SAR and QSAR
in Environmental Research

Taylor & Francis
Taylor & Francis Group

SAR and QSAR in Environmental Research

ISSN: 1062-936X (Print) 1029-046X (Online) Journal homepage: http://www.tandfonline.com/loi/gsar20

N-tuple topological/geometric cutoffs for 3D N-linear algebraic molecular codifications: variability, linear independence and QSAR analysis

C. R. García-Jacas, Y. Marrero-Ponce, S. J. Barigye, T. Hernández-Ortega, L. Cabrera-Leyva & A. Fernández-Castillo

To cite this article: C. R. García-Jacas, Y. Marrero-Ponce, S. J. Barigye, T. Hernández-Ortega, L. Cabrera-Leyva & A. Fernández-Castillo (2016): N-tuple topological/geometric cutoffs for 3D N-linear algebraic molecular codifications: variability, linear independence and QSAR analysis, SAR and QSAR in Environmental Research, DOI: 10.1080/1062936X.2016.1231714

To link to this article: http://dx.doi.org/10.1080/1062936X.2016.1231714

View supplementary material

Published online: 06 Oct 2016.

Submit your article to this journal

View related articles

View Crossmark data

SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2016
http://dx.doi.org/10.1080/1062936X.2016.1231714

Taylor & Francis
Taylor & Francis Group

N-tuple topological/geometric cutoffs for 3D *N*-linear algebraic molecular codifications: variability, linear independence and QSAR analysis

C. R. García-Jacas[a,b,c], Y. Marrero-Ponce[d,e], S. J. Barigye[g], T. Hernández-Ortega[c], L. Cabrera-Leyva[f] and A. Fernández-Castillo[c]

[a]Escuela de Sistemas y Computación, Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE), Esmeraldas, Ecuador; [b]Grupo de Investigación de Bioinformática, Instituto de Química, Universidad Nacional Autónoma de México (UNAM), Ciudad de México, D.F, México; [c]Grupo de Investigacion de Bioinformatica, Universidad de las Ciencias Informaticas (UCI), La Habana, Cuba; [d]Grupo de Medicina Molecular y Traslacional (MeM&T), Universidad San Francisco de Quito (USFQ), Quito, Ecuador; [e]Instituto de Simulación Computacional (ISC-USFQ), Universidad San Francisco de Quito (USFQ), Quito, Ecuador; [f]Grupo de Investigación de Inteligencia Artificial (AIRES), Universidad de Camagüey, Camagüey, Cuba; [g]Department of Chemistry, McGill University, Montréal, Québec, Canada

ABSTRACT

Novel *N*-tuple topological/geometric cutoffs to consider specific inter-atomic relations in the QuBiLS-MIDAS framework are introduced in this manuscript. These molecular cutoffs permit the taking into account of relations between more than two atoms by using (dis-)similarity multi-metrics and the concepts related with topological and Euclidean-geometric distances. To this end, the *k*th two-, three- and four-tuple topological and geometric neighbourhood quotient (NQ) total (or local-fragment) spatial-(dis)similarity matrices are defined, to represent 3D information corresponding to the relations between two, three and four atoms of the molecular structures that satisfy certain cutoff criteria. First, an analysis of a diverse chemical space for the most common values of topological/Euclidean-geometric distances, bond/dihedral angles, triangle/quadrilateral perimeters, triangle area and volume was performed in order to determine the intervals to take into account in the cutoff procedures. A variability analysis based on Shannon's entropy reveals that better distribution patterns are attained with the descriptors based on the cutoffs proposed (QuBiLS-MIDAS NQ-MDs) with regard to the results obtained when all inter-atomic relations are considered (QuBiLS-MIDAS KA-MDs – 'Keep All'). A principal component analysis shows that the novel molecular cutoffs codify chemical information captured by the respective QuBiLS-MIDAS KA-MDs, as well as information not captured by the latter. Lastly, a QSAR study to obtain deeper knowledge of the contribution of the proposed methods was carried out, using four molecular datasets (steroids (STER), angiotensin converting enzyme (ACE), thermolysin inhibitors (THER) and thrombin inhibitors (THR)) widely used as benchmarks in the evaluation of several methodologies. One to four variable QSAR models based on multiple linear regression

ARTICLE HISTORY

Received 7 June 2016
Accepted 30 August 2016

KEYWORDS

3D molecular descriptors; *N*-tuple molecular cutoffs; QSAR; QuBiLS-MIDAS; TOMOCOMD-CARDD

CONTACT C. R. García-Jacas ✉ cesarrjacas1985@gmail.com

The Supplemental material for this article is available online at http://dx.doi.org/10.1080/1062936X.2016.1231714

were developed for each compound dataset following the original division into training and test sets. The results obtained reveal that the novel cutoff procedures yield superior performances relative to those of the QuBiLS-MIDAS KA-MDs in the prediction of the biological activities considered. From the results achieved, it can be suggested that the proposed *N*-tuple topological/geometric cutoffs constitute a relevant criteria for generating MDs codifying particular atomic relations, ultimately useful in enhancing the modelling capacity of the QuBiLS-MIDAS 3D-MDs.

Introduction

In recent reports, the *N*-linear molecular descriptors (MDs) based on *N*-tuple spatial (dis)-similarity matrices and atomic weightings (QuBiLS-MIDAS) were introduced as a novel mathematical method for computing 3D molecular structural features [1–3]. Specifically, the duplex QuBiLS-MIDAS MDs encode atom-pair relations using bilinear, quadratic and linear algebraic maps [1, 2], while the ternary and quaternary QuBiLS-MIDAS MDs codify information regarding the relations among three and four atoms by means of three-linear and four-linear algebraic maps, respectively [3]. To carry out this molecular codification the *k*th two-tuple, three-tuple and four-tuple spatial-(dis)similarity matrices were defined, which were used to represent the chemical information on the inter-atomic relations using several (dis-)similarity metrics (see Table 1 in [2] for atom-pair relations) and multi-metrics (see Table 1 in [3] for relations among three and four atoms).

Several studies aimed at assessing the quality of this novel descriptor family were performed, and these included an evaluation of the information content (variability) and linear independence [2, 3]. The results demonstrated that the novel MDs have superior variability over 3D DRAGON MDs and another approaches implemented in software applications [1, 4–7]. Furthermore, the results revealed that the novel MDs not only codify all of the information contained in the 3D DRAGON MDs, but capture orthogonal information to the latter. The QuBiLS-MIDAS MDs were also used to carry out a comparative study in QSAR modelling. First, a pilot QSAR study on the modelling of the binding affinity to corticosteroid-binding globulin (CBG) was developed, achieving better results with respect to other methodologies (see Tables 8 and 9 in [2] and Tables 9 and 10 in [3]). A deeper QSAR experimentation employing eight benchmark chemical datasets (angiotensin converting enzyme (ACE), acetylcholinesterase (AchE) inhibitors, benzodiazepine receptor (BZR), cyclooxygenase-2 (COX2) inhibitors, dihydrofolate reductase inhibitors (DHFR), glycogen phosphorylase b (GPB), thermolysin inhibitors (THER), thrombin inhibitors (THR) – see [8] for datasets description) was performed [9], attaining statistically superior outcomes (based on non-parametric analysis) with respect to 11 alignment-dependent or alignment-free QSAR procedures reported in the literature [1, 10–13].

To date, QuBiLS-MIDAS MDs take into account all relations between atoms of a molecule or those corresponding to chemical regions of interest (atom-types, local-fragments), denoting them as 'keep all MDs' (QuBiLS-MIDAS KA-MDs – in this manuscript the QuBiLS-MIDAS MDs are considered as QuBiLS-MIDAS KA-MDs, unless otherwise specified). However, several procedures reported in the literature have used topological and/or geometrical neighbourhood criteria in order to increase the chemical information codified and, consequently, improve the performance of QSAR models [14–16]. Also, these criteria have been employed

Table 1. Examples of two-tuple and three-tuple total spatial-(dis)similarity matrices calculated with some (dis-)similarity metrics and multi-metrics from the chemical structure of chloro(methoxy)methane (see Figure 1) (a) Examples of the two-tuple total spatial-(dis)similarity matrices for $k = 1$ calculated from Euclidean, Lance-Williams, Soergel and angular separation metrics; (b) Example of the three-tuple total spatial-(dis)similarity matrix for $k = 1$ calculated from bond angle multi-metric.

(a) Two-tuple total spatial-(dis)similarity matrices, $\mathbb{G}^1$.

$\mathbb{G}^1$ *based on Euclidean metric.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	2.408	1.439	3.939
C2	2.408	0.000	1.438	1.757
O3	1.439	1.438	0.000	2.598
Cl4	3.939	1.757	2.598	0.000

$\mathbb{G}^1$ *based on Lance-Williams metric.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.000	0.973	1.000
C2	1.000	0.000	0.954	0.293
O3	0.973	0.954	0.000	0.973
Cl4	1.000	0.293	0.973	0.000

$\mathbb{G}^1$ *based on Soergel metric.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.158	1.003	1.709
C2	1.158	0.000	1.234	1.359
O3	1.003	1.234	0.000	2.235
Cl4	1.709	1.359	2.235	0.000

$\mathbb{G}^1$ *based on angular separation metric.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.354	0.558	1.875
C2	1.354	0.000	0.318	0.237
O3	0.558	0.318	0.000	0.952
Cl4	1.875	0.237	0.952	0.000

(b) Three-tuple total spatial-(dis)similarity matrix, $\mathbb{GT}^1$.

$\mathbb{GT}^1$ slide 1*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	0.000
C2	0.000	0.000	0.578	2.470
O3	0.000	1.985	0.000	2.682
Cl4	0.000	0.390	0.163	0.000

$\mathbb{GT}^1$ slide *2ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.578	0.281
C2	0.000	0.000	0.000	0.000
O3	1.985	0.000	0.000	0.697
Cl4	0.390	0.000	0.553	0.000

$\mathbb{GT}^1$ slide *3ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.578	0.000	0.297
C2	0.578	0.000	0.000	1.892
O3	0.000	0.000	0.000	0.000
Cl4	0.163	0.553	0.000	0.000

$\mathbb{GT}^1$ slide *4ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.281	0.297	0.000
C2	2.470	0.000	1.892	0.000
O3	2.682	0.697	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

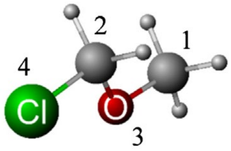


Figure 1. Chemical structure of chloro(methoxy)methane and its labelled molecular scaffold.

to compute matrix representations where only information related with specific interactions is considered, e.g. a topological cutoff on the geometric matrix produces a sparse matrix where the atom-pairs, not so distant from each other, are accounted for [17]. These kinds of matrices are well known as neighbourhood matrices, denoted by ^{N}M. On the other hand, other matrices have been defined using several combinations of 'topological', 'topographical' or 'geometrical' distances with the purpose of including diverse chemical information in the same representation. In this sense, it can be mentioned the quotient matrices (denoted by M1/M2), whose coefficients are the ratio of the off-diagonal elements of M1 over the corresponding values in M2, e.g. geometric distance/topological distance quotient matrix (G/D) [17, 18].

As may be perceived, all previous strategies can only be applied to bi-dimensional (duplex) matrices due to the fact that only atom-pair relations are taken into account in the descriptors calculation and, thus, these concepts cannot be used on the QuBiLS-MIDAS MDs when more than two atoms are considered. Therefore, this report is aimed at defining novel *N*-tuple molecular cutoff procedures with the purpose of considering relations among *N* atoms according to their topological and/or Euclidean geometric distances (for duplex), as well as with respect to multi-metrics analysis (for *N*-tuples), for use with the QuBiLS-MIDAS MDs computation. In addition, variability, linear independence and QSAR studies are performed in order to assess the feasibility of taking into account the novel cutoffs in the codification of molecular structures with QuBiLS-MIDAS MDs.

Overview of the QUBILS-MIDAS molecular descriptors

The *k*th two-linear, three-linear and four-linear algebraic indices (also known as QuBiLS-MIDAS) are computed as *N*-linear (multi-linear) algebraic forms (maps) in $\mathbb{R}^n$, in a canonical basis set, when relations among two ($N = 2$), three ($N = 3$) and four ($N = 4$) atoms are taken into account, respectively [1–3]. These indices are mathematically defined as follows:

$$ {}_{b(F)}L_a = b^{a,k}_{(F)}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} g^{a,k}_{ij(F)}x^i y^j = [X]^T\mathbb{G}^{\mathrm{a,k}}_{(\mathrm{F})}[Y] \tag{1} $$

$$ {}_{tr(F)}L_a = tr^{a,k}_{(F)}(\bar{x},\bar{y},\bar{z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} gt^{a,k}_{ijl(F)}x^i y^j z^l = \mathbb{GT}^{\mathrm{a,k}}_{(\mathrm{F})}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z} \tag{2} $$

$$ {}_{qu(F)}L_a = qu^{a,k}_{(F)}(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} gq^{a,k}_{ijlh(F)}x^i y^j z^l w^h = \mathbb{GQ}^{\mathrm{a,k}}_{(\mathrm{F})}\cdot\bar{x}\cdot\bar{y}\cdot\bar{z}\cdot\bar{w} \tag{3} $$

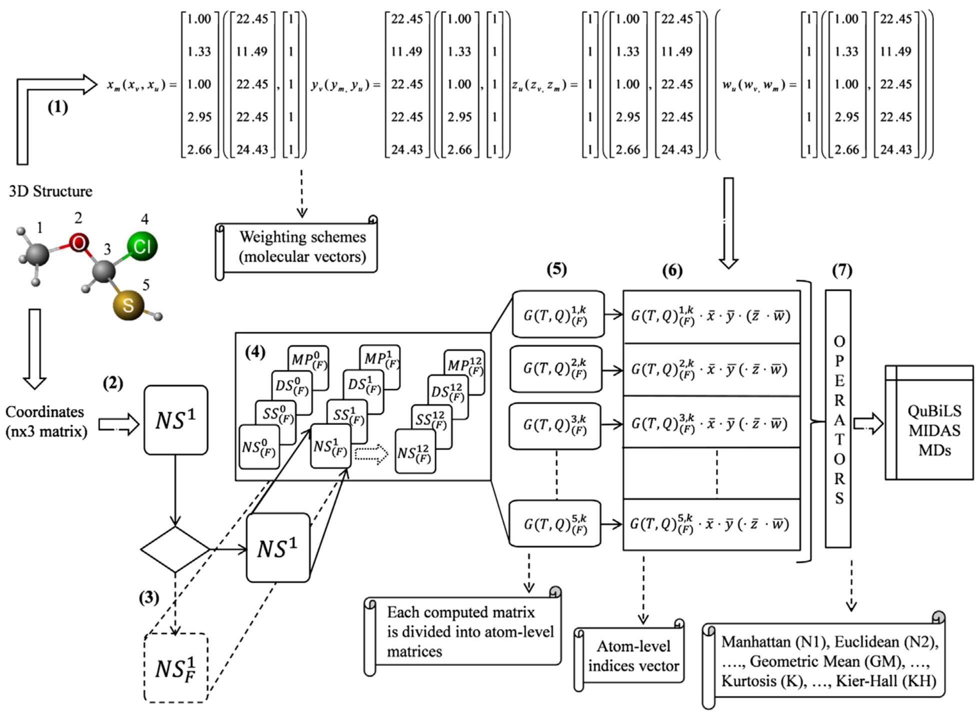
where, '*a*' is a particular atom (a = 1,2,…, n), n is the total number of atoms, is the value corresponding to the contribution of the atom '*a*' in the vector of atom-level indices , F is a local fragment that may or not be taken into consideration during the calculation, and $x^1,\ldots,x^n$, $y^1,\ldots, y^n$, $z^1,\ldots,z^n$ and $w^1,\ldots, w^n$ are the components of the molecular vectors $\bar{x}$, $\bar{y}$, $\bar{z}$ and $\bar{w}$, respectively. The $g^{a,k}_{ij(F)}$, $g^{a,k}_{ijl(F)}$ and $gq^{a,k}_{ijlh(F)}$ coefficients are the values of the kth two-tuple, three-tuple and four-tuple atom-level total (or local-fragment) spatial-(dis)similarity matrices ($\mathbb{G}^{a,k}_{(F)}$, $\mathbb{GT}^{a,k}_{(F)}$ and $\mathbb{GQ}^{a,k}_{(F)}$), which are computed from the respective kth two-tuple, three-tuple and four-tuple total (or local fragment) spatial-(dis)similarity matrices [$\mathbb{G}^{k}_{(F)}$, $\mathbb{GT}^{k}_{(F)}$ and $\mathbb{GQ}^{k}_{(F)}$]. Lastly, the k (±1,±2,…,±12) values constitute the power to which the matrices are raised by using the Hadamard product [1–3].

The total matrices ($\mathbb{G}^{k}$, $\mathbb{GT}^{k}$ and $\mathbb{GQ}^{k}$) are the basis for the computation of the QuBiLS-MIDAS MDs. Specifically, when $k = 1$ (matrix of order 1), the g^1_{ij}, gt^1_{ijl} and gq^1_{ijlh} entries corresponding to the $\mathbb{G}^1$, $\mathbb{GT}^1$ and $\mathbb{GQ}^1$ matrices represent the chemical information codified on relations among '*N*' atoms of a molecule, computed by means of several (dis)-similarity metrics and multi-metrics (see Table 1 for examples of two-tuple and three-tuple total spatial-(dis) similarity matrices). From the total matrix approaches, local-fragment matrices ($\mathbb{G}^{k}_{F}$, $\mathbb{GT}^{k}_{F}$ and $\mathbb{GQ}^{k}_{F}$) may be computed in order to consider atom-types or chemical regions (F) of interest, i.e. hydrogen-bond donors, hydrogen-bond acceptors, carbon atoms in aliphatic chains, halogens, terminal methyl groups, carbon atoms in aromatic portions and heteroatoms (see Table 2 for examples of two-tuple local-fragment spatial-(dis)similarity matrices). Also, in order to obtain normalized matrix representations, simple-stochastic, double-stochastic and mutual probability schemes could be used [2, 3] (see Table 3 for examples of normalized two-tuple spatial-(dis)similarity matrices).

Finally, from the non-stochastic (simple-stochastic, double-stochastic or mutual-probability) total (or local-fragment) matrices ($\mathbb{G}^{k}_{(F)}$, $\mathbb{GT}^{k}_{(F)}$ and $\mathbb{GQ}^{k}_{(F)}$), the corresponding atom-level matrices ($\mathbb{G}^{a,k}_{(F)}$, $\mathbb{GT}^{a,k}_{(F)}$ and $\mathbb{GQ}^{a,k}_{(F)}$) may be computed, whose values are employed in the QuBiLS-MIDAS MDs calculation (see Equations 1–3). Each atom-level matrix determines an atom-level descriptor for atom '*a*' of a molecule, which constitutes an entry of the ${}_{(F)}L$ vector. Once the ${}_{(F)}L$ vector is calculated, then the global kth two-linear, three-linear and four-linear QuBiLS-MIDAS MDs are determined using one or several aggregation operators (see Table 5 in [2]) over the coefficients of ${}_{(F)}L$. To compute these descriptors a multi-core, distributed and fully cross-platform software application was developed [19, 20], which is freely available at http://tomocomd.com/. Schema 1 shows a general flowchart regarding the overall calculation process for the QuBiLS MIDAS KA-MDs, while Schema 2 is a graphical representation of each step to be performed during the computation of a specific total duplex QuBiLS-MIDAS KA-MD.

Novel *N*-tuple topological and geometric cutoffs

The QuBiLS-MIDAS MDs exclusively codify information related with 3D molecular conformations and configurations, taking into account all relations between atoms of a molecule or according to certain atom-types or chemical fragments (F). Therefore, with the purpose of establishing a relation between the topological and geometrical aspects for each group of '*N*' atoms considered, and in this way take into account some short-, middle- and large-relations, two procedures are defined:

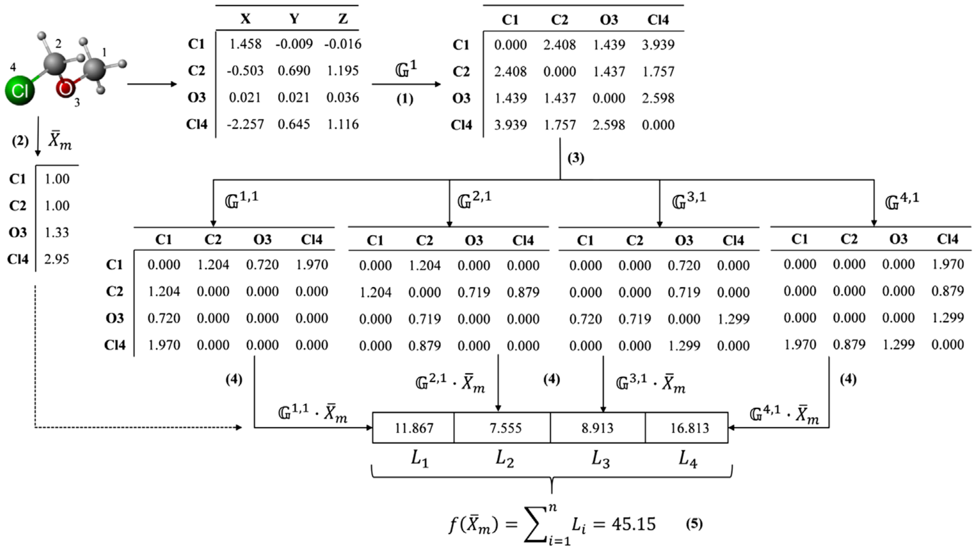


Schema 1. General workflow for calculating the QuBiLS MIDAS molecular descriptors. (1) Computation of the molecular vectors according to the selected atomic properties; (2) computation, from the 3D Cartesian coordinates of each atom, of the non-stochastic two-tuple, three-tuple or four-tuple total spatial-(dis)similarity matrices for $k = 1$; (3) consideration of atom types or local fragments (optional); (4) computation of the simple-stochastic, double-stochastic and mutual probability matrices, as well as determination of the kth matrices through the Hadamard product up to the selected k value; (5) splitting the calculated matrices into atom-level matrices; (6) computation of the atom-level indices (descriptors) using the molecular vectors calculated in step (1); (7) application of the selected aggregation operators over the vector of the atom-level descriptors.

- *N*-tuple graph theoretical cutoff (*P*), known as 'path cutoff', based on the topological distance. These cutoffs are denoted as: *lag P* for $N = 2$, *lag 3P* for $N = 3$ and *lag 4P* for $N = 4$.
- *N*-tuple Euclidean geometric cutoff (*L*), known as 'length cutoff', based on the Euclidean-geometric distance. These cutoffs are denoted as: *lag L* for $N = 2$, *lag 3L* for $N = 3$ and *lag 4L* for $N = 4$.

The application of one or both molecular cutoffs on the $\mathbb{G}^1_{(F)}$, $\mathbb{GT}^1_{(F)}$ and $\mathbb{GQ}^1_{(F)}$ matrices permits the computation of the two-tuple, three-tuple and four-tuple topological and geometric neighbourhood quotient total (or local-fragment) spatial-(dis)similarity matrices, $\mathbb{NQG}^1_{(F)}$, $\mathbb{NQGT}^1_{(F)}$ and $\mathbb{NQGQ}^1_{(F)}$, respectively. The coefficients of these novel approaches are computed by multiplying the $\mathbb{G}^1_{(F)}$, $\mathbb{GT}^1_{(F)}$ and $\mathbb{GQ}^1_{(F)}$ matrices by a ratio obtained as the number of relations between the *N* atoms considered that present a topological and/or Euclidean geometric distance smaller or equal to predefined *p* and/or *l* thresholds. Thus, the ${}^{NQ}g^1_{ij(F)}$, ${}^{NQ}gt^1_{ijl(F)}$ and ${}^{NQ}gq^1_{ijlh(F)}$ entries belonging to the $\mathbb{NQG}^1_{(F)}$, $\mathbb{NQGT}^1_{(F)}$ and $\mathbb{NQGQ}^1_{(F)}$ matrices are mathematically defined as follows:

$$\begin{aligned} {}^{NQ}g^1_{ij(F)} &= g^1_{ij(F)} \text{ if } p_{\min} \le p_{ij} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij} \le l_{\max} \\ &= 0 \text{ otherwise} \end{aligned} \quad (4)$$



2
1
4
Cl
O
3
$\mathbb{G}^1$
(1)
(2)
$\bar{X}_m$
(3)
$\mathbb{G}^{1,1}$
$\mathbb{G}^{2,1}$
$\mathbb{G}^{3,1}$
$\mathbb{G}^{4,1}$
(4)
$\mathbb{G}^{1,1}\cdot\bar{X}_m$
$\mathbb{G}^{2,1}\cdot\bar{X}_m$
(4)
$\mathbb{G}^{3,1}\cdot\bar{X}_m$
(4)
$\mathbb{G}^{4,1}\cdot\bar{X}_m$
L_1
L_2
L_3
L_4
$f(\bar{X}_m)=\sum_{i=1}^{n} L_i = 45.15$
(5)

	X	Y	Z
C1	1.458	-0.009	-0.016
C2	-0.503	0.690	1.195
O3	0.021	0.021	0.036
Cl4	-2.257	0.645	1.116

	C1	C2	O3	Cl4
C1	0.000	2.408	1.439	3.939
C2	2.408	0.000	1.437	1.757
O3	1.439	1.437	0.000	2.598
Cl4	3.939	1.757	2.598	0.000

C1	1.00
C2	1.00
O3	1.33
Cl4	2.95

	C1	C2	O3	Cl4
C1	0.000	1.204	0.720	1.970
C2	1.204	0.000	0.000	0.000
O3	0.720	0.000	0.000	0.000
Cl4	1.970	0.000	0.000	0.000

	C1	C2	O3	Cl4
C1	0.000	1.204	0.000	0.000
C2	1.204	0.000	0.719	0.879
O3	0.000	0.719	0.000	0.000
Cl4	0.000	0.879	0.000	0.000

	C1	C2	O3	Cl4
C1	0.000	0.000	0.720	0.000
C2	0.000	0.000	0.719	0.000
O3	0.720	0.719	0.000	1.299
Cl4	0.000	0.000	1.299	0.000

	C1	C2	O3	Cl4
C1	0.000	0.000	0.000	1.970
C2	0.000	0.000	0.000	0.879
O3	0.000	0.000	0.000	1.299
Cl4	1.970	0.879	1.299	0.000

11.867	7.555	8.913	16.813
L_1	L_2	L_3	L_4

Schema 2. General workflow for the calculation of a specific total duplex QuBiLS-MIDAS KA-MD based on the linear algebraic form, Euclidean metric, non-stochastic matrix approach, atomic mass as property and Manhattan aggregation operator. (1) Computation of the non-stochastic matrix for $k = 1\left(\mathbb{G}^1\right)$ from the 3D coordinate matrix using the Euclidean metric; (2) computation of the molecular vector based on the atomic mass property, $\bar{X}_m$; (3) splitting of the $\mathbb{G}^1$ matrix into n (number of atoms) atom-level matrices, $\mathbb{G}^{a,1}$, where a is an atom of the molecule; (4) Computation of the atom-level descriptors and saving them into the $\bar{L}$ vector; (5) application of the Manhattan aggregation operator over the entries of the $\bar{L}$ vector, this being the value of the molecular descriptor.

$$
\begin{aligned}
{}^{NQ}gt^1_{ijl(F)} &= gt^1_{ijl(F)} \text{ if } p_{\min} \le p_{ij}, p_{jl}, p_{li} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij}, l_{jl}, l_{li} \le l_{\max} \\
&= \tfrac{2}{3} gt^1_{ijl(F)} \begin{cases} \text{if } p_{\min} \le p_{ij}, p_{jl(li)} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij}, l_{jl(li)} \le l_{\max} \\ \text{if } p_{\min} \le p_{jl}, p_{li} \le p_{\max} \text{ and/or } l_{\min} \le l_{jl}, l_{li} \le l_{\max} \end{cases} \\
&= \tfrac{1}{3} gt^1_{ijl(F)} \text{ if } p_{\min} \le p_{ij(jl,li)} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij(jl,li)} \le l_{\max} \\
&= 0 \quad \text{otherwise}
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
{}^{NQ}gq^1_{ijlh(F)} &= gq^1_{ijlh(F)} \text{ if } p_{\min} \le p_{ij}, p_{jl}, p_{lh}, p_{hi} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij}, l_{jl}, l_{lh}, l_{hi} \le l_{\max} \\
&= \frac{3}{4} gq^1_{ijlh(F)} \begin{cases} \text{if } p_{\min} \le p_{ij}, p_{jl(lh)}, p_{lh(hi)} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij}, l_{jl(lh)}, l_{lh(hi)} \le l_{\max} \\ \text{if } p_{\min} \le p_{jl}, p_{lh}, p_{hi} \le p_{\max} \text{ and/or } l_{\min} \le l_{jl}, l_{lh}, l_{hi} \le l_{\max} \end{cases} \\
&= \frac{2}{4} gq^1_{ijlh(F)} \begin{cases} \text{if } p_{\min} \le p_{ij}, p_{jl(lh,hi)} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij}, l_{jl(lh,hi)} \le l_{\max} \\ \text{if } p_{\min} \le p_{jl}, p_{lh(hi)} \le p_{\max} \text{ and/or } l_{\min} \le l_{jl}, l_{lh(hi)} \le l_{\max} \\ \text{if } p_{\min} \le p_{lh}, p_{hi} \le p_{\max} \text{ and/or } l_{\min} \le l_{lh}, l_{hi} \le l_{\max} \end{cases} \\
&= \frac{1}{4} gq^1_{ijlh(F)} \text{ if } p_{\min} \le p_{ij(jl,li,hi)} \le p_{\max} \text{ and/or } l_{\min} \le l_{ij(jl,lh,hi)} \le l_{\max} \\
&= 0 \quad \text{otherwise}
\end{aligned} \tag{6}
$$

where, the g^1_{ij}, gt^1_{ijl}, gq^1_{ijlh} coefficients represent the relations among two, three and four atoms of a molecule and correspond to the total (or local-fragment) matrices $\mathbb{G}^1_{(F)}$, $\mathbb{GT}^1_{(F)}$ and $\mathbb{GQ}^1_{(F)}$,

Table 2. Examples of the two-tuple local-fragment spatial-(dis)similarity matrices. (a) Two-tuple total spatial-(dis)similarity matrix for k = 1, $\mathbb{G}^1$, computed from the 3D coordinates of the chloro(methoxy) methane molecule (see Figure 1); (b) Examples of the two-tuple local fragment spatial-(dis)similarity matrices, $\mathbb{G}^1_F$, obtained with different chemical fragments.

(a) Two-tuple total spatial-(dis)similarity matrices, $\mathbb{G}^1$.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	2.408	1.439	3.939
C2	2.408	0.000	1.438	1.757
O3	1.439	1.438	0.000	2.598
Cl4	3.939	1.757	2.598	0.000

(b) Two-tuple local-fragment spatial-(dis)similarity matrices, $\mathbb{G}^1_F$.

$\mathbb{G}^1_F$ based on halogen fragment.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	1.969
C2	0.000	0.000	0.000	0.878
O3	0.000	0.000	0.000	1.299
Cl4	1.969	0.878	1.299	0.000

$\mathbb{G}^1_F$ based on methyl group fragment

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.204	0.719	1.969
C2	1.204	0.000	0.000	0.000
O3	0.719	0.000	0.000	0.000
Cl4	1.969	0.000	0.000	0.000

$\mathbb{G}^1_F$ based on heteroatom fragment.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.719	1.969
C2	0.000	0.000	0.719	0.878
O3	0.719	0.719	0.000	2.598
Cl4	1.969	0.878	2.598	0.000

respectively. In addition, p_{xy} and l_{xy} represent the topological and Euclidean-geometric distance between two atoms of a molecule, while $[p_{\min}, p_{\max}]$ and $[l_{\min}, l_{\max}]$ constitute the user-defined topological and Euclidean-geometric intervals, respectively. Table 4 illustrates the computation of the two-tuple topological and geometric neighbourhood quotient total spatial-(dis)similarity matrices obtained for the chemical structure of chloro(methoxy)methane applying the graph-theoretical (topological) and Euclidean-geometric cutoffs.

Other molecular cutoff procedures are also proposed in order to consider only the ternary (N = 3) and quaternary (N = 4) relations among atoms of a molecule whose values are consistent with a specific multi-metric. These procedures are denominated as the N-tuple geometric cutoffs based on multi-metrics, and their mathematical definitions are:

$$\begin{aligned} {}^{NQ}gt^1_{ijl(F)} &= gt^1_{ijl(F)} \text{ if} tv_{\min} \leq tv_{ijl} \ \leq tv_{\max} \\ &= 0 \quad \text{ otherwise} \end{aligned} \tag{7}$$

$$\begin{aligned} {}^{NQ}gq^1_{ijlh(F)} &= gq^1_{ijlh(F)} \text{ if } qv_{\min} \leq qv_{ijlh} \ \leq qv_{\max} \\ &= 0 \qquad \text{ otherwise} \end{aligned} \tag{8}$$

where, tv_{ijl} and qv_{ijlh} are the values corresponding to the calculation of a ternary and quaternary multi-metric, respectively (see Table 1 in [3]). In addition, $[tv_{\min}, tv_{\max}]$ and $[qv_{\min}, qv_{\max}]$ are the predefined intervals when cutoffs based on relations among three and four atoms are applied, respectively. Specifically, the ternary multi-metrics that may be used include

Table 3. Example of probabilistic transformations over the non-stochastic two-tuple total spatial-(dis)similarity matrix for $k = 1$, ${}_{ns}\mathbb{G}^1$, computed from 3D coordinates of the chloro(methoxy)methane molecule (see Figure 1) using the Euclidean metric.

Non-stochastic matrix, ${}_{ss}\mathbb{G}^1$

	C1	*C2*	*O3*	*Cl4*
C1	0.000	2.408	1.439	3.939
C2	2.408	0.000	1.438	1.757
O3	1.439	1.438	0.000	2.598
Cl4	3.939	1.757	2.598	0.000

Simple-stochastic matrix, ${}_{ns}\mathbb{G}^1$

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.309	0.185	0.506
C2	0.430	0.000	0.257	0.314
O3	0.263	0.263	0.000	0.475
Cl4	0.475	0.212	0.313	0.000

Double-stochastic matrix, ${}_{ds}\mathbb{G}^1$

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.387	0.246	0.368
C2	0.387	0.000	0.368	0.246
O3	0.246	0.368	0.000	0.387
Cl4	0.368	0.246	0.387	0.000

Mutual probability matrix, ${}_{mp}\mathbb{G}^1$

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.089	0.053	0.145
C2	0.089	0.000	0.053	0.065
O3	0.053	0.053	0.000	0.096
Cl4	0.145	0.065	0.096	0.000

the triangle area (*lag A*), bond angle (*lag BA*) and ternary (or triangle) perimeter (*lag TP*); while the quaternary multi-metrics that may be used include volume (*lag V*), dihedral angle (*lag DA*) and quaternary (or quadrilateral) perimeter (*lag QP*). An example of molecular cutoff based on the bond angle multi-metric is shown in Table 5.

On one hand, it is important to highlight that the molecular cutoffs defined for the same number of atoms could be simultaneously applied, e.g. in a relation among three distinct atoms ($i \neq j \neq l$) if any permutation of three-tuple cutoffs (*lag 3P*, *lag 3L*, *lag A*, *lag BA* and *lag TP*) is used, then all the criteria considered must be fulfilled (see Table 6). On the other hand, the molecular cutoffs for relations among two, three and four atoms can also be concurrently applied on the same matrix representation. Therefore, on four-tuple matrix approaches when four distinct atoms are analysed ($i \neq j \neq l \neq h$) then four-tuple cutoffs can be applied; if three distinct atoms are analysed [$(i = j) \neq l \neq h$] then three-tuple cutoffs can be applied; and if two distinct atoms are analysed [$(i = j = l) \neq h$] then two-tuple cutoffs can be applied. Likewise, this previous strategy is employed on three-tuple matrix approaches when three-tuple cutoffs and two-tuple cutoffs are computed for relations among three ($i \neq j \neq l$) and two [$(i = j) \neq l$] distinct atoms, respectively (see Table 7).

So far, only topological and geometric neighbourhood quotient total (or local-fragment) spatial-(dis)similarity matrices for order 1 ($k = 1$) have been defined ($\mathbb{NQG}^1_{(F)}$, $\mathbb{NQGT}^1_{(F)}$ and $\mathbb{NQGQ}^1_{(F)}$). However, these matrices constitute classes of generalized matrices [17] as well,

Table 4. Examples of the two-tuple topological and geometric neighbourhood quotient total spatial-(dis)similarity matrices computed for the chloro(methoxy)methane molecule (see Figure 1). (a) Euclidean distance between all atoms of the molecule; (b) two-tuple KA spatial-(dis)similarity matrix for $k = 1$ (order) calculated with the angular separation metric; (c) two-tuple topological and geometric neighbourhood quotient matrix computed with topological cutoff (*lag P*) for $P = 1$; (d) two-tuple topological and geometric neighbourhood quotient matrix computed with the Euclidean geometric cutoff (*lag L*) for $L = 2.5$.

(a) Euclidean distance between all atoms.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	2.408	1.439	3.939
C2	2.408	0.000	1.438	1.757
O3	1.439	1.438	0.000	2.598
Cl4	3.939	1.757	2.598	0.000

(b) $\mathbb{G}^1$ *based on angular separation metric.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.354	0.558	1.875
C2	1.354	0.000	0.318	0.237
O3	0.558	0.318	0.000	0.952
Cl4	1.875	0.237	0.952	0.000

(c) $\mathbb{NQG}^1$ *based on topological cutoff.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.558	0.000
C2	0.000	0.000	0.318	0.237
O3	0.558	0.318	0.000	0.000
Cl4	0.000	0.237	0.000	0.000

(d) $\mathbb{NQG}^1$ *based on Euclidean geometric cutoff.*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.354	0.558	0.000
C2	1.354	0.000	0.318	0.237
O3	0.558	0.318	0.000	0.000
Cl4	0.000	0.237	0.000	0.000

where the coefficients for representations of higher orders ($k \geq 2$) are computed through the Hadamard product. Thus, these are the basis for calculating topological and geometric neighbourhood quotient descriptors (QuBiLS-MIDAS NQ-MDs) using Equations 1–3, but employing the $\mathbb{NQG}^k_{(F)}$, $\mathbb{NQGT}^k_{(F)}$ and $\mathbb{NQGQ}^k_{(F)}$ matrices in place of the $\mathbb{G}^k_{(F)}$, $\mathbb{GT}^k_{(F)}$ and $\mathbb{GQ}^k_{(F)}$ matrices, respectively. To conclude, it may be stated that with the incorporation of these cutoffs to the QuBiLS-MIDAS formalism, chemical information regarding particular topological and geometric aspects of the molecules is codified. This constitutes an important advance in the QuBiLS-MIDAS framework, since it is known that some chemical and/biological properties are more dependent on atomic interactions at particular distances, and thus the consideration of specific atomic separations/relations should enhance the modelling capacity of these 3D MDs. These novel molecular cutoffs have been built into the QuBiLS-MIDAS software (http://tomocomd.com/).

Table 5. Example of molecular cutoff based on the bond angle multi-metric. (a) Three-tuple KA matrix for $k = 1, \mathbb{GT}^1$, computed for the chloro(methoxy)methane molecule (see Figure 1) using the bond angle multi-metric; (b) three-tuple topological and geometric neighbourhood quotient matrix for $k = 1$, $\mathbb{NQGT}^1$, applying on $\mathbb{GT}^1$ the three-tuple cutoff based on bond angle multi-metric (*lag BA*) for $BA \leq 1$.

(a) Three-tuple KA spatial-(dis)similarity matrix, $\mathbb{GT}^1$

$\mathbb{GT}^1$ slide 1*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	0.000
C2	0.000	0.000	0.578	2.470
O3	0.000	1.985	0.000	2.682
Cl4	0.000	0.390	0.163	0.000

$\mathbb{GT}^1$ slide 2*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.578	0.281
C2	0.000	0.000	0.000	0.000
O3	1.985	0.000	0.000	0.697
Cl4	0.390	0.000	0.553	0.000

$\mathbb{GT}^1$ slide 3*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.578	0.000	0.297
C2	0.578	0.000	0.000	1.892
O3	0.000	0.000	0.000	0.000
Cl4	0.163	0.553	0.000	0.000

$\mathbb{GT}^1$ slide 4*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.281	0.297	0.000
C2	2.470	0.000	1.892	0.000
O3	2.682	0.697	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

(b) Three-tuple topological and geometric neighbourhood quotient spatial-(dis)similarity matrix, $\mathbb{NQGT}^1$*, corresponding to* $\mathbb{GT}^1$

$\mathbb{NQGT}^1$ slide 1*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	0.000
C2	0.000	0.000	0.578	0.000
O3	0.000	0.000	0.000	0.000
Cl4	0.000	0.390	0.163	0.000

$\mathbb{NQGT}^1$ slide 2*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.578	0.281
C2	0.000	0.000	0.000	0.000
O3	0.000	0.000	0.000	0.697
Cl4	0.390	0.000	0.553	0.000

$\mathbb{NQGT}^1$ slide 3*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.578	0.000	0.297
C2	0.578	0.000	0.000	0.000
O3	0.000	0.000	0.000	0.000
Cl4	0.163	0.553	0.000	0.000

$\mathbb{NQGT}^1$ slide 4*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.281	0.297	0.000
C2	0.000	0.000	0.000	0.000
O3	0.000	0.697	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

Validation of the novel *N*-tuple molecular cutoffs

First, in order to guide the assessment of the molecular cutoffs defined according to their usefulness, an exploration of the most common values of topological/Euclidean-geometric distances, bond/dihedral angles, triangle/quadrilateral perimeters, triangle area and volume in a diverse chemical compound space was performed. To this end, the PrimScreen1 dataset (http://www.otavachemicals.com/download-compound-libraries/cat_view/110-diversity-sets/128-primscreen-1) comprising 985 structurally diverse compounds was employed. Figure 2 shows the frequency of the different molecular cutoffs obtained for the PrimScreen1 dataset. As may be observed, the most frequent values for the aforementioned criteria correspond to the following ranges: topological and Euclidean-geometric distances [1, 10], triangle perimeter [9, 24], triangle area [0,15], quadrilateral perimeter [15, 30] and volume [0,12]. In this analysis, all previous ranges were split in size one intervals, while 10 equal intervals were considered for the bond- and dihedral-angle multi-metrics using the range [0,*π*] (in radians). The most frequent intervals in each case were used to evaluate the performance of the proposed molecular cutoffs by using variability analysis, principal component analysis and QSAR studies.

Variability analysis of the N-tuple topological and geometric molecular cutoffs

The variability analysis method is based on the Shannon's entropy (SE) parameter [21] and quantifies the information content codified by variables [22]. In this way, suitable MDs (variables) for cheminformatics studies can be identified under the principle that high-entropy values correspond to those MDs with good capacity to discriminate among structurally different compounds, while low-entropy values are indicative of redundant MDs [23]. This unsupervised procedure was performed using IMMAN software [24], which is freely available at http://mobiosd-hub.com/imman-soft/. For this study, the PrimScreen3 dataset (http://www.otavachemicals.com/download-compound-libraries/cat_view/110-diversity-sets/130-primscreen-3) comprising 2879 molecules was used to compute the QuBiLS-MIDAS NQ-MDs. The binning scheme adopted is equal to the number of molecules in the dataset, and thus the highest entropy value is 11.49 bits ($\log_2^{2879}$). Figure 3 shows the Shannon's entropy distributions corresponding to the molecular cutoffs defined. In all cases 30 NQ-MDs (see Supplementary Information SI1 for the projects configuration) were computed for each cutoff procedure, and a comparison with respect to total QuBiLS-MIDAS KA-MDs (all relations are considered) was carried out.

First, the contribution of the strategy to consider specific atom-pair relations was assessed. As can be observed in Figure 3(a), if the distributions above 9 bits (78% of the highest entropy) are analysed, it may be stated that the NQ-MDs based on two-tuple graph theoretical cutoffs (*lag P*) for *P* = 1; 2; 3; 4; 5; 6 present the best behaviour, with at least 50% of the variables with entropy values superior to the threshold considered. However, it is important to remark that the NQ-MDs calculated for *P* = 5; 6 show a better distribution pattern than the duplex QuBiLS-MIDAS KA-MDs, while the other distributions aforementioned have a comparable behaviour with respect to the latter. On the other hand, Figure 3(b) shows the entropy distributions corresponding to duplex QuBiLS-MIDAS NQ-MDs when computed using the two-tuple Euclidean-geometric cutoffs (*lag L*). In this sense, it can be detailed that the best distributions belong to the intervals *L* = [2, 3]; [3, 4]; [5, 6] due to the fact that they

Table 6. Example of computation of a three-tuple neighbourhood quotient matrix simultaneously applying two ternary multi-metrics. The compound used for all calculations is chloro(methoxy)methane (see Figure 1). (a) Three-tuple KA matrix for $k = 1$ computed using bond angle multi-metric; (b) three-tuple KA matrix for $k = 1$ computed using triangle area multi-metric with Euclidean distance; (c) three-tuple topological and geometric neighbourhood quotient matrix for $k = 1, \mathbb{NQGT}^1$, applying on $\mathbb{GT}^1$ of (a) the three-tuple cutoffs based on bond angle (*lag BA*, $BA \leq 1$ and triangle area (*lag A*, $A \leq 1$ multi-metrics.

(a) Three-tuple KA spatial-(dis)similarity matrix, $\mathbb{GT}^1$, computed with bond angle multi-metric.

$\mathbb{GT}^1$ slide 1*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	0.000
C2	0.000	0.000	0.578	2.470
O3	0.000	1.985	0.000	2.682
Cl4	0.000	0.390	0.163	0.000

$\mathbb{GT}^1$ slide 2*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.578	0.281
C2	0.000	0.000	0.000	0.000
O3	1.985	0.000	0.000	0.697
Cl4	0.390	0.000	0.553	0.000

$\mathbb{GT}^1$ slide 3*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.578	0.000	0.297
C2	0.578	0.000	0.000	1.892
O3	0.000	0.000	0.000	0.000
Cl4	0.163	0.553	0.000	0.000

$\mathbb{GT}^1$ slide 4*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.281	0.297	0.000
C2	2.470	0.000	1.892	0.000
O3	2.682	0.697	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

(b) Three-tuple KA spatial-(dis)similarity matrix, $\mathbb{GT}^1$, computed with triangle area multi-metric and Euclidean distance.

$\mathbb{GT}^1$ slide 1*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	0.000
C2	0.000	0.000	0.947	1.316
O3	0.000	0.947	0.000	0.829
Cl4	0.000	1.316	0.829	0.000

$\mathbb{GT}^1$ slide 2*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.947	1.316
C2	0.000	0.000	0.000	0.000
O3	0.947	0.000	0.000	1.198
Cl4	1.316	0.000	1.198	0.000

$\mathbb{GT}^1$ slide 3*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.947	0.000	0.829
C2	0.947	0.000	0.000	1.198
O3	0.000	0.000	0.000	0.000
Cl4	0.829	1.198	0.000	0.000

(*Continued*)

Table 6. (*Continued*)

$\mathbb{GT}^1$ slide 4*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	1.316	0.829	0.000
C2	1.316	0.000	1.198	0.000
O3	0.829	1.198	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

(c) Three-tuple topological and geometric neighbourhood quotient spatial-(dis)similarity matrix, $\mathbb{NQGT}^1$*, corresponding to* $\mathbb{GT}^1$ *of (a) applying cutoffs based on bond angle and triangle area multi-metrics.*

$\mathbb{NQGT}^1$ slide 1*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.000	0.000
C2	0.000	0.000	0.578	0.000
O3	0.000	0.000	0.000	0.000
Cl4	0.000	0.000	0.163	0.000

$\mathbb{NQGT}^1$ slide 2*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.578	0.000
C2	0.000	0.000	0.000	0.000
O3	0.000	0.000	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

$\mathbb{NQGT}^1$ slide 3*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.578	0.000	0.297
C2	0.578	0.000	0.000	0.000
O3	0.000	0.000	0.000	0.000
Cl4	0.163	0.000	0.000	0.000

$\mathbb{NQGT}^1$ slide 4*ij*

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.281	0.297	0.000
C2	0.000	0.000	0.000	0.000
O3	0.000	0.000	0.000	0.000
Cl4	0.000	0.000	0.000	0.000

are the only patterns with values superior to 10 bits (86% of the highest entropy). In addition, from 9 bits, these generally present a comparable-to-superior behaviour with respect to the duplex QuBiLS-MIDAS KA-MDs, except for a few variables.

Secondly, the information content codified by ternary QuBiLS-MIDAS NQ-MDs using the three-tuple molecular cutoffs was assessed. Figure 3(c) shows the SE trend achieved using three-tuple geometric cutoffs based on the bond angle multi-metric (*lag BA*). As can be observed, if the 10 best overall variables according to their entropy values are considered, it can be said that the ternary QuBiLS-MIDAS KA-MDs possess a better performance with respect to the distributions based on *lag BA*, which have a comparable behaviour. However, the patterns computed with *BA* = [0.0,0.314]; [0.314,0.628]; [1.570,1.884] can be highlighted, due to the fact that they have at least 15 NQ-MDs with better variability than the respective KA-MDs. On the other hand, Figure 3(d) shows the entropy distributions related with the three-tuple geometric cutoffs based on the triangle area multi-metric (*lag A*). As can be appreciated, the

Table 7. Example of computation of a three-tuple neighbourhood quotient matrix applying ternary multi-metrics and topological and Euclidean geometric distances. The compound used for all calculations is chloro(methoxy)methane (see Figure 1). (a) Three-tuple KA matrix for $k = 1$ computed using complete triangle area multi-metric with Euclidean distance. When the analyzed atoms ($i \neq j \neq l$) are similar then the Euclidean distance is computed [$(i = j) \neq l$] using this metric (see [3]); (b) three-tuple topological and geometric neighbourhood quotient matrix for $k = 1, \mathbb{NQGT}^1$, simultaneously applying the three-tuple cutoffs based on bond angle (*lag BA*, $BA \leq 1$) and triangle area (*lag A*, $A \leq 1$) multi-metrics, as well as the two-tuple cutoffs based on topological (*lag p*, $p = 1$) and Euclidean geometric (*lag l*, $l \leq 2.5$) distance.

(a) Three-tuple KA spatial-(dis)similarity matrix, $\mathbb{GT}^1$, computed with complete triangle area multi-metric using Euclidean distance.

$\mathbb{GT}^1$ slide 1*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	2.408	1.439	3.939
C2	2.408	2.408	0.947	1.316
O3	1.439	0.947	1.439	0.829
Cl4	3.939	1.316	0.829	3.939

$\mathbb{GT}^1$ slide 2*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	2.408	2.408	0.947	1.316
C2	2.408	0.000	1.438	1.757
O3	0.947	1.438	1.438	1.198
Cl4	1.316	1.757	1.198	1.757

$\mathbb{GT}^1$ slide 3*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	1.439	0.947	1.439	0.829
C2	0.947	1.438	1.438	1.198
O3	1.439	1.438	0.000	2.598
Cl4	0.829	1.198	2.598	2.598

$\mathbb{GT}^1$ slide 4*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	3.939	1.316	0.829	3.939
C2	1.316	1.757	1.198	1.757
O3	0.829	1.198	2.598	2.598
Cl4	3.939	1.757	2.598	0.000

(b) Three-tuple topological and geometric neighbourhood quotient spatial-(dis)similarity matrix, $\mathbb{NQGT}^1$, applying three-tuple cutoffs based on bond angle and triangle area multi-metrics and two-tuple topological and Euclidean-geometric cutoffs.

$\mathbb{NQGT}^1$ slide 1*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	1.439	0.000
C2	0.000	0.000	0.947	0.000
O3	1.439	0.000	1.439	0.000
Cl4	0.000	0.000	0.829	0.000

$\mathbb{NQGT}^1$ slide 2*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.947	0.000
C2	0.000	0.000	1.438	1.757
O3	0.000	1.438	1.438	0.000
Cl4	0.000	1.757	0.000	1.757

(*Continued*)

Table 7. (*Continued*)

$\mathbb{NQGT}^1$slide 3*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	1.439	0.947	1.439	0.829
C2	0.947	1.438	1.438	0.000
O3	1.439	1.438	0.000	0.000
Cl4	0.829	0.000	0.000	0.000

$\mathbb{NQGT}^1$slide 4*ij*.

	C1	*C2*	*O3*	*Cl4*
C1	0.000	0.000	0.829	0.000
C2	0.000	1.757	0.000	1.757
O3	0.000	0.000	0.000	0.000
Cl4	0.000	1.757	0.000	0.000

best distributions correspond to the cutoffs *A* = [4, 5]; [5, 6]; [6, 7]; [7, 8]; [8, 9], which show comparable-to-superior behaviour with regard to the ternary QuBiLS-MIDAS KA-MDs, while the other entropy patterns based on *lag A* have an inferior performance relative to the latter. Lastly, Figures 3(e)–(g) represent the results achieved by applying the three-tuple geometric cutoffs based on the triangle perimeter multi-metric (*lag TP*), three-tuple graph-theoretical cutoffs (*lag 3P*) and three-tuple Euclidean-geometric cutoffs (*lag 3L*), respectively. In these cases it can be appreciated that, in a general sense, the ternary QuBiLS-MIDAS KA-MDs have the best performance and thus NQ-MDs with better variability may not be obtained with these cutoffs.

Finally, an analysis of the contribution of the four-tuple molecular cutoffs was performed. Figure 3(h) shows the distributions corresponding to the four-tuple geometric cutoffs based on the dihedral angle multi-metric (*lag DA*), where it can be noted that all NQ-MDs based on this cutoff present a superior behaviour with respect to the quaternary QuBiLS-MIDAS KA-MDs. These patterns, based on *lag DA*, have a comparable performance amongst themselves, with the exception of those related to the NQ-MDs computed with *DA* = [2.826,3.140], which show the worst behaviour. According to the four-tuple geometric cutoffs based on the quadrilateral perimeter multi-metric (*lag QP*), whose entropy distributions are shown in Figure 3(i), it can be inferred that NQ-MDs with good variability are obtained with this cutoff, and in all cases better performances with respect to the quaternary QuBiLS-MIDAS KA-MDs are achieved. Nonetheless, it is important to remark that the distributions with the best behaviour correspond to *QP* = [19, 20]; [20, 21]; [21, 22]; [22, 23]; [23, 24]. On the other hand, Figure 3(j) represents the performance of the four-tuple geometric cutoffs based on the volume multi-metric (*lag V*). This analysis reveals that the distributions for *V* = [0,1]; [1, 2]; [2, 3]; [3, 4]; [4, 5]; [5, 6]; [6, 7] yield the best overall performance, while the entropy patterns based on the other *lag V* values present an inferior behaviour with respect to the quaternary QuBiLS-MIDAS KA-MDs. To conclude this variability study, the performance of the four-tuple graph-theoretical cutoffs (*lag 4P*) and four-tuple Euclidean-geometric cutoffs (*lag 4L*) were analysed. In this sense, Figures 3(k) and (l) show that the SE distributions for 4*P* = 1; 2; 3; 4; 5;6; 7 and 4*L* = [1, 2]; [2, 3]; [3, 4]; [4, 5]; [5, 6]; [6, 7]; [7, 8] achieve better behaviour than the corresponding quaternary QuBiLS-MIDAS KA-MDs, respectively.

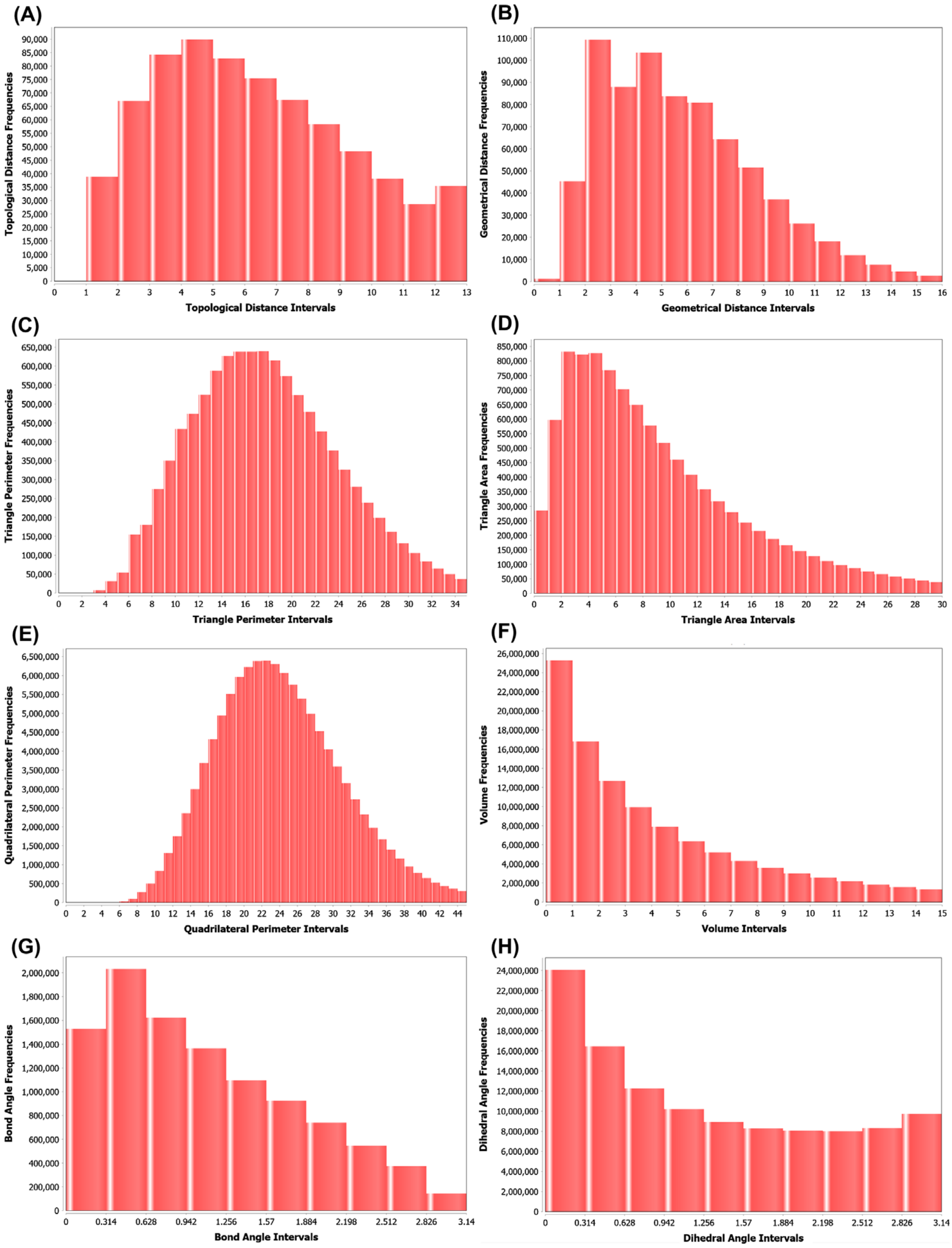


Figure 2. Histograms corresponding to the values calculated on the PrimScreen1 dataset, with the metrics and multi-metrics considered to compute molecular cutoffs.

It is important to highlight that when novel molecular parameters are defined, a desirable attribute is that these codify orthogonal structural information [23]. However, the variability analysis method based on Shannon's entropy does not provide information on the existing redundancy or correlation among the variables. Therefore, the next section is devoted to analysing the possible linear independence of the molecular cutoffs proposed.

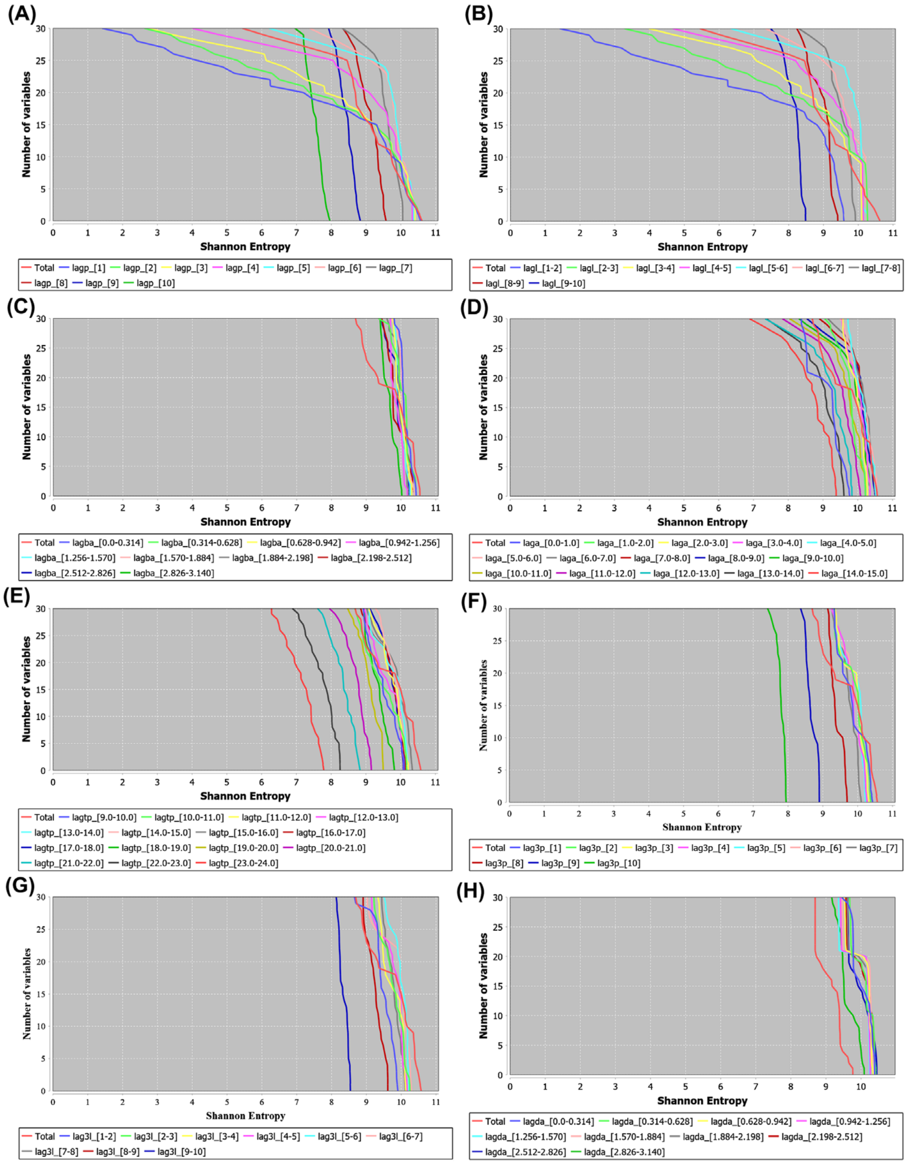


Figure 3. Shannon's entropy distribution of the QuBiLS-MIDAS NQ-MDs. In all figures, total QuBiLS-MIDAS KA-MDs are graphically represented. (a) Two-tuple graph-theoretical cutoffs; (b) two-tuple Euclidean geometric cutoffs; (c) three-tuple geometric cutoffs based on the bond angle multi-metric; (d) three-tuple geometric cutoffs based on the triangle area multi-metric; (e) three-tuple geometric cutoffs based on the triangle perimeter multi-metric; (f) three-tuple graph-theoretical cutoffs; (g) three-tuple Euclidean geometric cutoffs; (h) four-tuple geometric cutoffs based on the dihedral angle multi-metric; (i) four-tuple geometric cutoffs based on the quadrilateral perimeter multi-metric; (j) four-tuple geometric cutoffs based on the volume multi-metric; (k) four-tuple graph-theoretical cutoffs; (l) four-tuple Euclidean geometric cutoffs.

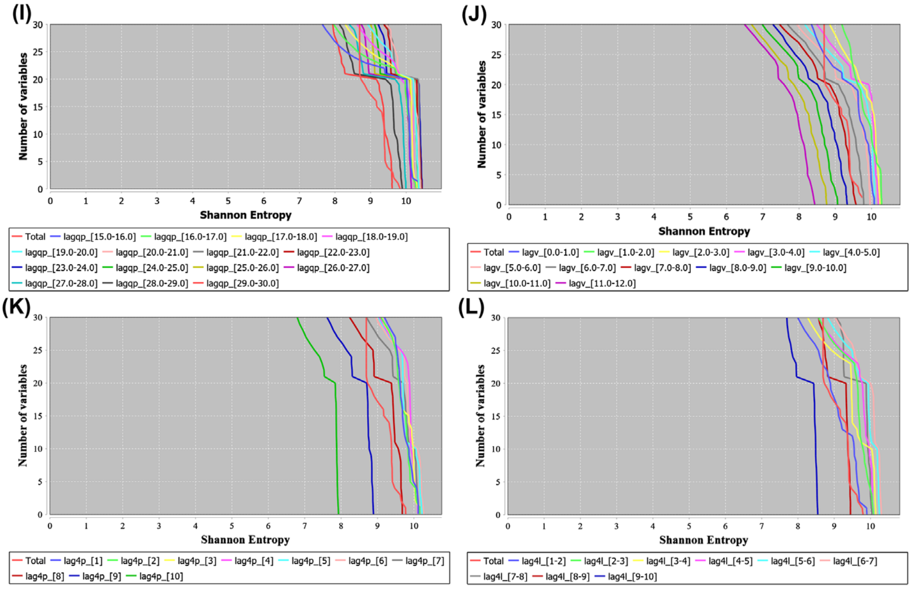


Figure 3. (*Continued*)

Orthogonality analysis of the N-tuple topological and geometric molecular cutoffs

The linear independence analysis of the novel molecular cutoffs was performed using the principal component analysis method [25]. This procedure is used to reduce features in high-dimensionality datasets, by computing orthogonal projections (principal components) that codify the highest variance possible in the original data matrix. In this method the orthogonal variables are loaded into different projections, while correlated variables are loaded into the same component. For all studies, the PrimScreen3 dataset and the 30 NQ-MDs considered for each cutoff in the variability analysis were employed (see Supplementary Information SI1 for the projects configuration). In this case, the MDs calculated for the same number of atoms were joined in a single dataset in order to perform a better analysis according to the cutoff types. A comparison with respect to total QuBiLS-MIDAS KA-MDs was also performed.

First, a study to evaluate the possible linear independence of the QuBiLS-MIDAS NQ-MDs according to the two-tuple topological and Euclidean-geometric cutoffs was performed. The eigenvalues and percentages of the explained variance by the five first principal components computed in the analysis, which approximately explain 82.91% of the cumulative variance, are available as Supplementary Information SI2. In this study, it is observed that the duplex QuBiLS-MIDAS KA-MDs are loaded in factor 1 (49.22%) and factor 5 (3%), with this last factor being exclusive for these indices. Also, factor 1 shows robust loadings for several NQ-MDs computed from the two-tuple topological (*lag P*) and Euclidean-geometric (*lag L*) cutoffs. Nonetheless, NQ-MDs calculated with *lag P* for $P = 1; 2; 3; 4$ and *lag L* for $L = [1, 2]; [2, 3]; [3, 4]; [4, 5]$ are strongly loaded in factor 2 (13.18%). Likewise, factor 3 (11.78%) exhibits robust loadings for NQ-MDs computed with *lag P* for $P = 5; 6$ and *lag L* for $L = [5, 6]; [6, 7]$, as well as for some NQ-MDs calculated with *lag P* for $P = 4$ and *lag L* for $L = [4, 5]$. Finally, NQ-MDs determined with *lag P* for $P = 7; 8; 9$ and *lag L* for $L = [8, 9]; [9, 10]$ present strong loadings in factor 4 (5.72%). Thus, it could be inferred that the introduction of two-tuple topological and

Euclidean-geometric cutoffs (NQ-MDs) permits the codification of orthogonal structural information with respect to that captured by the duplex QuBiLS-MIDAS KA-MDs.

In the second experiment, an analysis of the information codified by three-tuple molecular cutoffs was performed (see Supplementary Information SI3 for the eigenvalues and the percentages of the explained variance by the five first principal components obtained, which

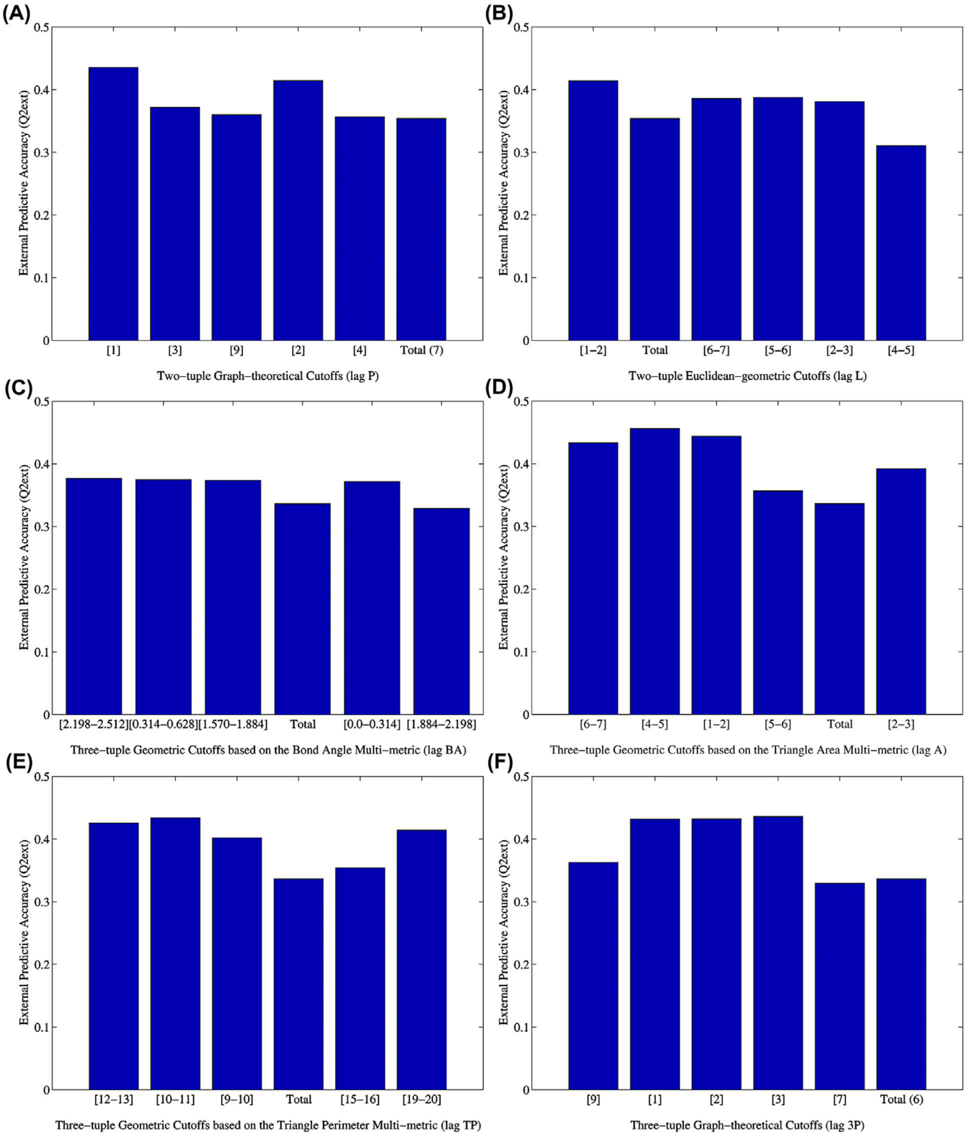


Figure 4. (a) Two-tuple graph-theoretical cutoffs; (b) two-tuple Euclidean geometric cutoffs; (c) three-tuple geometric cutoffs based on the bond angle multi-metric; (d) three-tuple geometric cutoffs based on the triangle area multi-metric; (e) three-tuple geometric cutoffs based on the triangle perimeter multi-metric; (f) three-tuple graph-theoretical cutoffs; (g) three-tuple Euclidean geometric cutoffs; (h) four-tuple geometric cutoffs based on the dihedral angle multi-metric; (i) four-tuple geometric cutoffs based on the quadrilateral perimeter multi-metric; (j) four-tuple geometric cutoffs based on the volume multi-metric; (k) four-tuple graph-theoretical cutoffs; (l) four-tuple Euclidean geometric cutoffs.

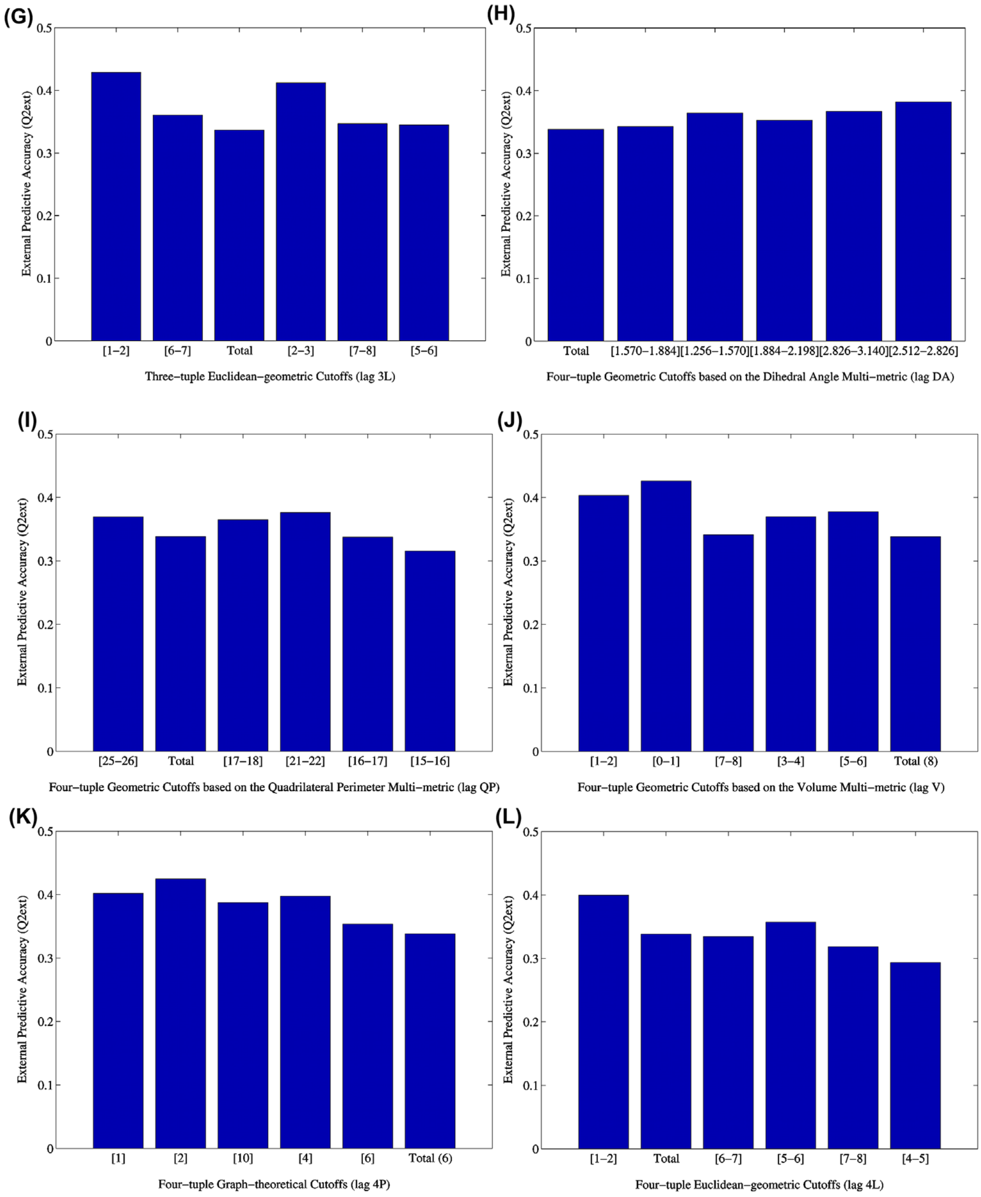


Figure 4. (*Continued*)

collectively explain 83.07% of the cumulative variance). As can be observed, the ternary QuBiLS-MIDAS KA-MDs are loaded in factor 1 (53.27%), and present unique loadings in factor 5 (3.52%). Also, factor 1 presents strong loadings for the majority of the NQ-MDs computed with the three-tuple cutoffs proposed, but, some exceptions may be highlighted. Specifically, factor 2 (15.26%) exhibits robust loadings for NQ-MDs determined with the three-tuple geometric cutoffs based on the bond angle multi-metric (*lag BA*) for all *BA* values, with this factor being unique for this type of indices. Factor 3 (5.77%) shows strong loadings for NQ-MDs based on the three-tuple geometric cutoffs based on the triangle area multi-metric (*lag A*) for *A* = [3, 4]; [4, 5]; [5, 6], as well as for NQ-MDs computed from the three-tuple geometric cutoffs based on the triangle perimeter multi-metric (*lag TP*) for TP = [16, 17]; [17, 18]; [18, 19]; [19, 20]. Lastly, factor 4 (5.23%) is exclusive for NQ-MDs calculated with *lag A* for *A* = [0,1]. Therefore, it can be stated that the introduction of three-tuple molecular cutoffs (NQ-MDs) allows the codification of linearly independent information relative to the ternary QuBiLS-MIDAS KA-MDs.

Finally, the orthogonality of the information captured by the four-tuple molecular cutoffs was examined. The eigenvalues and the percentages of the explained variance by the four first principal components calculated in the analysis, which collectively explain 89.62% of the cumulative variance, are available in Supplementary Information SI4. Factor 1 (62.84%) presents robust loadings for all quaternary QuBiLS-MIDAS KA-MDs, as well as for all NQ-MDs computed from the non-stochastic (NS) matrix and the four-tuple geometric cutoffs based on the dihedral angle (*lag DA*, for all *DA* values), quadrilateral perimeter (*lag QP*, for all *QP* values) and volume (*lag V*, for all *V* values) multi-metrics, respectively. On the other hand, factor 2 (13.31%) shows strong loadings for those NQ-MDs calculated with matrices distinct to the NS approach, *lag DA* for all *DA* values and *lag V* for *V* = [1, 2]; [2, 3]; [3, 4]; [4, 5]; [5, 6]. Lastly, the NQ-MDs were not computed with the NS matrix and based on *lag QP* are strongly loaded in factor 3 (7.96%) and factor 4 (5.50%). Thus, it can be concluded that orthogonal information with respect to quaternary QuBiLS-MIDAS KA-MDs is codified by using four-tuple molecular cutoffs.

These results justify the theoretical and practical contribution of the novel *N*-tuple molecular cutoffs, and thus complement QuBiLS-MIDAS KA-MDs theory as previously reported in the literature [1–3]. Accordingly, QSAR models with greater predictive power with respect to QuBiLS-MIDAS KA-MDs could be developed. However, it is important to note that in practice it may be advisable to combine both approaches (i.e. use KA- and NQ-MDs) for greater modelling power.

Predictive ability analysis of the N-tuple topological and geometric molecular cutoffs

In order to assess the correlation ability of the novel *N*-tuple molecular cutoffs when used to compute QuBiLS-MIDAS NQ-MDs, QSAR studies over four well-known chemical datasets were performed. These datasets are composed of steroids (STER), angiotensin converting enzyme (ACE) inhibitors, thermolysin inhibitors (THER) and thrombin inhibitors (THR). The former was proposed by Cramer et.al. when CoMFA methodology was introduced [26], while the others were employed by Sutherland et al. in a comparative study of QSAR methods commonly used in chemo-informatics analysis [8]. Currently, these are utilized as benchmarks to compare results obtained in several approaches [9–13]. In this study, the 3D coordinates of the molecular structures were generated using the CORINA software, and the datasets were partitioned into training and test sets as previously used in the literature. These chemical datasets were selected due to the fact that they had previously been employed to evaluate the predictive ability of the QuBiLS-MIDAS KA-MDs through a comparative and statistical analysis [9], demonstrating the superior behaviour of the latter according to other QSAR procedures reported. Thus, this study is only aimed at assessing the performance of the QuBiLS-MIDAS NQ-MDs with respect to the QuBiLS-MIDAS KA-MDs and, in this way, to appraise the feasibility of using the novel molecular cutoffs (NQ-MDs) in the codification of chemical structures.

QuBiLS-MIDAS KA- and NQ-MDs (see Supplementary Information SI1 for the projects configuration) were computed for each of the considered chemical datasets and used to build QSAR models employing the multiple linear regression (MLR) statistical technique and the genetic algorithm (GA) meta-heuristic as the feature selection strategy (search method). To carry out the search process the leave-one-out cross-validation (Q^2_{loo}) was used as the fitness function, and the GA procedure was configured as follows: (1) number of iterations equal to 50000; (2) reproduction/mutation trade-off equal to 0; (3) selection bias equal to 0 (indicates random selection). In this way, from each population, one to four variable regression models were built for the respective biological activity, and the best model for each dimension according to the coefficients' leave-one-out cross-validation (Q^2_{loo}), bootstrapping validation (Q^2_{boot}) and external validation (Q^2_{ext}) was retained. These model-building procedures and validation parameters are implemented in the MobyDigs software (version 1.0) [27], which was employed in this study.

With the aim of comparing the performance of the different molecular cutoffs in QSAR modelling, the statistical parameters obtained for each model were used to build data matrices X_{nxk}, where the ***n*** rows denote the Q^2_{loo}, Q^2_{boot} and Q^2_{ext} coefficients calculated for the predictive models built on each chemical dataset, and the ***k*** columns constitute the considered intervals (see Supplementary Information SI5). In this way, the ranks $\left(r_{ij}\right)$ within each row may be computed. In the case of tied scores, the average rank without including the tied scores is assigned. Lastly, the average of each cutoff k_j is calculated as:

$$ave_{k_j} = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$$

which indicates the average performance in QSAR modelling according to the three coefficients taken into account. This previous calculation constitutes the first step of the Friedman test [28], which was applied in this study using KEEL software [29]. Its purpose was not to determine global statistical differences, but to compute the average rank of the molecular cutoffs analysed.

Therefore, from the results obtained with the previous analysis (for details see Supplementary Information SI6 and to SI17), it can be stated that the novel QuBiLS-MIDAS NQ-MDs have a better average performance in QSAR modelling with respect to the QuBiLS-MIDAS KA-MDs. Specifically, NQ-MDs based on two-tuple graph-theoretical cutoffs (*lag P*) for *P* = 1,3,9,2,4,5 and two-tuple Euclidean-geometric cutoff (*lag L*) for *L* = [1, 2] yield models with greater performance than the duplex QuBiLS-MIDAS KA-MDs. On the other hand, NQ-MDs for relations among three atoms and calculated with the three-tuple geometric cutoffs based on the bond angle multi-metric (*lag BA*) for *BA* = [2.198,2.512]; [0.314,0.628]; [1.570,1.884], three-tuple geometric cutoffs based on the triangle area multi-metric (*lag A*) for A = [6, 7]; [4, 5]; [1, 2]; [5, 6], and three-tuple geometric cutoffs based on the triangle perimeter multi-metric (*lag TP*) for *TP* = [12, 13]; [10, 11]; [9, 10] exhibit better performance in QSAR modeling than the respective ternary QuBiLS-MIDAS KA-MDs. Likewise, this behaviour is presented by the three-tuple graph-theoretical cutoffs (*lag 3P*) for 3*P* = 9; 1; 2; 3; 7 and three-tuple Euclidean-geometric cutoffs (*lag 3L*) for 3L = [1, 2];[6, 7]. Also, superior performance is evident in the NQ-MDs calculated with the four-tuple geometric cutoffs based on the quadrilateral perimeter (*lag QP*) and volume (*lag V*) multi-metrics for *QP* = [25, 26] and *V* = [1, 2]; [0,1]; [7, 8]; [3, 4]; [5, 6]; [2, 3]; [10, 11], respectively, with respect to the quaternary

QuBiLS-MIDAS KA-MDs. In addition, this superior performance is shown for those NQ-MDs computed from the four-tuple graph-theoretical (*lag 4P*) and Euclidean-geometric (*lag 4L*) cutoffs for 4*P* = 1; 2; 10; 4; 6 and 4*L* = [1, 2], respectively.

Finally, an analysis regarding the Q^2_{ext} coefficient of the best five molecular cutoffs according to their overall average performance (i.e. considering Q^2_{loo}, Q^2_{boot}, Q^2_{ext}) on the four chemical compound datasets was carried out. In this study, the results obtained with the QuBiLS-MIDAS KA-MDs are also included. Figure 4 shows the average Q^2_{ext} corresponding to the cutoffs considered. As can be observed in Figure 4(a), the two-tuple graph-theoretical cutoffs (*lag P*) for *P* = 1,2 yield superior external predictive abilities (Q^2_{ext}) to the duplex QuBiLS-MIDAS KA-MDs, while for *P* = 3,4,9 the Q^2_{ext} attained is comparable to the latter. In the case of the two-tuple Euclidean-geometric cutoffs (*lag L*) (see Figure 4(b)), it can be appreciated that for *L* = [1, 2]; [6, 7]; [5, 6]; [2, 3] the NQ-MDs yield QSAR models with a better (Q^2_{ext}) than the corresponding KA-MDs, despite the fact that these last possess the second-best overall average behaviour.

Figure 4(c) shows the performance of the QSAR models for the three-tuple geometric cutoffs based on the bond angle multi-metric (*lag BA*), where it can be noted that for *BA* = [2.198,2.512];0.314,0.628];[1.570,1.884];[0.0,0.314] the results obtained by external prediction are superior to those attained by the KA-MDs. On the other hand, Figure 4(d) and (e) represent the behaviour of the three parameters considered when QSAR models are developed with the three-tuple geometric cutoffs based on the triangle area (*lag A*) and triangle perimeter (*lag TP*) multi-metrics, respectively. In this sense, it can be stated that the indices calculated with *lag A* for *A* = [6, 7]; [4, 5]; [1, 2]; [2, 3] and *lag TP* for *TP* = [12, 13]; [10, 11]; [9, 10]; [19, 20] allow the building of better predictive models than the respective KA-MDs. In the case of the NQ-MDs computed from the three-tuple graph theoretical (*lag 3P*) and Euclidean geometric (*lag 3L*) cutoffs, whose results are shown in Figures 4(f) and (g), it can be stated that for 3*P* = 1;2;3 and 3*L* = [1, 2]; [2, 3] the performance obtained by external prediction (Q^2_{ext}) is superior with respect to those achieved by the ternary QuBiLS-MIDAS KA-MDs, respectively.

Lastly, the behaviour of the four-tuple geometric cutoffs was assessed. According to the results represented in Figure 4(h), it can be said that, although the quaternary QuBiLS-MIDAS KA-MDs show the best average performance (see Supplementary Information SI11), the four-tuple geometric cutoffs based on the dihedral angle multi-metric (*lag DA*) present a higher external predictive accuracy (Q^2_{ext}) for the intervals *DA* = [1.256,1.570];[2.826,3.140];[2.512,2.826]. Also, with regard to the outcomes shown in Figure 4(i), it can be concluded that the NQ-MDs computed with the four-tuple geometric cutoffs based on the quadrilateral perimeter multi-metric (*lag QP*) for QP = [25, 26]; [17, 18]; [21, 22] produce QSAR models with superior performance (Q^2_{ext}) than the respective KA-MDs. Figure 4(j) indicates that the four-tuple geometric cutoffs based on the volume multi-metric (*lag V*) for *V* = [1, 2]; [0,2]; [3, 4]; [5, 6] yield QSAR models with superior predictive ability (Q^2_{ext}) if are compared to those based on the quaternary QuBiLS-MIDAS KA-MDs. Figures 4(k) and (l) show the performance of the NQ-MDs computed from the four-tuple graph-theoretical (*lag 4P*) and Euclidean-geometric (*lag 4L*) cutoffs, respectively. In this sense, it can be said that the NQ-MDs determined with *lag 4P* for 4*P* 1;2;10;4 and *lag 4L* for 4L = [1, 2]; [5, 6] yield QSAR models with better external prediction than the respective KA-MDs.

As a conclusion of this study, it can be suggested that the novel *N*-tuple topological and geometric cutoffs are suitable for extracting structural information, and thus contribute to the development of QSAR models with better predictive ability than when QuBiLS-MIDAS KA-MDs are employed alone.

Conclusions

Novel *N*-tuple topological/geometric cutoffs to characterize molecular structures are introduced. In this way, the *N*-tuple graph-theoretical cut-off and *N*-tuple Euclidean-geometric cut-off are defined in order to consider short, middle and large-relations according to the shortest path and Euclidean distance between atom-pairs ($N = 2$) of a molecule, respectively. Also, these two procedures are generalized to consider relations among three ($N = 3$) and four ($N = 4$) atoms, which constitutes an important contribution to the use of molecular cutoffs, due to the fact that, up to this time, procedures that relate to only two atoms have been defined in the literature. In this sense, by the use of (dis-)similarity multi-metrics, other molecular cutoff procedures are introduced with the purpose of taking into account ternary ($N = 3$) and quaternary ($N = 4$) atomic relations that satisfy particular multi-metrics.

It was demonstrated that, with the novel cutoff procedures, QuBiLS-MIDAS NQ-MDs that possess superior variability and codify structural information not captured by the corresponding KA-MDs are obtained. It was also demonstrated that QSAR models with greater robustness and predictive power may be developed, when compared with models built exclusively with QuBiLS-MIDAS KA-MDs.

In general, it can be suggested that according to the results obtained in the cheminformatics studies performed, good QuBiLS-MIDAS NQ-MDs may be computed with the following *N*-tuple cutoffs: two-tuple topological (*lag P*) for $P = 1{,}2{,}5{,}9$; two-tuple Euclidean-geometric (*lag L*) for $L = [1, 2]$; [2, 3]; [5, 6]; [8, 9]; three-tuple geometric based on the bond angle multi-metric (*lag BA*) for $BA = [0.0{,}0.314]$;[0.314,0.628];[1,570,1.884]; three-tuple geometric based on the triangle area multi-metric (*lag A*) for $A = [0{,}1]$; [1, 2]; [4, 5]; [6, 7]; three-tuple geometric based on the triangle perimeter multi-metric (*lag TP*) for $TP = [9, 10]$; [10, 11]; [12, 13]; [19, 20]; three-tuple graph-theoretical cutoffs (*lag 3P*) for $3P = 1;2;3$; three-tuple Euclidean-geometric cutoffs (*lag 3L*) for $3L = [1, 2]$; [2, 3]; [6, 7]; four-tuple geometric based on the dihedral angle multi-metric (*lag DA*) for DA = [1.256,1.570]; [2.512,2.826]; four-tuple geometric based on the quadrilateral perimeter multi-metric (*lag QP*) for $QP = [19, 20]$; [21, 22];[25, 26]; four-tuple geometric based on the volume multi-metric (*lag V*) for $V = [0{,}1]$; [1, 2]; [3, 4]; [5, 6]; four-tuple graph-theoretical cutoffs (*lag 4P*) for $4P = 1;2;4;10$; and four-tuple Euclidean-geometric cutoffs (*lag 4L*) for $4L = [1, 2]$; [5, 6].

Therefore, it can be stated that the novel *N*-tuple topological/geometric cutoffs constitute relevant criteria for generating NQ-MDs codifying particular atomic relations, ultimately useful in enhancing the modelling capacity of the QuBiLS-MIDAS 3D-MDs.

Future outlook

In forthcoming reports, the authors intend to apply the mathematical definitions of the novel molecular cutoff procedures defined in this report in the characterizing of macromolecules, e.g. proteins [30]. In addition, the results obtained in this research will be applied in the definition of contact-type matrices as a novel approach to extract structural features of organic compounds using QuBiLS-MIDAS 3D-MDs.

Acknowledgements

CR García-Jacas thanks the support from "Dirección General Asuntos del Personal Académico" (DGAPA) for the postdoctoral fellowship at "Universidad Nacional Autónoma de México" (UNAM) in 2016-2017. Y Marrero-Ponce acknowledges the partial financial support from 'Colegio de Medicina, USFQ'.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] N. Cubillán, Y. Marrero-Ponce, H. Ariza-Rico, S. Barigye, C. García-Jacas, J. Valdes-Martini, and Y. Alvarado, *Novel global and local 3D atom-based linear descriptors of the Minkowski distance matrix: Theory, diversity–variability analysis and QSPR applications*, J. Math. Chem. 53 (2015), pp. 2028–2064.

[2] Y. Marrero-Ponce, C.R. García-Jacas, S.J. Barigye, J.R. Valdés-Martiní, O.M. Rivera-Borroto, R.W. Pino-Urias, N. Cubillán, Y.J. Alvarado, and H. Le-Thi-Thu, *Optimum search strategies or novel 3D molecular descriptors: Is there a stalemate?*, Curr. Bioinform. 10 (2015), pp. 533–564.

[3] C.R. García-Jacas, Y. Marrero-Ponce, S.J. Barigye, J.R. Valdés-Martiní, O.M. Rivera-Borroto, and J. Olivero-Verbel, *N-linear algebraic maps for chemical structure codification: A suitable generalization for atom-pair approaches?*, Curr. Drug Metb. 15 (2014), pp. 441–469.

[4] C.W. Yap, *PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints*, J. Comput. Chem. 32 (2011), pp. 1466–1474.

[5] G. Hinselmann, *BlueDesc-Molecular Descriptor Calculator*. University of Tübingen, Tübingen, Germany, 2008; software available at http://www.ra.cs.uni-tuebingen.de/software/bluedesc/.

[6] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong, *Mold2, Molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics*, J. Chem. Inf. Comput. Sci. 48 (2008), pp. 1337–1344.

[7] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, *DRAGON software: An easy approach to molecular descriptor calculations*, Match 56 (2006), pp. 237–248.

[8] J.J. Sutherland, L.A. O'Brien, and D.F. Weaver, *A comparison of methods for modeling quantitative structure–activity relationships*, J. Med. Chem. 47 (2004), pp. 5541–5554.

[9] C.R. García-Jacas, E. Contreras-Torres, Y. Marrero-Ponce, M. Pupo-Meriño, S.J. Barigye, and L. Cabrera-Leyva, *Examining the predictive accuracy of the novel 3D N-linear algebraic molecular codifications on benchmark datasets*, J. Cheminf. 8 (2016), pp. 1–16.

[10] F. Bonachéra and D. Horvath, *Fuzzy tricentric pharmacophore fingerprints. 2. Application of topological fuzzy pharmacophore triplets in quantitative structure–activity relationships*, J. Chem. Inf. Model. 48 (2008), pp. 409–425.

[11] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner, and A. Zell, *jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints*, J. Cheminf. 3 (2011), pp. 1–14.

[12] A. Klamt, M. Thormann, K. Wichmann, and P. Tosco, *COSMOsar3D: Molecular field analysis based on local COSMO σ-profiles*, J. Chem. Inf. Model. 52 (2012), pp. 2157–2164.

[13] P. Tosco and T. Balle, *A 3D-QSAR-driven approach to binding mode and affinity prediction*, J. Chem. Inf. Model. 52 (2011), pp. 302–307.

[14] D. Gupta-Ostermann, V. Shanmugasundaram, and J. Bajorath, *Neighborhood-based prediction of novel active compounds from SAR matrices*, J. Chem. Inf. Model. 54 (2014), pp. 801–809.

[15] R. Guha, D. Dutta, P.C. Jurs, and T. Chen, *Local Lazy Regression: Making use of the neighborhood to improve QSAR predictions*, J. Chem. Inf. Model. 46 (2006), pp. 1836–1847.

[16] N. Fechner, A. Jahn, G. Hinselmann, and A. Zell, *Atomic local neighborhood flexibility incorporation into a structured similarity measure for QSAR*, J. Chem. Inf. Model. 49 (2009), pp. 549–560.

[17] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, Weinheim, 2009.

[18] M. Randic, A.F. Kleiner, and L.M. De Alba, *Distance/distance matrixes*, J. Chem. Inf. Model. 34 (1994), pp. 277–286.

[19] C.R. García-Jacas, L. Aguilera-Mendoza, R. González-Pérez, Y. Marrero-Ponce, L. Acevedo-Martínez, S.J. Barigye, and T. Avdeenko, *Multi-server approach for high-throughput molecular descriptors calculation based on multi-linear algebraic maps*, Mol. Inform. 34 (2015), pp. 60–69.

[20] C.R. García-Jacas, Y. Marrero-Ponce, L. Acevedo-Martínez, S.J. Barigye, J.R. Valdés-Martiní, and E. Contreras-Torres, *QuBiLS-MIDAS: A parallel free-software for molecular descriptors computation based on multilinear algebraic maps*, J. Comput. Chem. 35 (2014), pp. 1395–1409.

[21] S.J. Barigye, Y. Marrero-Ponce, F. Pérez-Giménez, and D. Bonchev, *Trends in information theory-based chemical structure codification*, Mol. Divers. 18 (2014), pp. 673–686.

[22] J.W. Godden, F.L. Stahura, and J. Bajorath, *Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations*, J. Chem. Inf. Model. 40 (2000), pp. 796–800.

[23] M. Randić, *Generalized molecular descriptors*, J. Math. Chem. 7 (1991), pp. 155–168.

[24] R.P. Urias, S. Barigye, Y. Marrero-Ponce, C. García-Jacas, J. Valdes-Martiní, and F. Perez-Gimenez, *IMMAN: Free software for information theory-based chemometric analysis*, Mol. Divers. 19 (2015), pp. 305–319.

[25] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.

[26] R.D. Cramer, D.E. Patterson, and J.D. Bunce, *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc. 110 (1988), pp. 5959–5967.

[27] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, MobyDigs: software for regression and classification models by genetic algorithms, in *Data Handling in Science and Technology*, R. Leardi ed., Elsevier, 2003, pp. 141–167.

[28] M. Friedman, *A Comparison of alternative tests of significance for the problem of m rankings*, Ann. Math. Stat. 11 (1940), pp. 86–92.

[29] J. Alcalá-Fdez, L. Sánchez, and S. García, M.J.d. Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, and F. Herrera, *KEEL: A software tool to assess evolutionary algorithms for data mining problems*, Soft Comput. 13 (2009), pp. 307–318.

[30] Y. Marrero-Ponce, E. Contreras-Torres, C.R. García-Jacas, S.J. Barigye, N. Cubillán, and Y.J. Alvarado, *Novel 3D bio-macromolecular bilinear descriptors for protein science: Predicting protein structural classes*, J. Theor. Biol. 374 (2015), pp. 125–137.