QUBILS-MIDAS 3D-DESCRIPTORS

1. Mathematical Definition.

The QuBiLS-MIDAS molecular 3D-indices[1, 2] are computed from the atomic contribution of each atom in a molecule. In this way, if a molecule consists of *n* atoms, then the k^{th} *two-linear*, *three-linear* and *four-linear* indices for atom *"a"* are calculated as *n-linear algebraic maps*[3, 4] (forms) in $\mathbb{R}^n$, in a canonical basis set, and are expressed by the following equations, respectively:

$$ {}_{m}L_a = m^{a,k}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^{a,k} x^i y^j = [X]^T\,\mathbb{G}^{a,k}[Y] \qquad (\mathbf{1}) $$

$$ {}_{tr}L_a = tr^{a,k}(\bar{x},\bar{y},\bar{z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} gt_{ijl}^{a,k} x^i y^j z^l = \mathbb{G}\mathbb{T}^{a,k}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z} \qquad (\mathbf{2}) $$

$$ {}_{qu}L_a = qu^{a,k}(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} gq_{ijl}^{a,k} x^i y^j z^l w^h = \mathbb{G}\mathbb{Q}^{a,k}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z}\cdot\bar{w} \qquad (\mathbf{3}) $$

where, *"a"* indicates the atom ($a = 1,2,\ldots,n$), *n* is the number of atoms in a molecule, L_a is the entry corresponding to the contribution of the atom *"a"* in the vector of atom-level indices $\bar{L}$ [designated here by the well-known acronym: LOVI (LOcal Vertex Invariant)][5, 6] and $x^1,\ldots,x^n$, $y^1,\ldots,y^n$, $z^1,\ldots,z^n$ and $w^1,\ldots,w^n$ are the coordinates or components of the molecular vectors $\bar{x}$, $\bar{y}$, $\bar{z}$ and $\bar{w}$ in a system of canonical ('natural') basis vectors of $\mathbb{R}^n$.

The use of molecular vectors based on atomic properties as representation of the chemical structures has been used in other works.[7-10] As can be noticed, these molecular vectors are weighted with different "standard" atom- and fragment-based properties (weights) for atoms in a molecule and thus several combinations of algebraic forms are obtained (see Table 1). The weighting schemes (properties) used are the following: 1) atomic mass (M), 2) the van der Waals volume (V), 3) the atomic polarizability (P), 4) atomic electronegativity in Pauling scale (E), 5) atomic Ghose-Crippen LogP (A),[11-13] 6) atomic charge (C) (Gasteiger-Marsili),[14] 7) atomic polar surface area (PSA),[15] 8) atomic refractivity (R),[11-13] 9) atomic hardness (H) and 10) atomic softness (S).

Table 1. N-linear algebraic forms implemented in the QuBiLS program.

1. Two-linear [$\boldsymbol{m^k(\bar{x},\bar{y})}$]
- Linear (X, Y = 1)
- Bilinear (X <> Y)
- Quadratic (X = Y)

2. Three-linear [$\boldsymbol{tr^k(\bar{x},\bar{y},\bar{z})}$]
- Threelinear (X <> Y <> Z)
- Threelinear-Quadratic-Bilinear ((X = Y) <> Z)
- Threelinear-Bilinear (X <> Y, Z = 1)
- Threelinear-Linear (X, Y = 1, Z = 1)
- Threelinear-Cubic (X = Y = Z)

3. Four-linear [$\boldsymbol{qu^k(\bar{x},\bar{y},\bar{z},\bar{w})}$]
- Fourlinear (X <> Y <> Z <> W)
- Fourlinear-Quadratic-Threelinear ((X = Y) <> Z <> W)
- Fourlinear-Threelinear (X = 1, Y <> Z <> W)
- Fourlinear-Cubic-Bilinear ((X = Y = Z) <> W)
- Fourlinear-Bilinear (X = Y = 1, Z <> W)
- Fourlinear-Linear (X = Y = Z = 1, W)
- Fourlinear-Quadruple (X = Y = Z = W)

Used symbols
1: Using the unitary vector
<>: Using different properties
=: Using equal properties

The coefficients $g_{ij}^{a,k}$, $gt_{ijl}^{a,k}$ and $gq_{ijlh}^{a,k}$ are the elements of the k^{th} *two-tuples*, *three-tuples* and *four-tuples atom-level spatial-(dis)similarity matrices*, $\mathbb{G}^{a,k}$, $\mathbb{GT}^{a,k}$ and $\mathbb{GQ}^{a,k}$ for atom "*a*", respectively. In this way, if each one of the k^{th} *two-tuples [three-tuples, four- tuples] atom-level matrices* for a molecule are summed, then is obtained the corresponding k^{th} *two-tuples [three-tuples, four-tuples] total spatial-(dis)similarity matrix*, $\mathbb{G}^{k}$ [$\mathbb{GT}^{k}$, $\mathbb{GQ}^{k}$] (see section **2** for mathematical definition). Therefore, each *atom-level matrix* define an *atom-level index* for atom *"a"* (see Eqs. **1-3**). Lastly, the coefficients $g_{ij}^{a,k}$, $gt_{ijl}^{a,k}$ and $gq_{ijlh}^{a,k}$ are obtained from the coefficients g_{ij}^{k} of the $\mathbb{G}^{k}$, $gt_{ijl}^{a,k}$ of the $\mathbb{GT}^{k}$ and $gq_{ijlh}^{a,k}$ of the $\mathbb{GQ}^{k}$, respectively, as follows:

$$\begin{aligned} g_{ij}^{a,k} &= g_{ij}^{k} \quad \textit{if the two atoms are equals atom "a"} \\ &= \frac{1}{2} g_{ij}^{k} \quad \textit{if one atom is equal to atom "a"} \\ &= 0 \quad \textit{otherwise} \end{aligned} \qquad (\mathbf{4})$$

$$
\begin{aligned}
gt_{ijl}^{a,k} &= gt_{ijl}^{k} && \text{if the three atoms are equals to atom "a"} \\
&= \frac{2}{3} gt_{ijl}^{k} && \text{if two atoms are equals to atom "a"} \\
&= \frac{1}{3} gt_{ijl}^{k} && \text{if one atom is equal to atom "a"} \\
&= 0 && \text{otherwise}
\end{aligned} \qquad (5)
$$

$$
\begin{aligned}
gq_{ijlh}^{a,k} &= gq_{ijlh}^{k} && \text{if the four atoms are equals to atom "a"} \\
&= \frac{3}{4} gq_{ijlh}^{k} && \text{if three atoms are equals to atom "a"} \\
&= \frac{2}{4} gq_{ijlh}^{k} && \text{if two atoms are equals to atom "a"} \\
&= \frac{1}{4} gq_{ijlh}^{k} && \text{if one atom is equal to atom "a"} \\
&= 0 && \text{otherwise}
\end{aligned} \qquad (6)
$$

Finally and right from the previous definitions (see Eqs. **1-3**), the k^{th} *total (whole-molecule) bilinear, quadratic, linear, three-linear* and *four-linear* indices (QuBiLS-MIDAS MDs) can be calculated applying a set of *aggregation operators* (also called *invariants*) defined in the reports[16, 17], to the vector of atomic contributions, $\bar{L}$, for instance: the sum of the atom-level indices (components of $\bar{L}$) to aggregate the information captured by them.

2. N-tuples Spatial-(Dis) Similarity Matrices to Represent 3D-Information of the Chemical Structures.

The codification of 3D information of the chemical structures to compute the proposed indices is performed through the k^{th} *two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices* [$\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$] for the relations among two, three and four atoms respectively (see Eqs. **1-3**). The superscript *k* indicates the power to which $\mathbb{G}$, $\mathbb{GT}$ and $\mathbb{GQ}$ are raised. In this way, for *k = 0* all entries of the matrices $\mathbb{G}^0$, $\mathbb{GT}^0$ and $\mathbb{GQ}^0$ have value 1 and for *k = 1* the coefficients g_{ij}^{1}, gt_{ijl}^{1} and gq_{ijlh}^{1} corresponding to the matrices $\mathbb{G}^1$, $\mathbb{GT}^1$ and $\mathbb{GQ}^1$ represent the information of the interactions among two, three and four atoms respectively. The definition of the coefficients g_{ij}^{1}, gt_{ijl}^{1} and gq_{ijlh}^{1} is shown below:

$$\begin{aligned} g_{ij}^{1} &= D_{ij} \quad i \neq j \\ &= L_{ij} \quad i = j \text{ and lone-pairs are considered } (\text{or } D_{io}) \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{7}$$

$$\begin{aligned} gt_{ijl}^{1} &= TT_{ijl} \quad \text{if atoms } i, j \text{ and } l \text{ are not equal} \\ &= L_{ijl} \quad i = j = l \text{ and lone-pairs are considered } (\text{or } D_{io}) \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{8}$$

$$\begin{aligned} gq_{ijlh}^{1} &= QQ_{ijlh} \text{ if atoms } i, j, l \text{ and } h \text{ are not equal} \\ &= L_{ijlh} \quad i = j = l = h \text{ and lone-pairs are considered } (\text{or } D_{io}) \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{9}$$

where, D_{ij} is the (dis)-similarity between atomic nuclei *i* and *j* (see Table 1), TT_{ijl} is an measure for ternary relations of atoms and QQ_{ijlh} is an measure for quaternary relations of atoms. The coefficients L_{ij}, L_{ijl} and L_{ijlh} represents the diagonal entries of the matrices $\mathbb{G}^1$, $\mathbb{GT}^1$ and $\mathbb{GQ}^1$ respectively, which for a greater discrimination of the molecular structures could have assigned two different values: 1) the number of lone-pairs electrons for atoms, or 2) the Euclidean spatial distance, D_{io} for each atom *i* and center of the molecule, *o*.

Table 1. Metrics used to compute the "distance" between two atoms of a molecule.

Metrics	Formula[a]	Range[b]	Average	Range
Minkowsky (**m1-m7**) p = 0.25, 0.5, 1, 1.5, 2, 2.5, 3, and ∞ [where, when p= 1 it is the Manhattan, city-block or taxi distance (also known as Hamming distance between binary vectors) and p = 2 is Euclidean distance)	$d_{XY} = \left(\sum_{j=1}^{h} \lvert x_j - y_j \rvert^{p}\right)^{\frac{1}{p}}$	$[0, \infty)$	$\bar{d} = \dfrac{d_{XY}}{n^{1/p}}$	$[0, \infty)$
Chebyshev/Lagrange (**m8**) (Minkowsky formula when p = ∞)	$d_{XY} = max\{\lvert x_j - y_j \rvert\}$			
Canberra (**m10**)	$d_{XY} = \sum_{j=1}^{h} \dfrac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$	$[0, n]$	$\bar{d} = \dfrac{d_{XY}}{n}$	$[0,1]$
Lance - Williams/Bray-Curtis (**m11**)	$d_{XY} = \dfrac{\sum_{j=1}^{h} \lvert x_j - y_j \rvert}{\sum_{j=1}^{h}(\lvert x_j \rvert + \lvert y_j \rvert)}$	$[0,1]$	$\bar{d} = \dfrac{d_{XY}}{n}$	$\left[0, \dfrac{1}{n}\right]$
Clark/Coefficient of Divergence (**m12**)	$d_{XY} = \sqrt{\sum_{j=1}^{h} \left(\dfrac{x_j - y_j}{\lvert x_j \rvert + \lvert y_j \rvert}\right)^{2}}$	$[0, n]$	$\bar{d} = \dfrac{d_{XY}}{\sqrt{n}}$	$[0, \sqrt{n}]$

Soergel (**m13**)	$d_{XY} = \frac{1}{n}\sum_{j=1}^{h}\frac{\lvert x_j - y_j\rvert}{max\{x_j, y_j\}}$	[0,1]	$\bar{d} = \frac{d_{XY}}{n}$	$\left[0, \frac{1}{n}\right]$
Bhattacharyya (**m14**)	$d_{XY} = \sqrt{\sum_{j=1}^{h}\left(\sqrt{x_j} - \sqrt{y_j}\right)^2}$	$[0, \infty)$	$\bar{d} = \frac{d_{XY}}{\sqrt{n}}$	$[0, \infty)$
Wave – Edges (**m15**)	$d_{XY} = \sum_{j=1}^{h}\left(1 - \frac{min\{x_j, y_j\}}{max\{x_j, y_j\}}\right)$	$[0, n]$	$\bar{d} = \frac{d_{XY}}{n}$	[0,1]
Angular Separation/[1-Cosine (**Ochiai**)] (**m16**)	$d_{XY} = 1 - Cos_{XY}$ where, $Cos_{XY} = \frac{\boldsymbol{XY}}{\lVert\boldsymbol{X}\rVert\lVert\boldsymbol{Y}\rVert} = \frac{\sum_{j=1}^{h} x_j y_j}{\sqrt{\sum_{j=1}^{h} x_j^2 \sum_{j=1}^{h} y_j^2}}$	[0,2]		

---

[a] The variable $x_j(y_j)$ is the value of the coordinate *j* of the atom *s* and the atom *t*, corresponding to the molecule *X* (*Y*), respectively. The *h* value is the Cartesian coordinates (x, y, z) of an atom. The *p* values in Minkowsky metric are 0.25, 0.5, 1 (Manhattan), 1.5, 2 (Euclidean), 2.5 and 3 (Minkowsky). [b] "*Range*" refers to "range" and not to "rank" and is defined as $Range = max\{x_j\} - min\{x_j\}$.

As can be previously observed, the *two-tuples spatial-(dis)similarity matrix* of order 1 ($\mathbb{G}^1$) constitute a generalization of the *geometrical matrix*[18] where each entry only correspond to the Euclidean distance[19-22] between two atoms. On the other hand and as can be analyzed, the sub-indices *i*, *j*, *l* and *h* belonging to the ternary and quaternary measures (TT_{ijl}, QQ_{ijlh}) represent the atoms of the non-covalent interactions that are codified. Thus, the values of these sub-indices are not always different whereby the distinct combinations of them are considered. In this way, the *three-tuples (or four-tuples) spatial-(dis)similarity matrices* can be built using only ternary (or quaternary) measures or as from the reducing of ternary (or quaternary) measures to the corresponding inferior measures. Therefore, the following options are into accounted to build the n-tuples matrices:

- Ternary relations:

$$3nC\ (non-complete):\ TT_{ijl} = \begin{cases} T_{ijl} & three\ different\ atoms \\ 0 & otherwise \end{cases} \tag{\textbf{10}}$$

$$3C\ (complete):TT_{ijl} = \begin{cases} T_{ijl} & three\ different\ atoms \\ D_{ij} & two\ equal\ atoms\ and\ one\ different\ atom \\ 0 & otherwise \end{cases} \tag{\textbf{11}}$$

- Quaternary relations:

$$4nC\ (non-complete):\ QQ_{ijlh} = \begin{cases} Q_{ijlh} & four\ different\ atoms \\ 0 & otherwise \end{cases} \qquad \mathbf{(12)}$$

$$4C\ (complete):\ QQ_{ijlh} = \begin{cases} Q_{ijlh} & four\ different\ atoms \\ T_{ijl} & two\ equal\ atoms\ and\ two\ different\ atoms \\ D_{ij} & three\ equal\ atoms\ and\ one\ different\ atom \\ 0 & otherwise \end{cases} \qquad \mathbf{(13)}$$

where, 3*C* (or 4*C*) and 3*nC* (or 4*nC*) is the nomenclature assigned when the ternary (or quaternary) measures can be or not reduced, respectively. In addition, Q_{ijlh} is the measure used to establish the relation among four atoms (see Table 2B), T_{ijl} is the measure used to establish the relation among three atoms (see Table 2A), and D_{ij} is the distance between two atoms (see Table 1). Table 3 shown how the reduction of the ternary and quaternary measures is fulfilled. It is important to highlight, that to compute the ternary and quaternary measures is mandatory to select at least one *(dis)-similarity metric*, except for the calculation of the measures of Volume, Bond Angle and Dihedral Angle. This selected metric is also used when the n-way measures are reduced to considerer relations between two atoms.

Table 2. Measures used to compute the ternary (A) and quaternary (B) relations among atoms of a molecule.

A) Triple Measures *(TT_{XYZ})*	
Measure	**Formula**
Perimeter (**m19-m20**)	$T_{XYZ} = d_{xy} + d_{yz} + d_{zx}$
Triangle Area (**m21-m22**)	$T_{XYZ} = \sqrt{s(s-d_{XY})(s-d_{YZ})(s-d_{ZX})}$ $s = \frac{d_{XY}+d_{YZ}+d_{ZX}}{2}$
Summation Sides (**m25-m26**)	$T_{XYZ} = d_{XY} + d_{YZ}$

Bond angle (Angle between sides) (**m27-m28**)	A_X, A_Y, A_Z *coordinates of three atoms of a molecule* $U = A_X - A_Y, V = A_Z - A_Y$ $T_{XYZ} = \alpha = \arccos\left(\frac{U*V}{\lvert U\rvert * \lvert V\rvert}\right)$
B) Quaternary Measures *(QQ_{XYZW})*	
Perimeter (**m19-m20**)	$Q_{XYZW} = d_{XY} + d_{YZ} + d_{ZW} + d_{WX}$
Volume (**m23-m24**)	A_X, A_Y, A_Z, A_W *coordinates of four atoms of a molecule* $Q_{XYZW} = \frac{1}{6}\begin{pmatrix} A_{Y1} - A_{X1} & A_{Z1} - A_{X1} & A_{W1} - A_{X1} \\ A_{Y2} - A_{X2} & A_{Z2} - A_{X2} & A_{W2} - A_{X2} \\ A_{Y3} - A_{X3} & A_{Z3} - A_{X3} & A_{W3} - A_{X3} \end{pmatrix}$
Summation Sides (**m25-m26**)	$Q_{XYZW} = d_{XY} + d_{YZ} + d_{ZW}$
Dihedral Angle (**m29-m30**)	A_X, A_Y, A_Z *coordinates of three atoms of a molecule in the plane A* B_W, B_Y, B_Z *coordinates of three atoms of a molecule in the plane B* $U_A = (A_X - A_Y)\times(A_Z - A_y)$ $U_B = (B_W - A_Y)\times(B_Z - A_y)$ $Q_{XYZW} = \alpha = \arccos\left(\frac{U_A * U_B}{\lvert U_A\rvert * \lvert U_B\rvert}\right)$

Table 3. Reduction of the ternary and quaternary measures to compute the n-way relations among atoms of a molecule

Quaternary Measure *(Q_{ijlh})*		**Ternary Measure *(T_{ijl})***		**Distance Metric *(D_{ij})***
Perimeter (quadrilateral)	→	Perimeter (triangle)	→	Distance between two atoms
Volume	→	Triangle Area	→	Distance between two atoms
Summation Sides (three sides)	→	Summation Sides (two sides)	→	Distance between two atoms
Dihedral Angle	→	Bond Angle	→	0

The matrices $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$ for $k \geq 2$ are calculated multiplying the coefficients g_{ij}^{k-1}, gt_{ijl}^{k-1} and gq_{ijlh}^{k-1} of the matrices $\mathbb{G}^{k-1}$, $\mathbb{GT}^{k-1}$ and $\mathbb{GQ}^{k-1}$, respectively, by the corresponding coefficients g_{ij}^{1}, gt_{ijl}^{1} and gq_{ijlh}^{1} of the matrices $\mathbb{G}^1$, $\mathbb{GT}^1$ and $\mathbb{GQ}^1$, respectively. So, the elements of the matrix $\mathbb{G}^k$ will be equal to $\left(g_{ij}^{1}\right)^k$, the elements of the matrix $\mathbb{GT}^k$ will be equal to $\left(gt_{ijl}^{1}\right)^k$ and the elements of the matrix $\mathbb{GQ}^k$ will be equal to $\left(gq_{ijlh}^{1}\right)^k$. When algebraic transformations to normalize the elements of these matrices are not applied, then these are designated as k^{th} *non-stochastic two-tuples spatial-(dis)similarity matrix* (NS-SDSM, ${}_{ns}\mathbb{G}^k$), k^{th} *non-stochastic three-tuples spatial-(dis)similarity matrix* (NS-T-SDSM, ${}_{ns}\mathbb{GT}^k$) and k^{th} *non-stochastic four-tuples spatial-(dis)similarity matrix* (NS-Q-SDSM, ${}_{ns}\mathbb{GQ}^k$).

The proposed matrices $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$ can be considered as ***generalized matrices***.[18] These matrices are computed through the *Hadamard matrix product* and are obtained by raising the matrix elements both to positive or negative exponents. When the exponent *k* is negative then is computed the reciprocal to each entry of the n-tuples matrices, except for the diagonal elements when the numbers of lone-pairs is considered. The *k* values corresponds to non-covalent interactions among atoms of a molecule and its maximum value *(k = -12)* is related with the Lennard-Jones potential.

3. Normalization Formalisms based on Simple-Stochastic, Double-Stochastic and Mutual Probability Schemes.

With the purpose of normalize the non-stochastic n-tuples matrices [$\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$] are applied three probability schemes which are associated with inter-atomic interactions in the chemical structures. The probabilistic transformations have been used in other frameworks with successful results although these are not commonly employed in chemo-informatics studies.[7, 23-26]

In this work are used the k^{th} *simple-stochastic two-tuples spatial-(dis)similarity matrix* (SS-SDSM, ${}_{ss}\mathbb{G}^k$), k^{th} *simple-stochastic three-tuples spatial-(dis)similarity matrix* (SS-T-SDSM, ${}_{ss}\mathbb{GT}^k$) and k^{th} *simple-stochastic four-tuples spatial-(dis)similarity matrix* (SS-Q-SDSM, ${}_{ss}\mathbb{GQ}^k$). The coefficients ${}_{ss}g_{ij}^{k}$, ${}_{ss}gt_{ijl}^{k}$ and ${}_{ss}gq_{ijlh}^{k}$ corresponding to the matrices ${}_{ss}\mathbb{G}^k$, ${}_{ss}\mathbb{GT}^k$ and ${}_{ss}\mathbb{GQ}^k$, respectively, are calculated as follows:

$${}_{ss}g_{ij}^{k} = \frac{g_{ij}^{k}}{S_j} = \frac{g_{ij}^{k}}{\sum_{j=1}^{n} g_{ij}^{k}} \tag{14}$$

$${}_{ss}gt_{ijl}^{k} = \frac{{}_{ns}gt_{ijl}^{k}}{S_{jl}} = \frac{{}_{ns}gt_{ijl}^{k}}{\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt_{ijl}^{k}} \tag{15}$$

$${}_{ss}gq^{k}_{ijlh} = \frac{{}_{ns}gq^{k}_{ijlh}}{S_{jlh}} = \frac{{}_{ns}gq^{k}_{ijlh}}{\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq^{k}_{ijlh}} \tag{16}$$

where, ${}_{ns}g^{k}_{ij}$, ${}_{ns}gt^{k}_{ijl}$ and ${}_{ns}gq^{k}_{ijlh}$ are the elements of the matrices ${}_{ns}\mathbb{G}^{k}$, ${}_{ns}\mathbb{GT}^{k}$ and ${}_{ns}\mathbb{GQ}^{k}$ respectively, S_j is the summation of the coefficients of the row *i* in the matrix ${}_{ns}\mathbb{G}^{k}$ or the *spatial (dis)similarity vertex degree of order k* for the atom *i*, and S_{jl} and S_{jlh} is the summation of all entries of the two- and three-tuples matrices corresponding to each atom *i* in the non-stochastic n-tuples matrices ${}_{ns}\mathbb{GT}^{k}$ and ${}_{ns}\mathbb{GQ}^{k}$, respectively.

From these simple-stochastic algebraic transformations are obtained non-symmetric matrices and thus other approach to considerer is a double-stochastic scaling where the sum of the elements of each row and column is equal to 1. However, for the n-tuples matrices $(n > 2)$ does not exist reported algorithms to perform this transformation. Therefore, only have been employed the *k^th^ double-stochastic two-tuples spatial-(dis)similarity matrix* (DS-SDSM, ${}_{ds}\mathbb{G}^{k}$) which is computed through the *Sinkhorn and Knopp algorithm*.[27]

Lastly, the *k^th^ mutual-probability two-tuples spatial-(dis)similarity matrix* (MP-SDSM, ${}_{mp}\mathbb{G}^{k}$), *k^th^ mutual-probability three-tuples spatial-(dis)similarity matrix* (MP-T-SDSM, ${}_{mp}\mathbb{GT}^{k}$) and *k^th^ mutual-probability four-tuples spatial-(dis)similarity matrix* (MP-Q-SDSM, ${}_{mp}\mathbb{GQ}^{k}$) are used. With the mutual-probability transformation are obtained matrices where the summation of all entries is equal to 1. The coefficients ${}_{mp}g^{k}_{ij}$, ${}_{mp}gt^{k}_{ijl}$ and ${}_{mp}gq^{k}_{ijlh}$ corresponding to ${}_{mp}\mathbb{G}^{k}$, ${}_{mp}\mathbb{GT}^{k}$ and ${}_{mp}\mathbb{GQ}^{k}$, respectively, are computed as follows:

$${}_{mp}g^{k}_{ij} = \frac{g^{k}_{ij}}{S} = \frac{g^{k}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} g^{k}_{ij}} \tag{17}$$

$${}_{mp}gt^{k}_{ijl} = \frac{{}_{ns}gt^{k}_{ijl}}{S_{ijl}} = \frac{{}_{ns}gt^{k}_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt^{k}_{ijl}} \tag{18}$$

$${}_{mp}gq^{k}_{ijlh} = \frac{{}_{ns}gq^{k}_{ijlh}}{S_{ijlh}} = \frac{{}_{ns}gq^{k}_{ijlh}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq^{k}_{ijlh}} \tag{19}$$

where, S_{ij} , S_{ijl} and S_{ijlh} are the sample spaces belonging to the matrices ${}_{ns}\mathbb{G}^{k}$, ${}_{ns}\mathbb{GT}^{k}$ and ${}_{ns}\mathbb{GQ}^{k}$ respectively. The three sample spaces are computed by summing all elements of their respective matrices.

4. Local-Fragment (group, atom-type) *N*-tuples Spatial-(Dis)Similarity Matrices.

The previous n-tuples matrices used to represent the relations among two, three and four atoms (${}_{ns[ss,ds,mp]}\mathbb{G}^{k}$, ${}_{ns[ss,mp]}\mathbb{GT}^{k}$, ${}_{ns[ss,mp]}\mathbb{GQ}^{k}$) can be also employed to codify information related with groups or atom-types belonging to a specific molecular fragment *F*. In this way, are utilized the k^{th} *local-fragment two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, ${}_{ns[ss,mp]}\mathbb{GT}_{F}^{k}$, ${}_{ns[ss,mp]}\mathbb{GT}_{F}^{k}$ and ${}_{ns[ss,mp]}\mathbb{GQ}_{F}^{k}$, respectively. The elements of these local-fragment matrices are computed as shown below:

$$\begin{aligned} {}_{ns[ss,ds,mp]}g_{ijF}^{k} &= {}_{ns[ss,ds,mp]}g_{ij}^{k} && \textit{if the two atoms belongs to fragment F} \\ &= \frac{1}{2}\,{}_{ns[ss,ds,mp]}g_{ij}^{k} && \textit{if one atom belongs to fragment F} \\ &= 0 && \textit{otherwise} \end{aligned} \qquad \mathbf{(20)}$$

$$\begin{aligned} {}_{ns[ss,mp]}gt_{ijlF}^{k} &= {}_{ns[ss,mp]}gt_{ijl}^{k} && \textit{if the three atoms belongs to fragment F} \\ &= \frac{2}{3}\,{}_{ns[ss,mp]}gt_{ijl}^{k} && \textit{if two atoms belongs to fragment F} \\ &= \frac{1}{3}\,{}_{ns[ss,mp]}gt_{ijl}^{k} && \textit{if one atom belongs to fragment F} \\ &= 0 && \textit{otherwise} \end{aligned} \qquad \mathbf{(21)}$$

$$\begin{aligned} {}_{ns[ss,mp]}gq_{ijlhF}^{k} &= {}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{if the four atoms belongs to fragment F} \\ &= \frac{3}{4}\,{}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{if three atoms belongs to fragment F} \\ &= \frac{2}{4}\,{}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{if two atoms belongs to fragment F} \\ &= \frac{1}{4}\,{}_{ns[ss,mp]}gq_{ijlh}^{k} && \textit{if one atom belongs to fragment F} \\ &= 0 && \textit{otherwise} \end{aligned} \qquad \mathbf{(22)}$$

where, the coefficients ${}_{ns[ss,ds,mp]}g_{ijF}^{k}$, ${}_{ns[ss,mp]}gt_{ijlF}^{k}$ and ${}_{ns[ss,mp]}gq_{ijlhF}^{k}$ are the values of the *local-fragment matrices* ${}_{ns[ss,mp]}\mathbb{GT}_{F}^{k}$, ${}_{ns[ss,mp]}\mathbb{GT}_{F}^{k}$ and ${}_{ns[ss,mp]}\mathbb{GQ}_{F}^{k}$, respectively, and the elements

${}_{ns[ss,ds,mp]}g_{ij}^{k}$, ${}_{ns[ss,mp]}gt_{ijl}^{k}$ and ${}_{ns[ss,mp]}gq_{ijlh}^{k}$ are the (dis)similarity values represented in the total matrices ${}_{ns[ss,ds,mp]}\mathbb{G}^{k}$, ${}_{ns[ss,mp]}\mathbb{GT}^{k}$, ${}_{ns[ss,mp]}\mathbb{GQ}^{k}$, respectively.

It is important highlight that to the local-fragment matrices can be applied the algorithms specified in the Eqs. **4-6** and in this way determine the k^{th} *atom-level local-fragment matrices*, ${}_{ns[ss,mp]}\mathbb{GT}_{F}^{a,k}$, ${}_{ns[ss,mp]}\mathbb{GT}_{F}^{a,k}$ and ${}_{ns[ss,mp]}\mathbb{GQ}_{F}^{a,k}$. Therefore, these matrices can be used to compute the atom-level molecular indices for each atom *"a"* of a molecule, which are represented in the local-fragment LOVIs vector, ${}_{F}L_{a}$. In this way, the k^{th} *local-fragment bilinear, quadratic, linear, three-linear* and *four-linear indices* are calculated applying the aggregation operators over the atom-level local-fragment vector ${}_{F}\bar{L}$.

In this software, the local-fragment MDs can be calculated by seven chemical (or functional) groups in the molecule, these are: hydrogen bond acceptors (A), carbon atoms in aliphatic chains (C), hydrogen bond donors (D), halogens (G), terminal methyl groups (M), carbon atoms in aromatic portion (P) and heteroatoms (O, N and S in all valence states, denoted as X).

5. *N*-tuples Constraints to Consider Interactions According to Topological and/or Euclidean Geometric Distances.

The total (or local-fragment) matrices used to compute the total (or local-fragment) molecular descriptors always consider all interactions among atoms of a molecule. However, in the literature have been reported several transformations to the *geometry matrix* to take into account both topological as topographic aspects in the same representation. In this way, two different constraints both for the total and local QuBiLS-MIDAS indices are applied:

1) N-tuples Graph-theoretical cut-off (*p*) based on topological distance at a lag *p*, denoted as "*path cut-off*".
2) N-tuples Geometric cut-off (*l*), based on Euclidean distance at a lag *l* known as "*length cut-off*".

Through the application of these constraints are obtained the *two-*, *three- and four-tuples topological and geometric neighborhood quotient matrices*, denoted as, $\mathbb{NQG}^{1}$, $\mathbb{NQGT}^{1}$ and $\mathbb{NQGQ}^{1}$, respectively. The values of these matrices are the coefficients of the total (or local-fragment) matrices multiplied by a rate according to the amount interactions that have their topological and/or geometric distances smaller or equals to a predefined thresholds *p* and/or *l*. The *neighborhood quotient matrices* are computed as follows:

$$
\begin{aligned}
{}^{NQ}g_{ij}^{1} &= g_{ij}^{1} \text{ if } \mathrm{p_{min}} \le \mathrm{p_{ij}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij}} \le \mathrm{l_{max}} \\
&= 0 \;\; \text{otherwise}
\end{aligned}
\qquad \textbf{(23)}
$$

$$
\begin{aligned}
{}^{NQ}gt_{ijl}^{1} &= gt_{ijl}^{1} && \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij}}, \mathrm{p_{jl}}, \mathrm{p_{li}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij}}, \mathrm{l_{jl}}, \mathrm{l_{li}} \le \mathrm{l_{max}} \\
&= \frac{2}{3} gt_{ijl}^{1} && \begin{cases} \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij}}, \mathrm{p_{jl\,(li)}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij}}, \mathrm{l_{jl\,(li)}} \le \mathrm{l_{max}} \\ \text{if } \mathrm{p_{min}} \le \mathrm{p_{jl}}, \mathrm{p_{li}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{jl}}, \mathrm{l_{li}} \le \mathrm{l_{max}} \end{cases} \\
&= \frac{1}{3} gt_{ijl}^{1} && \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij\,(jl,li)}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij\,(jl,li)}} \le \mathrm{l_{max}} \\
&= 0 && \text{otherwise}
\end{aligned}
\qquad \textbf{(24)}
$$

$$
\begin{aligned}
{}^{NQ}gq_{ijlh}^{1} &= gq_{ijlh}^{1} && \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij}}, \mathrm{p_{jl}}, \mathrm{p_{lh}}, \mathrm{p_{hi}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij}}, \mathrm{l_{jl}}, \mathrm{l_{lh}}, \mathrm{l_{hi}} \le \mathrm{l_{max}} \\
&= \frac{3}{4} gq_{ijlh}^{1} && \begin{cases} \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij}}, \mathrm{p_{jl\,(lh)}}, \mathrm{p_{lh\,(hi)}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij}}, \mathrm{l_{jl\,(lh)}}, \mathrm{l_{lh\,(hi)}} \le \mathrm{l_{max}} \\ \text{if } \mathrm{p_{min}} \le \mathrm{p_{jl}}, \mathrm{p_{lh}}, \mathrm{p_{hi}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{jl}}, \mathrm{l_{lh}}, \mathrm{l_{hi}} \le \mathrm{l_{max}} \end{cases} \\
&= \frac{2}{4} gq_{ijlh}^{1} && \begin{cases} \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij}}, \mathrm{p_{jl\,(lh,hi)}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij}}, \mathrm{l_{jl\,(lh,hi)}} \le \mathrm{l_{max}} \\ \text{if } \mathrm{p_{min}} \le \mathrm{p_{jl}}, \mathrm{p_{lh\,(hi)}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{jl}}, \mathrm{l_{lh\,(hi)}} \le \mathrm{l_{max}} \\ \text{if } \mathrm{p_{min}} \le \mathrm{p_{lh}}, \mathrm{p_{hi}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{lh}}, \mathrm{l_{hi}} \le \mathrm{l_{max}} \end{cases} \\
&= \frac{1}{4} gq_{ijlh}^{1} && \text{if } \mathrm{p_{min}} \le \mathrm{p_{ij\,(jl,li,hi)}} \le \mathrm{p_{max}} \;\; or/and \;\; \mathrm{l_{min}} \le \mathrm{l_{ij\,(jl,lh,hi)}} \le \mathrm{l_{max}} \\
&= 0 && \text{otherwise}
\end{aligned}
\qquad \textbf{(25)}
$$

where, ${}_{ns[ss,ds,mp]}g_{ij}^{k}$, ${}_{ns[ss,mp]}gt_{ijl}^{k}$ and ${}_{ns[ss,mp]}gq_{ijlh}^{k}$ represents the pair-wise, ternary and quaternary relations among the corresponding atoms (see Eqs. **7-9**), and p_{XX} and l_{YY} are the user-defined thresholds corresponding to topological and geometrical (Euclidean) distances, respectively. Min and Max means the minimum and maximum cut-offs (rank).

In similar way to the reduction of the ternary and quaternary measures when the atoms taken into account are not all different (see Eqs. **11** and **13**), the calculation of the *three- and four-tuples topological and geometric neighborhood quotient matrices* can also be reduced. So, when the considered atoms in the four-tuples constraint (see Eq. **25**) are not distinct then can be applied the three-tuples constraint (see Eq. **24**), and if the three atoms in this last case are not different then can be applied the two-tuples constraint (see Eq. **23**).

These new truncated matrices, $\mathbb{NQG}^1$, $\mathbb{NQGT}^1$ and $\mathbb{NQGQ}^1$, also constitute classes of the ***generalized matrices*** (see above) and thus can be used to compute QuBiLS-MIDAS indices.

Moreover, it is important remark that the use of the constraints is not mandatory for the calculation of the described indices in this work.

References.

1. Marrero-Ponce, Y.; García-Jacas, C. R.; Barigye, S. J.; Valdés-Martiní, J. R.; Rivera-Borroto, O. M.; Pino-Urias, R. W.; Cubillán, N.; Alvarado, Y. J., Optimum Search Strategies or Novel 3D Molecular Descriptors: is there a Stalemate? **2013, In Press**.
2. García-Jacas, C. R.; Marrero-Ponce, Y.; Barigye, S. J.; Valdés-Martiní, J. R.; Rivera-Borroto, O. M.; Verbel, J. O., N-Linear Algebraic Maps to Codify Chemical Structures: is a suitable generalization to the atom-pairs approaches? **2013, In Press**.
3. Johnson, R. W.; Huang, C. H.; Johnson, J. R., Multilinear algebra and parallel programming. *J Supercomput* **1991**, 5, 189-217.
4. Hestenes, D.; Sobczyk, G., *Linear and Multilinear Functions*. Vol. 5.
5. Balaban, A., T., Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 398.
6. Todeschini, R., Consonni V., New Local Vertex Invariants and Molecular Descriptors Based on Functions of the Vertex Degrees. *MATCH Commun. Math. Comput. Chem* **2010**, 64, 359-372.
7. Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R.; Ortega-Broche, S. E.; Romero Zaldivar, V., Novel 2D TOMOCOMD-CARDD molecular descriptors: atom-based stochastic and non-stochastic bilinear indices and their QSPR applications. *J Math Chem* **2008**, 44, 650-673.
8. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F., Atom-based 3D-chiral quadratic indices. Part 2: Prediction of the corticosteroid-binding globulinbinding affinity of the 31 benchmark steroids data set. *Bioorg. Med. Chem.* **2006**, 14, 2398-2408.
9. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; Rotondo, R., Atom-based stochastic and non-stochastic 3D-chiral bilinear indices and their applications to central chirality codification. *Journal of Molecular Graphics and Modelling* **2007**, 26, 32-47.
10. Castillo-Garit, J. A.; Martinez-Santiago, O.; Marrero-Ponce, Y.; Casañola-Martín, G. M.; Torrens, F., Atom-based non-stochastic and stochastic bilinear indices: Application to QSPR/QSAR studies of organic compounds. *Chemical Physics Letters* **2008**, 464, 107-112.
11. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J., Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, 102, 3762–3772.
12. Balaban, A. T., *From Chemical Topology to Three-Dimensional Geometry*. Plenum Press: New York, 1997.
13. Balaban, A. T., *Steric fit in quantitative structure-activity relations*. Springer-Verlag: 1980; p 178.
14. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, 36, 3219-3288.
15. 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.

16. Barigye, S. J.; Marrero-Ponce, Y.; Martínez López, Y.; Artiles Martínez, L. M.; Pino-Urias, R. W.; Martínez Santiago, O.; Torrens, F., Relations Frequency Hypermatrices in Mutual, Conditional and Joint Entropy-Based Information Indices. *J. Comput. Chem.* **2013**, 34, 259-274.
17. Barigye, S. J.; Marrero-Ponce, Y.; Santiago, O. M.; López, Y. M.; Torrens, F., Shannon's, Mutual, Conditional and Joint Entropy-Based Information Indices. Generalization of Global Indices Defined from Local Vertex Invariants *Curr. Comput.-Aided Drug Des.* **2013**, 9.
18. Todeschini, R.; Consonni, V., *Molecular Descriptors for Chemoinformatics. Vol. 1. Alphabetical Listing; Vol. 2. Appendices, References*. Wiley-VCH: Weinheim, 2009; p 2125.
19. Andrew C. Good, S. S. S., W. Graham Richards, Structure-activity relationships from molecular similarity matrices. *Journal of Medicinal Chemistry* **1993**, 433-438.
20. Gasteiger, G.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V., Chemical Information in 3D Space. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 1030-1037.
21. Todeschini, R.; Consonni, V., *Molecular Descriptors for Chemoinformatics*. 1st ed.; WILEY-VCH: Weinheim, 2009; Vol. 1, p 667.
22. Bogdanov, B.; Nikolic, S.; Trinajstic, N., On the Three-Dimensional Wiener Number *J. Math. Chem.* **1989**, 3, 299-309.
23. Marrero-Ponce, Y.; Huesca-Guillén, A.; Ibarra-Velarde, F., Quadratic indices of the molecular pseudograph's atom adjacency matrix and their stochastic forms: a novel approach for virtual screening and in silico discovery of new lead paramphistomicide drugs-like compounds. *J. Mol. Struct.(THEOCHEM)* **2005**, 717, 67-79.
24. González-Díaz, H.; Uriarte, E., Proteins QSAR with Markov average electrostatic potentials. *Bioorg. Med. Chem. Lett.* **2005**, 15, 5088-5094.
25. Ramos de Armas, R.; González Díaz, H.; Molina, R.; Uriarte, E., Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. *PROTEINS: Struc. Funct. Bioinform.* **2004**, 56, 715–723.
26. Carbo-Dorca, R., Stochastic Transformation of Quantum Similarity Matrixes and Their Use in Quantum QSAR (QQSAR) Models. *Int. J. Quantum Chem.* **2000**, 79, 163-177.
27. Sinkhorn, R.; Knopp, P., Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* **1967**, 21, 343-348.