Morales, A.; Serrano, H.; González, J.; Ramos, R. *J. Mol. Model.* **2002**, *8*, 237.
60. González-Díaz, H.; Ramos de, A. R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
61. González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. *Bioorg. Med. Chem.* **2005**, *13*, 323.
62. González-Díaz, H.; Cruz-Monteagudo, M.; Viña, D.; Santana, L.; De Clercq, E.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1651.
63. Browder, A. *Mathematical Analysis. An Introduction*; Springer: New York, 1996, pp 176–296.
64. Axler, S. *Linear Algebra Done Right*; Springer: New York, 1996, pp 37-70.
65. Walker, P. D.; Mezey, P. G. *J. Am. Chem. Soc.* **1993**, *115*, 12423.
66. Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic: New York, 1999.
67. Estrada, E.; Molina, E. *J. Mol. Graphics Modell.* **2001**, *20*, 54.
68. Rien, B. *Comput. Chem.* **2002**, *26*, 223.
69. Eliel, E.; Wilen, S.; Mander, L. *Stereochemistry of Organic Compounds*; John Wiley & Sons: New York, 1994.
70. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de, A. R. *Bull. Math. Biol.* **2004**, *66*, 1285.
71. González-Díaz, H.; Tenorio, E.; Castañedo, N.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 1523.
72. Golbraikh, A.; Bonchev, D.; Tropsha, J. *Chem. Inf. Comput. Sci.* **2001**, *41*, 147.
73. González-Díaz, H.; Hernández-Sánchez, I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
74. de Julián-Ortiz, J. V.; de Alapont, C. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. *J. Mol. Graphics Modell.* **1998**, *16*, 14.
75. Marrero-Ponce, Y. Romero, V. TOMOCOMD software. Central University of Las Villas. 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 or ymarrero77@yahoo.es.
76. Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics,* Wiley: New York; 198.
77. Wold, S.; Erikson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1980; pp 309–318.
78. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.
79. Cronin, M. T. D.; Schultz, T. W. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 39.
80. Norinder, U. *J. Chemom.* **1996**, *10*, 533.
81. Silverman, B. D.; Platt, D. E. *J. Med. Chem.* **1996**, *39*, 2129.
82. STATISTICA version. 6.0, StatSoft.