Molecular Diversity
https://doi.org/10.1007/s11030-019-10002-3

ORIGINAL ARTICLE

When global and local molecular descriptors are more than the sum of its parts: *Simple, But Not Simpler?*

Yoan Martínez-López[1,2,3] · Yovani Marrero-Ponce[1,4] · Stephen J. Barigye[5,6] · Enrique Teran[1] · Oscar Martínez-Santiago[1,3,7] · Cesar H. Zambrano[4,7] · F. Javier Torres[4,7]

Received: 31 July 2019 / Accepted: 9 October 2019

Abstract

In this report, we introduce a set of aggregation operators (AOs) to calculate global and local (group and atom type) molecular descriptors (MDs) as a generalization of the classical approach of molecular encoding using the sum of the atomic (or fragment) contributions. These AOs are implemented in a new and free software denominated **MD-LOVIs** (http://tomocomd.com/md-lovis), which allows for the calculation of MDs from atomic weights vector and LOVIs (local vertex invariants). This software was developed in Java programming language and employed the Chemical Development Kit (CDK) library for handling chemical structures and the calculation of atomic weights. An analysis of the complexities of the algorithms presented herein demonstrates that these aspects were efficiently implemented. The calculation speed experiments show that the **MD-LOVIs** software has satisfactory behavior when compared to software such as Padel, CDKDescriptor, DRAGON and Bluecal software. Shannon's entropy (SE)-based variability studies demonstrate that **MD-LOVIs** yields indices with greater information content when compared to those of popular academic and commercial software. A principal component analysis reveals that our approach captures chemical information orthogonal to that codified by the DRAGON, Padel and Mold2 software, as a result of the several generalizations in **MD-LOVIs** not used in other programs. Lastly, three QSARs were built using multiple linear regression with genetic algorithms, and the statistical parameters of these models demonstrate that the **MD-LOVIs** indices obtained with AOs yield better performance than those obtained when the summation operator is used exclusively. Moreover, it is also revealed that the **MD-LOVIs** indices yield models with comparable to superior performance when compared to other QSAR methodologies reported in the literature, despite their simplicity. The studies performed herein collectively demonstrated that **MD-LOVIs** software generates indices as simple as possible, but not simpler and that use of AOs enhances the diversity of the chemical information codified, which consequently improves the performance of traditional MDs.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11030-019-10002-3) contains supplementary material, which is available to authorized users.

✉ Yovani Marrero-Ponce
ymarrero@usfq.edu.ec; ymarrero77@yahoo.es
http://www.uv.es/yoma/
http://ymponce.googlepages.com/home

[1] Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas, Universidad San Francisco de Quito (USFQ), Av. Interoceánica Km 12 ½ —Cumbayá, 170157 Quito, Ecuador

[2] Grupo de Investigación de Inteligencia Artificial (AIRES), Facultad de Informática, Universidad de Camagüey, Camagüey, Cuba

[3] Department of Computer Sciences, Faculty of Computer Sciences, Camagüey University, 74650 Camagüey City, Camagüey, Cuba

[4] Instituto de Simulación Computacional (ISC-USFQ), Universidad San Francisco de Quito (USFQ), Diego de Robles y vía Interoceánica, 17-1200-841, Quito, Ecuador

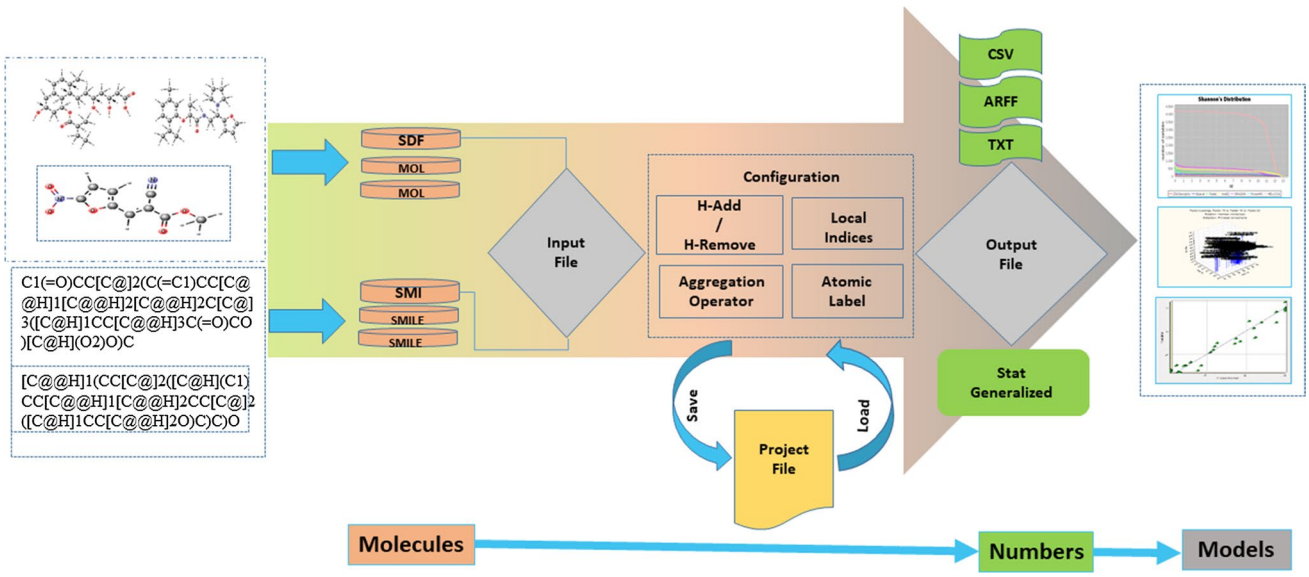
[5] Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain

[6] ProtoQSAR SL, Centro Europeo de Empresas Innovadoras (CEEI), Parque Tecnológico de Valencia, 46980 Paterna, Valencia, Spain

[7] Grupo de Química Computacional y Teórica (QCT-USFQ), Departamento de Ingeniería Química, Universidad San Francisco de Quito, Diego de Robles y Vía Interoceánica, 17-1200-841, Quito, Ecuador

Published online: 28 October 2019

Graphic abstract



Keywords Molecular descriptor · Aggregation operator · Atom weight vector · **MD-LOVIs** software · No free lunch theorem · QSP(A)R · PCA · Shannon entropy

> Everything Should Be Made as Simple as Possible, But Not Simpler.
>
> Albert Einstein, 1933.

Introduction

Molecular descriptors (MDs) are defined as "the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or result of some standardized experiment" [1]. The MDs allow the assignment of numbers to molecular structures based on their intrinsic properties [1, 2], and these may be obtained following different strategies, for example: (i) counting particular atom types or structural fragments (0D or 1D descriptors), (ii) applying mathematical algorithms to molecular graphs (the so-called topological or 2D MDs), (iii) from geometrical representations of molecular structures, that is, geometrical or 3D MDs, among others [1]. These MDs have been implemented in different software, such as DRAGON [3], CODESSA [4], Padel [5], QuBILS-MIDAS [6], QuBILS-MAS [7], Mold2 [8], CDKDescriptor Calculator [9] and ChemDes [10].

On the other hand, traditionally most indices use the sum of their parts (e.g., atom- or bond-based labels) as the mathematical formalism to obtain global descriptions of molecules, for instance, the first Zagreb index $Z_{g1}(G)$ in Eq. 1 is defined as the sum of squares of the vertex degrees, δ_i and δ_j, of vertices u and v in a molecular graph (G) [11],

$$Z_{g1}(G) = \sum_{b} (\delta_i^2 + \delta_j^2) \tag{1}$$

where b refers to the bond i–j and the sum is over all bonds B of the molecule. The Randic index $R(G)$ is also based on the vertex degrees and is defined as the sum all bonds in the molecule (Eq. 2), where δ_i and δ_j are the vertex degrees of the atoms incident to the considered bond b [12]:

$$R(G) = \sum_{b} \frac{1}{\sqrt{\delta_i \delta_j}} \tag{2}$$

Moreover, the Moreau–Broto autocorrelation (autocorrelation of a topological structure, ATS) [13] for a G is defined as (Eq. 3):

$$\mathrm{ATS}_d = \sum_{i=1}^{A} \sum_{j=1}^{A} \delta_{ij} \cdot \left(w_i \cdot w_j\right)_d = \mathbf{w}^{\mathbf{T}} \cdot^{\mathbf{m}} \mathbf{B} \cdot \mathbf{w} \tag{3}$$

where w is any atomic property, A is the atom number, d is the considered topological distance (i.e., the lag in autocorrelation terms) and δ_{ij} is Kronecker delta ($\delta_{ij} = 1$ if $\delta_{ij} = d$, zero otherwise). $^{\mathbf{m}}\mathbf{B}$ is the mth-order binary sparse matrix, and $\mathbf{w}$ the A-dimensional vector of atomic properties.

Also, the gravitational index (Eq. 4) characterizes the mass distribution in a molecule, where m_i and m_j are the atomic masses of the considered atoms, r_{ij} the corresponding interatomic distances, and A the number of atoms of the molecule [14]:

$$G_1 = \sum_{i=1}^{A} \sum_{j=i+1}^{A} \frac{m_i \cdot m_j}{r_{ij}^2} \tag{4}$$

Likewise, the total sum operator (TS), i.e., the sum of all the elements in a matrix representation of a molecule, is also used to calculate many indices (Eq. 5):

$$\mathrm{TS}(M) \equiv \sum_{i=1}^{n} \sum_{j=1}^{p} m_{ij} \tag{5}$$

Additionally, the Kier–Hall connectivity indices (Eq. 6), where k denotes the mth-order subgraph comprised of n atoms ($n = m + 1$ for acyclic subgraphs), K is the total number of mth-order subgraphs present in the G and δ_i is the vertex degrees [15]:

$${}^{m}\chi_q = \sum_{k=1}^{K} \left(\prod_{i=1}^{n} \delta_i \right)_k^{-1/2} \tag{6}$$

It is recognized that many observables such as the bond lengths or molecular van der Waals volumes are not equivalent to the sum of the corresponding atomic radii, or atomic van der Waals volumes, as these often require corrections [16]. From an algorithmic perspective, it is acknowledged that there is no universally superior algorithm for all optimization problems, evocative of the "no free lunch" theorem, which postulates that a method unsuitable for particular application may perform ideally in another [17]. Extrapolated to the definition of molecular descriptors, it may be postulated that an aggregation operator that yields descriptors with poor correlation for a given endpoint may suitably correlate with another.

Indeed, better correlations for determined bioactivities resulted with global MDs based on other aggregation operators (AOs) than their summation, further corroborating the notion that it is an overly simplified characterization of the collective contribution of atoms [18–20]. Similar principles are regularly followed in workflows for data structure analysis. For example, from a statistical perspective when one wishes to study the asymmetric tendency of data matrices, the variance offers no useful information as it is particularly related to the central tendency of data points, while skewness and kurtosis are deeply instructive for such cases [21]. Also, when one wishes to compare variables of different properties or ranges, the arithmetic mean may be misleading due to the disparity in the ranges, while the geometric mean "suppresses" this difference making them comparable [22]. Therefore, it is conceivable to suggest that not all global (molecular) chemical phenomena are ideally represented by the summation operator on atomic contributions.

In mathematics, AOs [23, 24] are algebraic functions, which assign a real number y to an n-dimensional vector of real numbers. These AOs are based on linear and/or additive measures.

AOs has been successfully used in diverse applications, such as face recognition [25], data mining [26], decision-making [27] and molecular structures encoding, which is an essential phase in cheminformatics tasks such as diversity analysis [28] molecular similarity searching [29] and quantitative structure–activity relationship (QSAR) studies [30].

The classical invariants that are traditionally used to obtain indices such as autocorrelation, Kier–Hall and gravitational indices use graph-theoretical labels as weights, for example the vertex degrees or some basic atomic properties. One way to extend these indices is to incorporate other atomic weights as labels. In this study, three groups of several atom-based properties are proposed, which expand on the few atomic properties or vertex degrees that have been traditionally used. Finally, with some exceptions, such as the electrotopological state indices, algebraic forms [31] and GT-STAF descriptors [18], it is common to calculate only global (total) indices for the whole molecule. However, it is known that some molecular properties (e.g., basicity, nucleophilic addition, polarity) are influenced by chemical fragments instead of being influenced by the whole molecular body. Therefore, in the present report we also propose indices that encode information on fragment types in molecular structures (i.e., local indices).

This report aims to propose a generalized scheme to obtain different MDs applying AOs to simple atomic weights vectors. Moreover, in order to guarantee the accessibility and usability of these MDs, we provide a free computational program for calculating these MDs, denominated **MD-LOVIs** (acronym for **M**olecular **D**escriptors from **LO**cal **V**ertex **I**nvariant**s** and Related Maps) software, http://tomocomd.com/md-lovis. To assess the utility of these MDs in cheminformatics tasks, we compare the performance of the **MD-LOVIs** descriptors with those of other software reported in the literature using several techniques, such as variability analysis, principal component analysis, as well as QSAR modeling. Finally, to facilitate the use of the **MD-LOVIs** software we are providing predefined sets of descriptors (in form of descriptor projects) containing lists of descriptors identified as most optimal based on the analyses performed herein.

Theoretical outline

A molecular graph $G(V, E)$ is defined as a 2D representation of a chemical structure comprising a set V of vertices (atoms) and the corresponding set E of edges, which represent the chemical bonds. Often, the vertices in a G are weighted with determined atomic labels, such as the electronegativity and Van der Waals radii to achieve greater discrimination of organic compounds, and these types of Gs are denoted as ***weighted Gs*** [32].

The ordered weighted averaging (OWA) are example of AOs [33], which in different types of distance measures has been studied by many authors. Merigó and Gil-Lafuente [33] proposed the OWA distance (OWAD) operator with the aim of introducing a parameterized family of distance operators between the minimum and the maximum distance. Xu also studied the use of fuzzy information [34], and other authors studied the use of Choquet integrals [35, 36]. Merigó [37] proposed the GOWAWA operator, while García-Jacas [19] applied this operator in QSAR studies, which is based on the fusion of the generalized ordered weighted averaging (GOWA) and the weighted generalized mean (WGM) functions in the same formulation for codifying the chemical information of molecules. The models built with variables derived from the GOWAWA operators were shown to possess greater predictive power relative to those built with the sum operator exclusively. Their main advantage is that they can integrate the OWA operator and the WA considering the degree of importance that each concept has in the aggregation.

The information codified in a weighted G may be projected in a $\mathbb{R}^n$ space, where $\mathbb{R}$ is a set of real numbers and n is the number of atoms in a molecule. If the molecular vector $\vec{W}$ in the $\mathbb{R}^n$ space is considered to start from the origin, then the vector components represent given atom, atom type or group atomic weights. Consequently, the AOs may be applied to $\vec{W}$ yielding a set of total and local MDs. The AOs, as molecular structure characterizing parameters have an important property of being invariant; thus, they yield the same result irrespective of the numbering of the vector components. In this manuscript, the term invariant and AO will be used interchangeably.

Let us take as an example the G of 2-methylbutanal. Firstly, this G may be represented in the $\mathbb{R}^6$ space by a six-dimensional vector, with components $[x_{C1}, x_{C2}, x_{C3}, x_{C4}, x_{C5}, x_{O6}]$ representing particular atomic properties (see Fig. 1).

When Pauling's electronegativity (E) for each atom is taken as atomic weight and is divided by vertex degree [i.e., $W_i = E\,(S_i)/\delta(S_i)$], the following vector of atomic labels is obtained (see Fig. 1c),

$\vec{W}$ **= [E(C_1)/ δ(C1), E(C_2)/ δ(C2), E(C_3)/ δ(C3), E(C_4)/ δ(C4), E(C_5)/ δ(C5), E(O_6)/ δ(O6)] = [2.55/1, 2.55/2, 2.55/3, 2.55/1, 2.55/3, 3.5/2] = [2.55, 1.275, 0.85, 2.55, 0.85, 1.75].**

Atomic weights (AW)

A group of atomic weights (AWs) [1] are used to generate the n-dimensional vector of atomic properties. These AWs are standard values or they are calculated from atoms of a molecule and are classified in three main groups.

The *first group* constitutes **Chemical** AWs, such as atomic number (Z), atomic mass (A), Van der Waals volume (VW), covalent radius (R), polarizability (P) and Pauling electronegativity (E). These Chemical AWs, in the formalism presented herein, will always be divided by the *bond vertex degree* (δ) of each vertex in the G.

The *second group* is comprised of **Physicochemical** AWs, such as topological surface area (T) [38], Ghose–Crippen

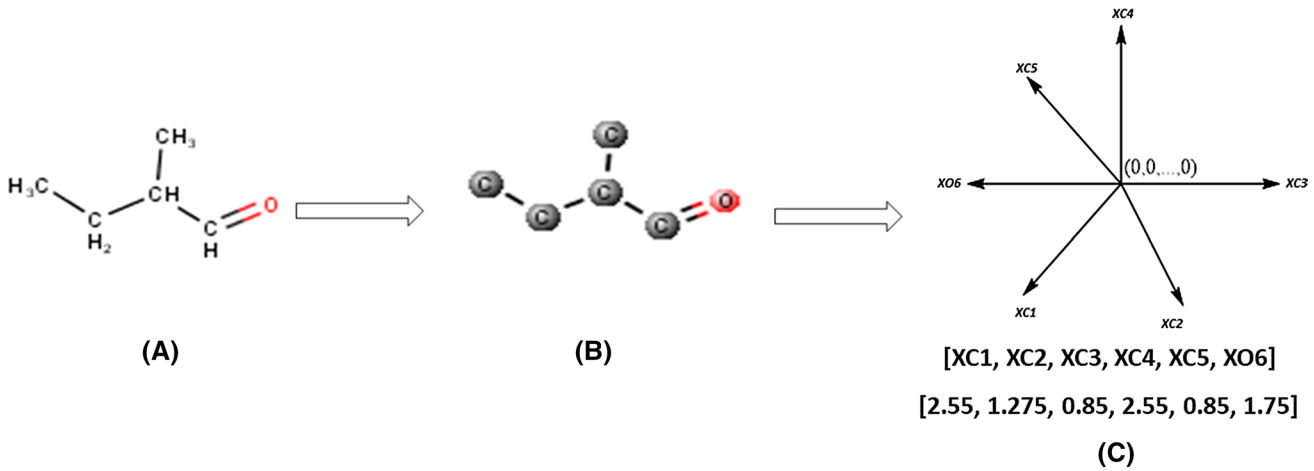


Fig. 1 **a** 2-Methylbutanal molecule, **b** H-suppressed molecular graph of 2-methylbutanal, G, **c** representation of the space of LOVIs of G

ALogP (*L*) and molar refractivity (*M*) [39], as well as the atomic charge (*C*) [40].

The *third group* of AWs is comprised of **vertex degrees** computed on *G* and includes the following:

The valence degree (*N*), defined by Kier and Hall [41], is mathematically expressed as (Eq. 7):

$$N_i = \frac{\left(Z_i^v - h_i\right)}{\left(Z_i - Z_i^v - 1\right)} \tag{7}$$

where Z_i is the total number of electrons of the *i*th atom (i.e., atomic number), Z_i^v is the number of valence electrons (σ electrons, π electrons and *n* lone pair electrons,) of the *i*th atom and h_i is the number of hydrogen atoms bonded to it. The intrinsic state (*I*) is a modification of the valence degree (*N*), also proposed by Kier and Hall, is defined as (Eq. 8):

$$I_i = \frac{\left(2/L_i\right)^2.N_i + 1}{\delta_i} \tag{8}$$

where L_i is the principal quantum number, N_i is the valence vertex degree and δ_i is the simple vertex degree of the *i*th atom. The L_i is used to account for the increase in the screening effect of the inner electrons, and δ_i is the sum of elements in the adjacency matrix of *G* [41]:

The electrotopological state (ES) or also called the E-state for atom a_i is expressed as [42]:

$$\mathrm{ES}_i = I_i + \sum_{j=1}^{A} \Delta I_{ij} \tag{9}$$

where I_i is the intrinsic state of atom a_i and ΔI_{ij} is the perturbation of I_i, which is in turn computed using Eq. 10, where d_{ij} is the topological distance between atoms a_i and a_j of the *G* [43].

$$\Delta I_{ij} = \frac{I_i - I_j}{d_{ij}^2} \tag{10}$$

The Kupchik's vertex degree (KU) is proposed to take into account the covalent radius of atoms in the *G* (Eq. 11), where R_c and R_i are the covalent radius of the carbon atom and the *i*th atom of the molecule, respectively, and N_i is the valence vertex degree [44]:

$$\mathrm{KU_i} = \frac{\mathrm{R_c}}{\mathrm{R_i}}.\mathrm{N_i} \tag{11}$$

The Bond's vertex degree (BD) accounts for bonding as well as the bond multiplicity (Eq. 12). It is calculated from the atom connectivity matrix **C** as sum of row entries, where *A* is the number of graph vertices (atoms) and C_{ij} are elements of **C**. The elements of this matrix are equal to one for simple bonds, two for double bonds, three for triple bonds (here, conjugated bonds are considered as double bonds) and zero no bonded atoms:

$$\mathrm{BD_i} = \sum_{\mathrm{j=1}}^{\mathrm{A}} \mathrm{C_{ij}} \tag{12}$$

The Hu–Xu's vertex degree (HX), proposed by Hu–Xu et al. [45] (Eq. 13), is expressed as the sum of the elements (δ_i) of the adjacency matrix of *G* and Z_i the atomic number of the considered atom, a_i:

$$\mathrm{HX}_i = \delta_i.\sqrt{Z_i} \tag{13}$$

The Li's vertex degree (LI) [46] is defined as (Eq. 14):

$$\mathrm{LI}_i = \frac{Z_i^v.\left(Z_i^v - h_i\right)}{L_i^2} \tag{14}$$

where Z_i^v is the number of valence electrons (σ electrons, π electrons and *n* lone pair electrons) of the *i*th atom, h^i is the number of hydrogen atoms bonded to it and L_i is the principal quantum number.

The Alikhanidi's vertex degree (Alk) is proposed as a modification of the Hu–Xu vertex degree and is expressed as [47] (Eq. 15):

$$\mathrm{Alk}_i = \partial_i \sqrt{Z_i'} \tag{15}$$

where Z_i' is a function of the atomic numbers Z_j of the atoms adjacent to the *i*th atom (also known as consecutive AT number) and is defined as (Eq. 16):

$$Z_i' = \left[\sum_{j=1}^{A} a_{ij}\sqrt{\left(\sqrt{2} + Z_j\right)}\right]^2 \tag{16}$$

where a_{ij} denotes the elements of the adjacency matrix.

The Ivanciuc's vertex degree (IN), proposed as a combination of topological distances and vertex degrees in *G*, is computed according to the following general expression [48] (Eq. 17):

$$\mathrm{IN}_i = \sum_{j=1}^{A} d_{ij}^{\alpha}\partial i^{\beta}\partial j^{\gamma} \tag{17}$$

where d_{ij} is the topological distance between vertices v_i and v_j; $\alpha = 0, 1$; $\beta = 0, 1, -1$; $\gamma = 0, 1, -1$; and ∂_i and ∂_j are the vertex degrees for v_i and v_j, respectively.

The distance count (DC) indicates the frequencies $\{{}^1f_i, {}^2f_i, {}^3f_i, \ldots, {}^Df_i\}$ of distances equal to $\{1, 2, 3, \ldots D\}$, respectively, from vertex v_i to any other vertex of a *G*; *D* is the eccentricity (maximum distance from v_i). The eccentric connectivity (*Y*) is defined as the sum of products between eccentricity

D_i and valence vertex degree δ_i of a G and is expressed as (Eq. 18):

$$Y_i = \sum_{i=1}^{A} D_i \delta_i \tag{18}$$

Note that the AWs may be computed from an H-suppressed or H-filled G. Therefore, with the set of AWs computed for a given molecular structure, the corresponding atomic weights vector, $\overrightarrow{\boldsymbol{W}}$, is constructed and possesses the following structure:

$$\overrightarrow{\boldsymbol{W}} = \left[w_1, w_2, w_3, \ldots., w_n\right] \tag{19}$$

where w_i is the weight for atom a_i and n is the number of atoms of G.

New generalized indices: aggregation operators applied to atomic weights vector

The AOs [33] are applied to the atomic weights vector,$\vec{\boldsymbol{W}}$, providing a generalized scheme of the classical approach of computing the global (or local) MDs by summation of the vector components. These AOs are classified in four main groups. The first group is the **Norms** (or **Metric**s) AOs, which include: Minkowski's norms [i.e., N1 (equivalent to the summation), N2, N3]. The second group is the **Means** AOs (***first statistical moment***) and comprises: geometric mean (G), arithmetic mean (M), quadratic mean ($P2$), power mean ($P3$) and harmonic mean (A). Finally, the third group is the **Statistical** operators (***highest statistical moments***), which includes: variance (V), skewness (S), kurtosis (K), standard deviation (SD), variation coefficient (VC), range (R), percentile 25 ($Q1$), percentile 50 ($Q2$), percentile 75 ($Q3$), inter-quartile range (I50), X max (MX) and X min (MN).

The fourth group is the **"Classical Algorithms"** invariants, which comprises autocorrelation (AC), gravitational (GI), total information content (TI), mean information content (MI), standardized information content (SI), total sum (TS), Ivanciuc–Balaban (IB), electrotopological state (ES) and Kier–Hall connectivity (CN). Note that most of the AOs in this group can be generalized by using the first three AOs groups (denoted here as non-classical AOs), given that these algorithms in turn involve the summation operator; see Table 1.

For example, if a first-order AC operator is applied to $\overrightarrow{\boldsymbol{W}}$ obtained from G of 2-methylbutanal, using Pauling's electronegativity (E) as atomic label, $\vec{\boldsymbol{W}}$ = **[2.55**, **1.275**, **0.85**, **2.55**, **0.85**, **1.75]**.

$$AC_1 = \sum_{i=1}^{n}\sum_{j\geq 1}^{m} L_i \times L_j * \left(\delta\left(d_{ij}\right)\right)$$

$$AC_1 = [\boldsymbol{AC}_1(\mathbf{C}1), \boldsymbol{AC}_1(\mathbf{C}2), \boldsymbol{AC}_1(\mathbf{C}3), \boldsymbol{AC}_1(\mathbf{C}4), \boldsymbol{AC}_1(\mathbf{C}5), \boldsymbol{AC}_1(\mathbf{O}6)]$$

$$AC_1(C1) = w1 * w2 = 2.55 * 1.275 = 3.25125$$

$$\begin{aligned} AC_1(C2) &= w1 * w2 + w2 * w3 \\ &= 2.55 * 1.275 + 1.275 * 0.85 = 4.335 \end{aligned}$$

$$\begin{aligned} AC_1(C3) &= w2 * w3 + w3 * w4 + w3 * w5 \\ &= 1.275 * 0.85 + 0.85 * 2.55 + 0.85 * 0.85 \\ &= 3.97375 \end{aligned}$$

$$AC_1(C4) = w3 * w4 = 0.85 * 2.55 = 2.1675$$

$$AC_1(C5) = w3 * w5 + w5 * w6 = 0.85 * 0.85 + 0.85 * 1.75 = 2.21$$

$$AC_1(\mathrm{O}6) = w5 * w6 = 0.85 * 1.75 = 1.4875$$

$$AC_1 = [3.25125, 4.335, 3.97375, 2.1675, 2.21, 1.4875]$$

If non-classical AOs are applied, then a series of indices can be obtained as generalization of the sum of their parts over classical invariants, for example:

$$\begin{aligned} N1\left(AC_1\right) &= \sum_{i=1}^{6} AC_{1\mathrm{i}} = (3.25125 + 4.335 + 3.97375 \\ &\quad +2.1675 + 2.21 + 1.4875) = 17.37 \end{aligned}$$

$$\begin{aligned} N2\left(AC_1\right) &= \sqrt{\sum_{i=1}^{6} AC_{1\mathrm{i}}^2} \\ &= \sqrt{3.25125^2 + 4.335^2 + 3.97375^2 + 2.1675^2 + 2.21^2 + 1.4875^2} \\ &= 6.18 \end{aligned}$$

$$\begin{aligned} MX\left(AC_1\right) &= \max(3.25125, 4.335, 3.97375, \\ &\quad 2.1675, 2.21, 1.4875) = 4.34 \end{aligned}$$

$$\begin{aligned} G\left(AC_1\right) &= \sqrt[6]{\prod_{i=1}^{6} AC_{1\mathrm{i}}} \\ &= \sqrt[6]{3.25125 * 4.335 * 3.97375 * 2.1675 * 2.21 * 1.4875} \\ &= 1.52 \end{aligned}$$

The same applies to all the indices of the fourth group. That is to say, the classical forms (fourth group) may as well be generalized using the non-classical AOs of the first three groups, as shown in the example of AC order 1, based on E as the atomic label.

Local and total molecular descriptors

Motivated by the understanding that the chemical, physicochemical or biological activity of compounds does not

Table 1 Norms, Means and Statistical AOs as generalizations of the linear combination of *W* and Classical Algorithms

No.	Group[a]	Name	Identifier	Formula[b]
20	Norms (Metrics)	Minkowski norm ($p=1$) *Manhattan norm*	$N1$	$N1=\sum_{i=1}^{n}\lvert L_i\rvert$
21		Minkowski norm ($p=2$) *Euclidean norm*	$N2$	$N2=\sqrt{\sum_{i=1}^{n}\lvert L_i\rvert^2}$
22		*Minkowski norm* ($p=3$)	$N3$	$N3=\sqrt[3]{\sum_{i=1}^{n}\lvert L_i\rvert^3}$
23	Mean (first statistical moment)	Geometric mean	G	$G=\sqrt[n]{\prod_{i=1}^{n}L_i}$
24		Arithmetic mean (power mean of degree β = 1)	M	$M_\beta=\left(\frac{L_1^\beta+L_2^\beta+\ldots+L_n^\beta}{n}\right)^{\frac{1}{\beta}}$
25		Quadratic mean (power mean of degree $\beta=2$)	$P2$	
26		Power mean of degree $\beta=3$	$P3$	
27		Harmonic mean (power mean of degree $\beta=-1$)	A	
28	Statistical (highest statistical moments):	Variance	V	$V=\frac{\sum_{i=1}^{n}(L_i-M)}{n-1}$
29		Skewness	S	$S=\frac{n*(X_3)}{(n-1)(n-2)(\mathrm{DE})^3}$ $X_3=\sum_{i=1}^{n}\left(L_i-M\right)^3$ M, arithmetic mean SD, standard deviation
30		Kurtosis	K	$K=\frac{n(n+1)X_4-3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(\mathrm{DE})^4}X_j=\sum_{i=1}^{n}\left(L_i-M\right)^j$ M, arithmetic mean SD, standard deviation
31		Standard deviation	SD	$\mathrm{SD}=\sqrt{\frac{\left(\sum_{i=1}^{n}L_i-M\right)^2}{n-1}}$
32		Variation coefficient	VC	$\mathrm{VC}={}^{\mathrm{DE}}/_{M}$
33		Range	R	$R=L_{\max}-L_{\min}$
34		Percentile 25	$Q1$	$Q1=\left[\frac{N}{4}+\frac{1}{2}\right]$ N, La number
35		Percentile 50	$Q2$	$Q2=\left[\frac{N}{2}+\frac{1}{2}\right]$ N, La number
36		Percentile 75	$Q3$	$Q3=\left[\frac{3N}{4}+\frac{1}{2}\right]$ N, La number
37		Inter-quartile range	I50	$I50=Q3-Q2$
38		Maximum value	MX	MX = La max
39		Minimum value	MN	MN = La min
40	Classical	Autocorrelation	AC_k	$\mathrm{AC}_k=\sum_{i=1}^{n}\sum_{j=1}^{n}L_i\times L_j\cdot\left(\delta\left(d_{ij},k\right)\right)$ $k=1,2,\ldots7$
41		Gravitational	GI_k	$\mathrm{GI}_k=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{L_iL_j}{d_{ij}^k}\cdot\left(\delta\left(d_{ij},k\right)\right)$ $k=1,2,\ldots7$
42		Total sum at lag k	TS_k	$\mathrm{TS}_k=\sum_{i=1}^{n}\sum_{j=1}^{n}L_{ij}\cdot\left(\delta\left(d_{ij},k\right)\right)$ $k=1,2,\ldots7$

Table 1 (continued)

No.	Group[a]	Name	Identifier	Formula[b]
43		Kier–Hall connectivity	CN_t^m	$CN_t^m = \sum_{k=1}^{K} \left(\prod_{i=1}^{nk} L_i \right)_k^{\lambda}$ where K is the number of subgraphs, nk is the number of atoms in a fragment, λ is equal to ½, m and t are the subgraph order and type, respectively
44		Mean information content	MI	$MI = -\sum_{g=1}^{G} \frac{N_g}{N_0} \cdot \log_2 \frac{N_g}{N_0}$ where N_g is the number of atoms with the same LOVI value. N_0 is the number of atoms in a molecule
45		Total information content	TI	$TI = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g$
46		Standardized information content	SI	$SI = \frac{TI}{N_0 \cdot \log_2 N_0}$
47		Electrotopological state (E-state index)	ES	$ES = I_i + \Delta I_i + \sum_{j=1}^{n} \frac{I_i - I_j}{(d_{ij}+1)^2}$ where I_i is the intrinsic state of the ith atom and ΔI_i is the field effect on the ith atom calculated as perturbation of the Ii of ith atom by all other atoms in the molecule, d_{ij} is the topological distance between the ith and the jth atoms, and n is the number of atoms. The exponent k is 2.
48		Ivanciuc–Balaban-type indices	IB	$IB = \frac{n^2 \cdot B}{n+C+1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \left[L_i \times L_j \right]^{\frac{-1}{2}}$ where the summation goes over all pairs of atoms, but only pairs of adjacent atoms are accounted for by means of the elements a_{ij} of the adjacency matrix. The n, B and C are the number of atoms, bonds and rings (cyclomatic number), respectively

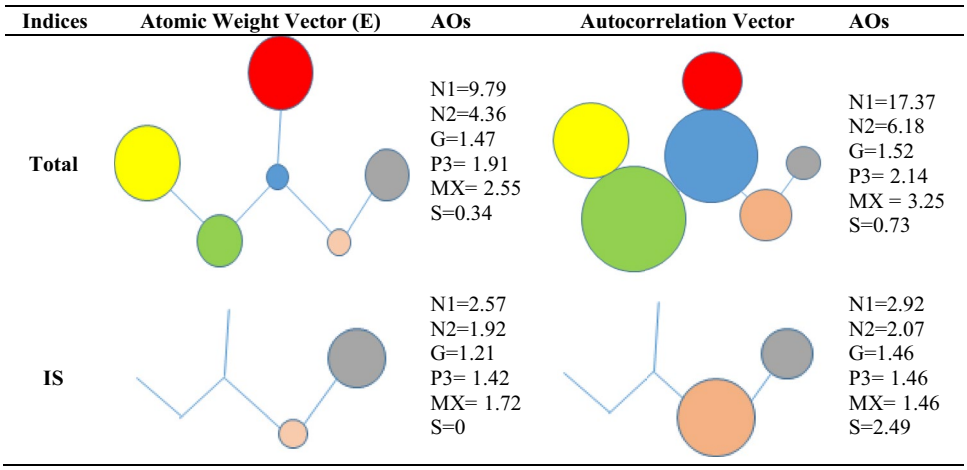
[a]The second group (AOs **20-22**) could be renamed as "location statistics" if percentiles and maximum (minimum) are taken into consideration in this group. In this case, the third group (AOs **23-39**) could be renamed as "spread and shape statistics"

[b]W for "*i*" atoms in molecule

always depend on the molecular structure as a whole, but also on particular portions, which may be functional groups, substructures or unsaturated regions, the present approach is further refined to allow the computation of total or local definitions of a molecular structure. In this sense, all or a subset of vector components may be considered, to obtain total or local MDs, respectively. Ten local types (or groups) are analyzed in the present report, i.e., hydrogen bond acceptors (HA), carbon atoms in aliphatic chains (LA), hydrogen bond donors (HD), halogens (HL), terminal methyl groups (MD), carbon atoms (CB), carbon atoms in aromatic systems (RA), heteroatoms (O, N and S in all valence states, denoted as HT), unsaturated bonds (IS) and groups at lag k (GL calculates a new vector $\vec{W}$ from the corresponding one, taking into account the topological distance for a determined step k between a pair of atoms, greater or equal to a determined cutoff value related to the number of atoms, found at determined distances and/or group of selected local types); for example, see Scheme 1. This scheme depicts the workflow followed in the computation of the total or local fragment indices, where different AOs are applied. In this case, IS and total were used as examples, applied to the $\vec{W}$ and AC vectors, respectively. Likewise, other local types may be considered yielding the corresponding vectors to which the AOs may be applied.

Software design

To calculate the MDs based on the application of AOs on $\vec{W}$ as previously described, the **MD-LOVIs** software was developed (http://tomocomd.com/md-lovis). This computational tool offers a user-friendly platform to allow for quick and straightforward calculation of MDs. Moreover, it was



Scheme 1 Graphical representation of calculation of total and local MDs from LOVIs (AW (using *E* values) vector and autocorrelation vector) by using some AOs in **MD-LOVIs** software

designed to run on any host operating systems which supports Java (TM) 7 Runtime Environment and it is, therefore, platform independent.

The **MD-LOVIs** software is based on the Chemical Development Kit (CDK) library [40], which is an open-source library of algorithms for structural chemo- and bioinformatics studies. It serves as a base for many other applications, including some parts of **MD-LOVIs**.

Algorithms for calculating MDs using AOs

The following algorithm is used to calculate the MDs presented herein:

To begin:

Step 1 Compute the $\vec{W}$ from the MS.
Step 2 Compute the local atomic weights vector ($\vec{W}^{L}$) from *W*, where $\vec{W}$ is a subset of $\vec{W}$.
Step 3 Compute the **MD** by applying the **AO** to $\vec{W}^{L}$: MD = AO ($\vec{W}^{L}$).

The **MD** names are defined based on the following scheme: $\left[\textbf{Aggregation Operator}\right]^{[\textbf{Property}]}_{[\textbf{fragment-type formalism}]}$, e.g., $\textbf{N1}^{\textbf{HX}}\,\textbf{HT}$ corresponds to a MD computed using the N1-norm applied to a vector $\vec{W}$ with the Hu–Xu's vertex degree values (as atomic labels) for all heteroatoms within a structure. As it may be observed, the computation of the MDs using AOs follows a simple and straightforward algorithm.

In order to evaluate the computational complexity of the algorithm, some experiments were carried out, where appropriate behavior observed, and the results are given in supporting information SI-2. Further analysis on the computational complexity is provided in the stress experiments carried out in subsequent sections; see Sect. 4.1.

MD-LOVIs software

Figure 2 shows the **MD-LOVIs**' flowchart collection, which includes the program architecture, main descriptors configuration sequence and the output descriptor headers.

Figure 3 shows the UML diagram [49] of most important classes implemented in the **MD-LOVIs** software to perform the descriptor computation. The UML diagram for the packages is provided as Supporting Information SI-3.

In Fig. 3, the ***MolecularDescriptor*** class defines the concept corresponding to the MDs based on $\vec{W}$. On the other hand, the ***Statistical***, ***Mean***, ***Norm*** and ***Classical Algorithm*** classes implement the algorithms for the AOs. Moreover, the ***AtomicWeight***, ***Chemical***, ***Physical*** and ***VertexDegree*** classes contain the implementation of AWs for the molecular structures. In addition, the ***LocalType*** and ***CDKfunction*** classes implement the local fragment and functions obtained from CDK, respectively.

All the requests performed by the users through the graphic user interface (GUI) are processed by the **MD-LOVIs** library. This component is organized in packages according to the goals of the functionalities, so as to facilitate its understanding. The principal package is Main, which does contain the packages ***CDK***, ***View***, ***Local***, ***Properties*** and ***AggregationOperators***; see **SI-3**. The ***Norms***, ***Means***, ***Statistics*** and ***Classical Algorithm*** packages contain the classes for the computation of AOs. Moreover, the ***Chemical***, ***Physical*** and ***VertexDegree*** packages contain the classes for calculating atomic weights. On the other hand, the ***View*** package includes the classes for working with the GUI.

The ***CDK*** package includes the classes, which use the functionalities of the CDK descriptor. The ***View*** package contains the objects responsible for building the visualization of the program. The ***Locals*** package contains the classes related to local fragments. The ***Properties*** package presents

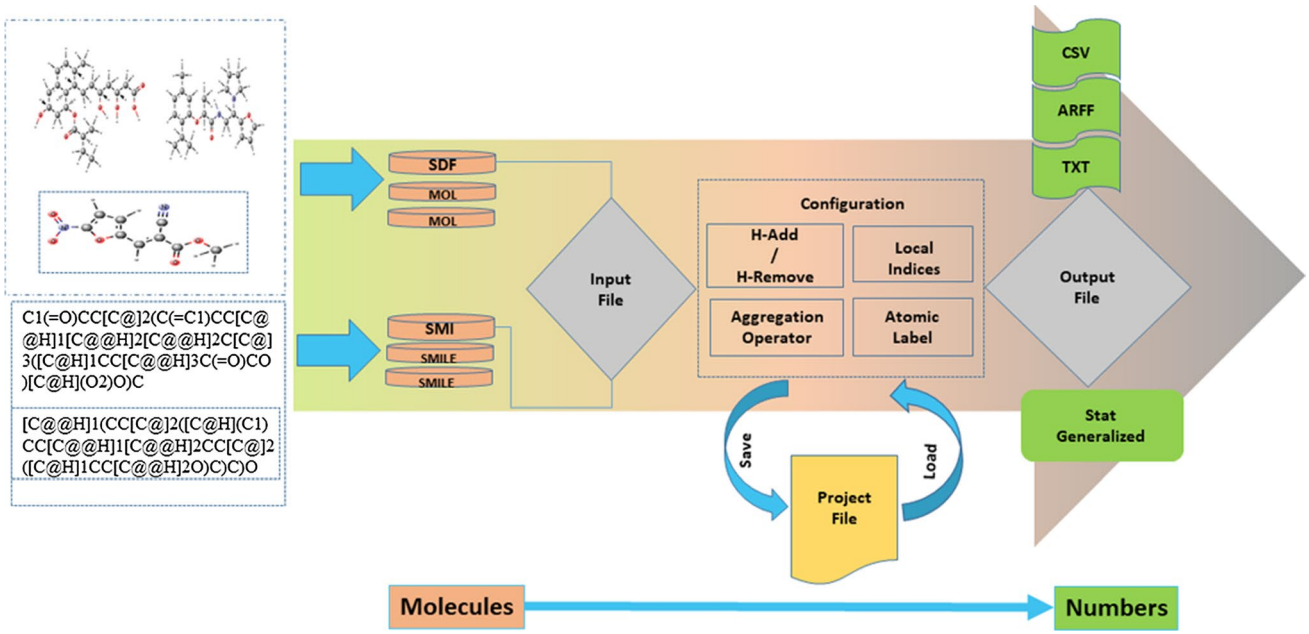


Fig. 2 General flowchart of the **MD-LOVIs** software for MDs calculation

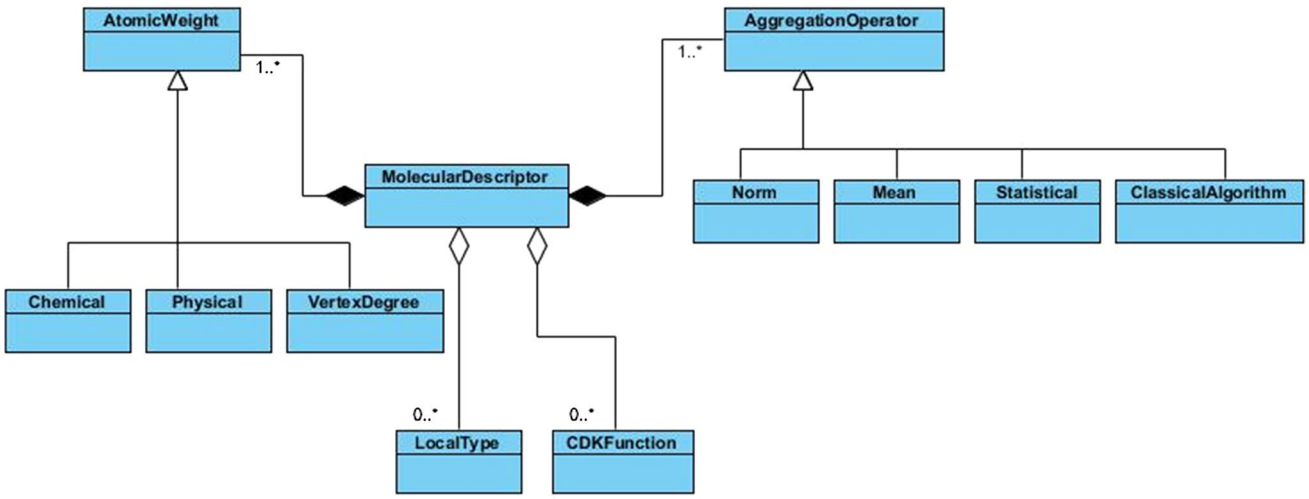


Fig. 3 UML Diagram showing the dependencies and inheritance hierarchy of fundamental classes of **MD-LOVIs** software

the classes corresponding to AW based on the physical, chemical and vertex degree properties. The ***Aggregation-Operators*** package includes the classes related to Statistics, Norms, Means and Classical Algorithm AOs. **SI-3** shows the GUI of the **MD-LOVIs** program, highlighting the different configuration parameters.

This software provides a total of 352,000 MDs and supports MOL, SDF and SMILE (and .smi) file formats as input files [50]. Moreover, this software supports CSV format (comma-separated value), TXT format (space-separated value) and ARFF (Attribute-Relation File Format) Weka file [51] as the output file formats.

Applications and experiments: what do aggregation operators have to offer in encoding chemical structural information?

Herein, we carried out four groups of experiments to evaluate the performance and utility of **MD-LOVIs** software: (i) speed experiments (using three datasets) to assess the applicability of the software in large-scale computations, (ii) entropy-based variability analysis, (iii) orthogonality analysis by means of factor analysis and (iv) applicability

in QSAR studies. Most importantly, the performance of the **MD-LOVIs** software was compared with that of commercial and open-source software in all the aforementioned experiments, in order to gain understanding on the merits and drawbacks of this software relative to the existing ones.

Descriptor calculation speed experiments

All the experiments for determining the speed of MD calculations were performed on an HP ProDesk 600 G1 TWR with Intel Inside(R) Core(TM) i7-4790 CPU @3.60 GHz 3.60 GHz processors and 16 GB RAM, with Windows 10 and 64 bit Operating System and with 1 TB of total size of HDD.

A total of 19,601 compounds distributed in three datasets (15,050 compounds from OtavaPrimScreen15 [52], 1545 compounds from NPACT [53] and 3006 compounds from DrugBank FDA-approved [54]) were used for the MD calculations (see **SI-5**). The average of speed of calculation was 0.19 s/molecule, delaying per molecule an average of 6.19 s, and the time of calculation of a descriptor was 2.261×10^{-5} s.

An additional analysis was carried out to estimate the average calculation time of the AOs used in **MD-LOVIs** software. Table 2 shows the results of the average time per molecule for the calculation of MDs using the different classes of AOs.

As may be seen, the Norms, Means and Statistical AOs have an average speed of calculation of 27 molecules/s, 27 molecules/s and 20 molecules/s, delaying for each molecule 0.043 s, 0.0447 s and 0.0557 s, respectively. Figure 4 shows the average calculation speed of the AOs (time for 100 molecules) by **MD-LOVIs** software and that of other software.

As can be seen, superior computation speeds are achieved for the Means, Norms and Statistical AOs implemented in **MD-LOVIs** (http://tomocomd.com/md-lovis) relative to other software, such as Padel version 2.4 [5], CDKDescriptor version 0.94 [9], DRAGON version 5.5 [3] and Bluecal [55]. **MD-LOVIs** is comparable with other software from the literature, since the average time for the calculation of a descriptor for 100 molecules, using the Classical (All AO) (0.002 s), Means (0.002 s), Norms (0.003 s) or Statistical (0.001 s) AOs, is lower than that for the CDKDescriptor (0.412 s), DRAGON (0.006 s), Bluecal (0.106 s) and Padel (0.031 s). *For more information, see SI01.*

Variability and degeneration

The Shannon entropy (SE)-based variability analysis was used for the internal comparison of the AOs, as well as the **MD-LOVIs** descriptors in general, with those generated from other software packages. For this study, the same three datasets (NPAC, OtavaPrimScreen15 and FDA-approved) were used to analyze the variability of the generated indices, in order to select the best descriptor types for optimal MD configuration profiles (denominated herein molecular descriptor projects) of the **MD-LOVIs** software. A total of 19,601 molecules were employed, of which 19,589 were correctly calculated, representing a 99.93% coverage.

The highest SE for this dataset is determined as equal to $\log_2 19{,}589 = 14.26$ bits ($SE = \log_2 N$, where N is the number of bins; in this case, N is equal to the number of molecules). The IMMAN software [56] was used for calculating the SE, where values equal to zero (0) correspond to maximum degeneracy, while high values are related to low degeneracy in that such descriptors are deemed to be sensitive to progressive changes in chemical structures and are thus generally suitable for correlation studies. For this study, a cutoff corresponding to 60% relative to the maximum SE (i.e., 8.55 bits, based on a 19,589 bins discretization scheme) was employed as a criterion for selecting variables with low degeneracy.

A set of four studies were carried out: the first one in order to compare only the atomic properties (AWs), the second in order to compare the AOs and a third one in order to analyze the variability of local fragment indices. Finally, an overall study was carried out to compare the **MD-LOVIs** MDs with other indices provided by other software reported in the literature.

Comparison between atomic weights (AWs)

In order to carry out this study, 96 MDs were calculated for each AW using the same descriptor configuration, and the

Table 2 Performance of executions of AOs using three datasets

OtavaPrimScreen15				NPACT			FDA-approved		
Descriptor	Time (s)	Time/mol	Mol/s	Time (s)	Time/mol	Mol/s	Time (s)	Time/mol	Mol/s
Classical (All AO)(273,680)	53,090	3.53	1	14,415	9.33	1	17,216	5.7	1
Norms (1320)	370	0.025	41	102	0.066	15	117	0.039	26
Means (2200)	397	0.026	38	97	0.063	16	125	0.042	24
Statistics (5280)	565	0.038	27	118	0.077	13	150	0.050	20

Time is given in seconds (s), speed in molecule by seconds (molecules/s) and time by molecule (s/mol)

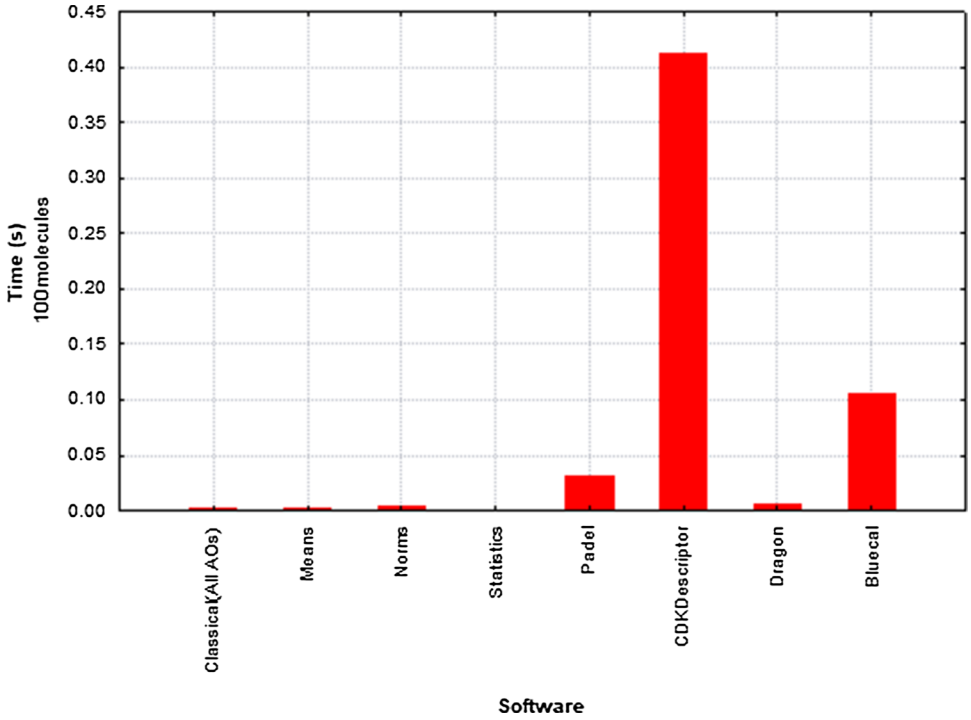


Fig. 4 Time (s) of the **MD-LOVIs** and other relevant software for calculating of DMs of 100 molecules

SE computed for each MD in the dataset. The SE distribution of each of the AWs is shown in **SI-6**.

As it can be observed, the best SE distribution was achieved for the AWs: DC and IN with SE values for all the generated variables greater than 11.80 and 12.47 bit/molecule, respectively, followed by P which had over 60% of its variables with SE values greater than 8.55 bit/molecule. The rest of the atomic labels generated MDs with only a few of them presenting SE values greater than 8.55 bit/molecule. For more information on all the results, see SI02.

In view of the results obtained in the performed studies, three groups of AW were formed: (i) AWs with high SE values (high_SE): IN, DC, P, Alk and Y, (ii) AW with medium SE values (medium_SE): S, A, Z, R, HX, LI, E, BD, M and VW and (iii) AWs with low SE (low_SE): T, I, C, L, KU and N. Each group had a total of 480 variables (size determined based on group with the lowest number of variables). For more information on all the results, see SI02.

Furthermore, the AWs were clustered into three groups according to their conceptual origin, that is, (1) *Chemical AWs*: Z, VW, P, A, R, E; (2) *physical* AWs: T, L, M, C; and (3) *VertexDegree AWs*: N, Y, S, KU, I, BD, LI, HX, Alk, IN, DC and subsequently the variability of each group analyzed. Each group comprised of 384 variables (cutoff determined by physical AW group as it possesses the least number of variables); the Shannon's entropy distribution for each group is shown in SI-6 c). As may be observed, the *vertexDregree* group presented the highest variability with all it variables yielding SE values superior to 12.03 bit/molecule, followed by the chemical group (with 60% of its variables having SE values superior to 9.23 bit/molecule), while the physical group had a least favorable SE distribution. For more information on all the results, see SI02.

In general, this study showed that the descriptors based on the AWs IN, DC, P, Alk and Y (mostly belonging to the *VertexDegree* group) possess high SE and are thus expected to yield less degenerate MDs, which could ultimately be useful in modeling tasks [56].

Comparison between aggregation operators (AOs)

The study presented herein was focused on the evaluation of the contribution in variability terms of the different aggregation operators to the **MD-LOVIs** descriptors. For this, 240 MDs were calculated for each AO using **MD-LOVIs** software and maintaining the same descriptor configuration. Subsequently, the SE was computed for these MDs and their distributions analyzed based on the corresponding AO groups. *SI-7* shows the SE values for the non-classical AOs: A, I50, R, Q3, N1, Q2, MX, Q1, P3, K, S, M, MN, P2, N2, N3, VC, SD, V and G.

As it can be seen, the A, N1, P3, M, P2, N2, N3, SD and G AOs presented the best performance with average SE values between 6 and 7 bit/molecule, while the rest of the AOs had average SE values below 6 bits. For more information on all the results, see SI03.

Based on the results of the aforementioned studies, the AOs were classified into three groups, where the first group high_SE (N2, N3, VC, SD, S, K) included the MDs with high SE values (> 12 bit/molecule), the second group

medium_SE (P3, P2, M, N1, A, G) had the indices with moderate values of SE (11–12 bit/molecule) and the third group low_SE (Q1, MN, I50, Q2, Q3, R, V) have the MDs with low values of SE (< 11 bit/molecule). *SI-7* b and *SI-7*c show the SE distribution of these groups, with each group comprising of 300 MDs. For more information on all the results, see SI03.

In general, this study showed that the non-classical AOs N2, N3, VC, SD, S and K possess high variability; these belong to the Norms and Statistical AO classes. It may be suggested that these AOs yield MDs with suitable chemical structure discriminating capacity.

Comparison between Classical Algorithms AOs For this study, 240 MDs were computed for each AO of the classical algorithms, following the same descriptor configuration for the AWs, total and local indices. *SI-8* shows the SE values for each classical AO.

As it may be observed, the AOs TS, GI, AC and CN yielded the most favorable SE distribution with averages superior to 11 bit/molecule, while SI, TI and MI yielded lower average SE values. For more information on all results, see SI04.

In order to assess the contribution of the aforementioned AOs relative to the summation (N1) operator, two sets of variables were built comprising of, on the one hand, the best 100 variables obtained with all Classical Algorithm AOs considering the N1 (sum) operator exclusively and, on the other hand, the best 100 MDs obtained with classical AOs using the rest of non-classical AOs. Posteriorly, the SE distribution of the two sets was compared as shown in *SI-8*. From this study, it was observed that MDs with higher variability are obtained when AOs other than the sum are employed, thus demonstrating the practical contribution of the proposed generalized scheme. For more information on all results, see SI04.

Comparison between local fragment types

For this study, the following configuration for the proposed local (group type) and total indices was considered: P and Y as AWs; N1, N2, P2 and K as AOs. A set of MDs were obtained for each local (group type) and the total indices; see *SI-9*.

The best MDs were global and local type's indices: GL, HA, HT, IS, CB and LA, whose SE values were above 11 bits on average. However, HL and MD had a low value of SE < 7 bit/molecule, while the RA and HD had an acceptable value. For more information on all the results, see SI05.

Comparison of MD-LOVIs with the other software

From the results of the previous studies, **MD-LOVIs** descriptor calculations were performed based on the best AWs, group types and AOs, yielding a total of 4200 MDs. Subsequently, the variability of these MDs was compared with that of indices computed with the CDKDescriptor version 0.94 [9], Bluecal [55], Mold2 [8], DRAGON version 5.5 [3], PowerMV version 02 [57] and Padel version 2.4 [5]. Figure 5 shows the Shannon's entropy distribution of each of these MDs computing software.

MD-LOVIs generate higher probabilities of containing high SE variables than the rest of the software, such as Padel, DRAGON, CDK and Mold2. Therefore, it may be deduced that the variability of **MD-LOVIs** indices is comparable to that of the most popular MD computing software.

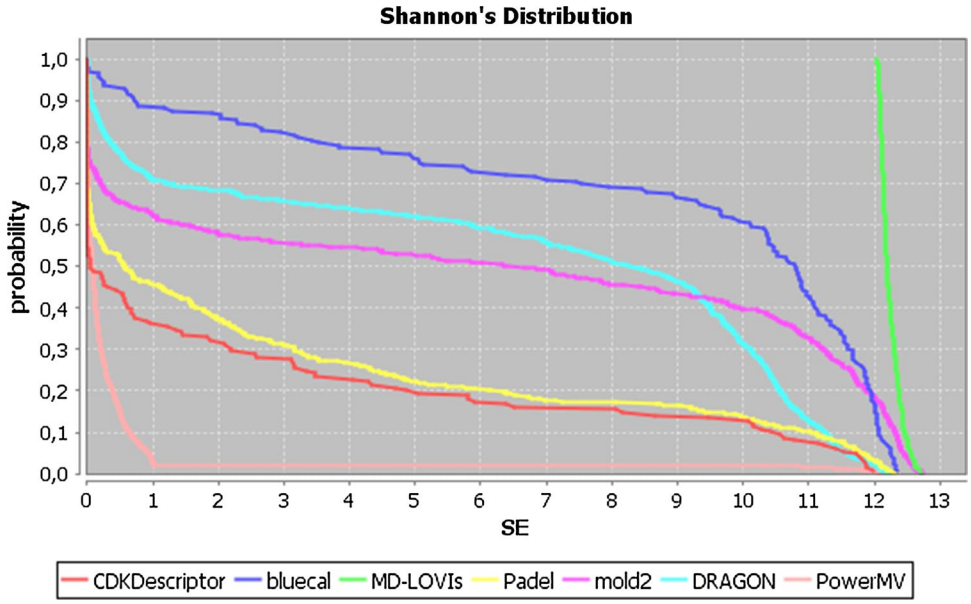


Fig. 5 Comparison among **MD-LOVIs** and other software reported in the literature taking into account SE variability analysis

This result suggests that **MD-LOVIs** software yields indices with greater information content and thus should contribute to greater modeling capacity. For more information on all results, see SI06.

Linear independence

In this section, the possible orthogonality between the MDs from **MD-LOVIs** and other software reported in the literature is examined using principal component analysis (PCA). The PCA is a mathematical technique that transforms several correlated variables into a reduced number of non-correlated variables, called *principal components* [31]. The extracted components have the following features: (1) The first component explains the highest variance of the analyzed dataset, (2) consecutive components explain the variance that previous components did not explain and (3) variables loaded in each component are linearly independent of the ones loaded in the other components. The varimax-normalized method was used as the rotation strategy and a cutoff of 0.5 or 0.7 for the factor loadings. For this study, the STATISTICA software was employed [58].

Comparison between AWs

For this experiment, groups of AWs were selected based on their classification in the preceding section and the orthogonality (linear independence) between them analyzed. The configuration of the MDs was as follows: Only *N*1 was chosen as the AO; HT and RA were chosen as the fragment types in addition to the total indices; and from each AW group (high, medium and low), representative AWs were chosen.

The high group MDs analyzed herein were loaded in four factors, collectively explaining approximately 94.46% of the total cumulative variance. Factor 1 (59.63%) possessed strong loadings for the all AWs and IN with HT; in factor 2 (24.46%), IN, Y and Alk with RA were loaded; factor 3 (8.45%) has loadings for the AWs: DC, P with RA and DC, P, Alk with HT.

The medium group was loaded in four factors, containing approximately 97.46% of the total cumulative variance. For more information on all the results, see SI07. Factor 1 (71.93%) had loadings for the AWs: Z, A, VW, M, S, BD, HX and LI with RA; in factor 2 (14.61%) was loaded the S with HT; factor 3 (7.22%) had loadings for the AWs: E, R, M, BD, HX, LI with HT and E, R with RA; and factor 4 (3.70%) was loaded the VW with HT.

The low SE group was also loaded in four factors, containing approximately 88.28% of the total cumulative variance (For more information on all the results see SI07). In factor 1 (41.87%) were loaded the following Aws: T, N, KU, with total, C, T, N, I and KU with HT; factor 2 (25.72%) had loadings for L, N, I and KU with RA; in factor 3 (11.53%) were loaded L, I with T and L with HT; and factor 4 (9.15%) had loadings for C and T with RA.

From this analysis, it is deduced that there is a degree of collinearity between the AWs of the high group and some of the medium group, but there is stronger collinearity between the medium and low groups. Nonetheless, some properties of the high group are orthogonal to the properties of the other groups. This indicates that the properties of the high group capture the same information as those of other groups, in addition to capturing orthogonal information relative to these groups. Consequently, the low and medium group would be discarded at this point.

Some of the most representative variables of each group and each factor belonging to each group (IN, P, HX, S, R, VW, N, KU, I, T) were selected for a further study conducted to analyze the linear independence between them. These variables were loaded into four factors, containing approximately 94.46% of the total cumulative variance; for more information on all the results, see SI07; in the first factor (58.70%) were loaded VW, P, S, HX and IN using total and IN with HT; the second factor (25.91%) had loadings for VW, P, N, S, HX, IN with RA; in the third factor (7.73%) were loaded the following AWs: T, N, VW, P, S, HX with HT; and in the fourth factor (2.13%) only T with RA was loaded. The AWs: IN, S, P, T were orthogonal to each other, being the most representative of each group.

Comparison between AOs

In order to perform the orthogonality study of the AOs, non-classical and classical AOs were selected and analyzed separately, as in the previous variability study (Sect. 4.4.2).

The non-classical AOs were loaded in ten factors, accounting for approximately 91.59% of the total cumulative variance. For more information on all the results, see SI08. In the first factor (33.49%) were loaded N2, N3, G, M, P2, P3, A, V, SD, R, Q1, Q2, Q3, MX as total indices; the second factor had loadings for (23.12%) N1, N2, N3, G, M, P2, P3, A, Q1, Q2, Q3, MX and MN with RA local indices; in the third factor (10.18%) were loaded N2, N3, M, P2, P3, V, SD, VC, R, Q3, I50, MX using HT local indices; in the fourth factor (5.26%) were loaded V, S, K, SD, VC and R using RA; the fifth factor (5.08%) had loadings for G, A, Q1, Q2 and MN using HT; in the sixth factor (4.79%) was loaded N1 using total and HT.

In the case of the classical AOs, these were loaded in 11 factors, explaining approximately 90.54% of the total cumulative variance (For more information on all the results see SI08); in the first factor (38.09%) were loaded the classical AOs with N1 and TS (6-T) using HT; the second factor (26.52%) had strong loadings for all AOs AC,

GI, TS, MI, TI, SI, IB, CN with N1 using RA local type indices; in the third factor (10.13%) were loaded the majority of the AC (2-T), GI, TS (2–5), MI, TI, ES using HT; factor 4 (3.66%) had loadings for 1AC, 1GI, (1–2) CN using HT; non-classical AOs were loaded in factor 5; in factor 6 (2.42%) were loaded (2-5) CN using HT; factor 7, had loadings for non-classical AOs; in factor 8 (1.58%) was loaded the total SI; non-classical AOs were loaded in factor 9; in factor 10 (1.15%) only ES using RA was loaded; and in factor 11 (1.07%) was loaded SI using HT. The most representative variables in each strongly loaded factor: 4TS as global indices, TI using RA, TGI and 1GI using HT, SI using T, ES using RA, SI using HT were thus selected to constitute the set of orthogonal descriptors. It may therefore be concluded that by applying different fragment types to these AOs, descriptors orthogonal to each other may be obtained.

Comparison between total and local fragment type MDs

For the analysis of the possible orthogonality between the local (fragment types) and the global (total) indices, the same configurations were selected for each local type (i.e., N1, N2 and P2 as the AOs, P and Y as AWs). These MDs were loaded into 17 factors, accounting for approximately 86.98% of the total cumulative variance. For more information on all results, see SI09.

F1 (35.07%) possess strong loadings for CB, GL, IS, LA, HT, MD, HD, HA and total (with P and Y as AWs) using N1, N2, P2 as non-classical AOs; F2 (12.44%) had loadings for IS, RA as fragment types using the P and Y as AWs, as well as the total and GL-type indices using Y as AW and the K as AO; in F3 (8.12%) were loaded HA, HL, HT using P and Y as AWs, and P2, N1, N2 as AOs; the fourth factor (4.78%) had loadings for CB, GL, IS, LA, as well as the total indices P as AWs and P2 as AO; F5 (4.08%) had loadings for MDs using Y and P as AWs and N1, N2, P2 as AOs; and in F6 (3.40%) only N1 AO is shown, with locals HA, HT calculated with P and Y AWs.

MD-LOVIs versus other commercial and open MDs' software

A set of 390 MDs was generated from the best configuration (best combination of total and locals, AOs and AWs) with the **MD-LOVIs** software. The MDs whose SE value <1 bit/molecule were excluded. Likewise, other MDs were calculated with DRAGON [3], Padel [5], Mold2, PowerMV [57], Bluecal [55] and CDKDescriptor [9] software and a 0.6 cutoff of the factor loadings employed.

A total of 75 factors were obtained, which explain 85.04% of the cumulative total variance, exclusive loadings are obtained for **MD-LOVIs** indices in the factors (3(3.82%), 8(2.25%), 10(1.96%), 13(1.21%), 15(1.05%), 18(0.89%), 22(0.69%), 29(0.489%), 33(0.44%), 39(0.37%), 60(0.22%), 67(0.18%)), while DRAGON indices are exclusively also loaded in the factors F9, F19, F23, F27, F34, F36, F43, F49, F53, Padel are exclusively loaded in the factors F7, F12, F16, F26, F28, F30, F37, F38, F41, F48, F50, F55, F58, F59, F65, F69, Mold2 MDs are exclusively loaded in few factors (F11 and F21), Bluecal MDs are exclusively loaded in the factors F5, F24, F25, F40, F56 and F63, CDKDescriptor variables are exclusively loaded in F19, F20 and F36, while PowerMV MDs are exclusively loaded only one factor (F44). Much of the information codified by the Mold2, CDKDescriptor, Bluecal and DRAGON MDs is equally captured in the factors F1, F2, F4, F6, F9, F17, F42, F47, which demonstrated the collinearity between the MDs from this software. It can be seen that the MDs obtained by **MD-LOVIs** are orthogonal to those of other software reported in the literature such as Padel, Mold2 and DRAGON. An important inference from this study is that **MD-LOVIs** indices codify structural information not described by indices from software reported in the literature, due to the diverse generalization schemes (AOs, weights, local indices) provided **MD-LOVIs**, which are not employed in any other software. For more information of all the results, see SI10.

Relevance in QSA(P)R studies

In order to evaluate the contribution of the indices obtained with the **MD-LOVIs** software in modeling the chemical, physicochemical and biological properties of molecules, a relevance study was performed using different benchmark datasets, such as (I) Cramer's steroid, (II) derivatives of 2-furylethylene and (III) alkyl alcohols.

Here the models were built using multiple linear regression coupled with the genetic algorithm as feature selection methodology (MLR-GA), implemented in the MobyDigs software [59]. The following configuration setup was employed: the statistical parameter Q^2_{loo} ("*leave-one-out*" cross-validation) as the optimization function, initial population size of 100 individuals and the reproduction/mutation ratio of 0.5. The validation techniques bootstrapping (Q^2_{boot}) and y-scrambling [a (Q^2)] were used to assess the models' predictive power and to verify the possibility of fortuitous correlations in the obtained QSPR models, respectively. The best model in each case was evaluated for its generalization ability using an "external validation" (Q^2ext) procedure.

Cramer's steroids

Cramer's steroid dataset consists of 31 steroids with the corresponding binding affinity to corticosteroid binding globulin (CBG) [60]. This dataset has been used in several studies to evaluate the performance of novel procedures or methodologies, proposed by several authors, by comparing them

with reported data [60]. For more information, see Support Information (SI-10). Therefore, the set can be considered as a "benchmark" for comparing molecular structure characterizing strategies.

In this study, the Cramer's steroid dataset was used to examine the local and total MDs obtained from the $\vec{W}$ and to compare the N1 invariant (sum of atom/fragment indices) with other AOs, based on their predictive capacity of the CBG binding affinity. Firstly, a comparison of all the AOs on the basis of their modeling capacity of the CGB of the Cramer's steroids was performed. For this study, all the 31 steroids are considered as training set. (The 31st steroid previously reported as an outlier in the literature was not excluded [61–63].) SI-11 shows the performance of the six-variable QSAR models, in terms of the Q^2_{loo}, built with the different AOs.

As can be observed, many mathematical AOs exhibited better performance than the N1 operator, for example, AC, IB, TS, ES, CN, TI, SI, MI and GI, with all these belonging to classical AOs group. Moreover, the following observations were taken for each group of AOs: for the Means, the quadratic mean (P2) had the best performance, while skewness (S) was the best Statistical AO and also the overall best among the non-classical AOs). As for the "Classical Algorithms," the best AO was GI using Norms, Means and Statistical AOs, respectively. Overall, the "Classical Algorithms" had better performance than the non-classical AOs (i.e., Norms, Means and Statistical), which is logical since the classical invariants additionally take into account the connectivity (topology) of atoms in the molecules. The best overall AO was the gravitational (GI), with $Q^2_{loo} = 0.963$. For more information, see Support Information (SI-12). This internal comparison showed that the global and local MDs from Norms, Mean, Statistical and "Classical Algorithms" AOs were better predictors of the Log*1/k (CGB) property* than just the sum of $\vec{W}$ components (N1); for more information, see Support Information (SI-13).

An important finding from this evaluation is the fact that AOs allows the calculation of MDs with varying performance in QSAR models, and their combinations should yield a better performance. SI-14 shows the best regression models based on 1–6 **MD-LOVIs** indices.

Generally, the obtained models are robust and not prone to chance correlation, being noticeable by the favorable cross-validation (Q^2_{loo} and Q^2_{boot}) and y-randomization [a (Q^2)] parameters, respectively.

The Cramer's dataset [60] was additionally split into training and test sets, in which 21 steroids are considered as training group and ten steroids as the prediction set.

In SI-15, the best QSARs are shown selected according the following statistical parameters: R^2, the determination coefficient, Q^2_{loo}, Q^2_{boot}, $a(Q^2)$, SDEC, the standard deviation of regression and F, the Fisher ratio.

All models comprising of one to four variables showed good predictive ability of the log $1/K$ (CBG), with Q^2_{loo} of 0.980, 0.967, 0.934 and 0.850, respectively. In this case, the AOs: AC using N3, V, S and R, ES using R, and DE were employed in the model building. The statistic parameters of these QSARs suggest that the indices from **MD-LOVIs** codify important chemical structural information, useful in the modeling of bioactivity endpoints.

2-Furylethylene derivatives

Modeling specific rate constants (log k). In order to evaluate the applicability of this new approach in the QSR(Reactivity) R studies, we selected a dataset of 34 derivatives of 2-furylethylene, for which the corresponding specific rate constants (log k) and partition coefficients (log P) have been reported [64]. The best log k models obtained using these $\vec{W}$ as MDs, together with their respective statistical parameters, are given in *SI-16.*

Modeling partition coefficients (log P). The partition coefficient n-octanol/water (log P) has an important role in the understanding of the biological behavior of these 2-furylethylene derivatives [64]. The best log P models obtained using the MD-LOVIs MDs, together with the corresponding statistical parameters, are given in SI-16.

As observed in Table 6, the **MD-LOVIs** yield models with good statistical quality, providing further confidence in the possible applicability of these AOs in QSPR modeling tasks.

Modeling boiling point (Bp) of 28 Alkyl Alcohols

This dataset comprises 28 alkyl alcohols (14 are primary, 6 secondary and 8 tertiary) for which the boiling point (Bp) has been previously reported [65]. The best MLR models obtained to describe the Bp of these compounds using **MD-LOVIs** are given in *SI-17.*

In general, most models presented herein include Classical Algorithms and Atatistical AOs (i.e., AC, ES, TS, S, V and R), AWs based on the IN, L and KU, local fragments IS, LA, HT, CB and HA, as well as the total indices (see SI-18). This result suggests that combining different MD-LOVIs parameter configurations yields information-rich descriptors useful in QSPR modeling.

It may therefore be concluded that the incorporation of the aforementioned generalization scheme in **MD-LOVIs** improves the performance of the derived global and/or local descriptors in modeling different endpoints, thus demonstrating the practical contribution of this framework.

Analysis for QSA(P)R comparative studies

Here, a comparison of the performance of the **MD-LOVIs** approach as a whole with respect to other software and approaches reported in the literature is conducted. In comparison with other approaches reported in the literature for Crammer's steroids dataset, it is observed that the **MD-LOVIs** indices yield better performance than the best model reported up to now, based on the combined electrostatic and shape similarity matrix (CESSM) method [66]; see supporting information SI-19, even with a much lower degree of freedom [i.e., Q^2_{loo} (**MD-LOVIs**) = 0.952, Q^2_{loo} (MD-LOVIs) = 0.958, Q^2_{loo} (MD-LOVIs) = 0.964 for 4, 5 and 6 variables, respectively, compared to Q^2_{loo} (CESSM) = 0.941 obtained with 6 variables]. The CESSM method employs the neural networks as the fitting method, which is known to yield better optimized models compared to other traditional regression procedures such as MLR, principal component regression or multiple logistic regression. Nonetheless, better statistical parameters are obtained with the **MD-LOVIs** approach using a much simpler technique, demonstrating the theoretical robustness of the proposed strategy, notwithstanding its simplicity (see Supporting Information SI-19).

In addition, the two size variable **MD-LOVIs** models computed on H-filled and H-suppressed Gs yielded superior performance ($Q^2_{loo} = 0.881$ and $Q^2_{loo} = 0.853$, respectively) than many other methods, whose models were constructed with a greater degree of freedom, for example, TQSI ($Q^2_{loo} = 0.848$), CoMMA($Q^2_{loo} = 0.828$) and SOMFA ($Q^2_{loo} = 0.74$).

The good performance of the **MD-LOVIs** indices in QSAR modeling supports the hypothesis that global definitions of chemical behavior from atomic characteristics may not after all imply their linear combinations, and therefore, other relations should also be contemplated. Additionally, the **MD-LOVIs** indices offer a remarkable advantage over most 3D techniques due to their inherent simplicity, in which procedures such as molecular alignment and structural optimization, which usually result in the inapplicability of some methods to structurally diverse datasets, are avoided [67]. SI-20 shows the statistical parameters of the different approaches reported in the literature.

The models obtained by **MD-LOVIs** (with three and four variables) with Q^2_{loo} of 0.980 and 0.967, respectively, have superior performances than the corresponding model based on kNN-MFA/simulated annealing (SA) [68], with Q^2_{loo} of 0.95, for six statistical variables, which is considered as the best model reported in the literature. Moreover, the model from **MD-LOVIs** (with two variables and $Q^2_{loo} = 0.934$) has better performance than some model such as kNN-MFA/genetic algorithm (GA) [68] (with eight variables and $Q^2_{loo} = 0.93$), CoMFA [69] (with two principal components and $Q^2_{loo} = 0.903$), G-WHIM [70] (with two variables and 0.893) and CoMSA [71] (with one principal component and $Q^2_{loo} = 0.88$). These methods are more complex than **MD-LOVIs** and mostly nonlinear in nature. The **MD-LOVIs** strategy offers a remarkable advantage over most complex techniques due to its inherent simplicity. This result suggests that indices from **MD-LOVIs** codify relevant chemical information useful in correlation studies despite their simplicity and low computational cost.

Moreover, a comparison of the results obtained with the **MD-LOVIs** indices in modeling the boiling point of alkyl alcohols, as well as the physicochemical properties of the 2-furylethylene derivatives, was made. The results of the respective global predictions obtained with all other approaches are shown in SI-21.

The models obtained by the **MD-LOVIs** indices also show comparable performance with respect to recently published DIVATI [20] and GT-STAF [72] results in the case of logP and log k, and they show superior performance than rest MD families reported in the literature such as linear indices, local spectral moments [92] and quantum chemical descriptors [73], despite their simplicity. It can be suggesting that indices of **MD-LOVIs** could be a promissory tool in the QSAR/QSPR studies. This conclusion is consistent with two recent studies, where **MD-LOVIs** software was used in comparative QSAR studies, based on Sutherland's [74] and benzene derivative datasets [75], respectively. In the first study, the models obtained with the **MD-LOVIs** descriptors using the MLR-GA approach were superior to CoMFA [60], CoMSIA [76], Hologram QSAR (HQSAR), QSAR by eigenvalues analysis (EVA), back-propagation feed-forward neural network implemented in Cerius2 using 2.5D descriptors (NN 2.5D) and ensemble neural network (NN-ens) using 2.5D descriptors by Sutherland et al. [77]. Also, MD-LOVIs descriptors were successfully used to obtain models for predicting the structure–toxicity relationships of benzene derivatives, comparing favorably with previously published models that used the same dataset, such as CoMSIA [76], CoMFA [60], VolSurf, ERM-VolSurf and Quantum chemical descriptors [78].

Concluding remarks

In this report, a set of AOs was introduced as a generalization of the sum of elements in an atomic weights vector, $\vec{W}$. These AOs are used as invariants and are classified in Norms, Means, Statistical AOs and "Classical Algorithms," respectively. To facilitate the use of the proposed AOs in QSAR modeling, these were implemented in the **MD-LOVIs** software, a free and user-friendly platform for the computation of descriptors. This software was implemented in the Java programming language and can thus be run on any operating system (Linux,

Windows, MacOS). Moreover, the **MD-LOVIs** software allows computation of total and/or local indices and the addition or removal of H atoms in molecular structures.

The variability and principal component analysis of the **MD-LOVIs** indices demonstrated that the proposed generalizations yield MDs with superior variability compared to the MDs based on the summation operator, as well as other indices reported in the literature, and capture chemical information not codified by the MDs derived from the sum of the atomic indices as well as those of other software in general, such as DRAGON [3], Padel [5] and Mold2 [8].

Also, the indices generated by **MD-LOVIs** software are used in the modeling of diverse properties of organic compounds, yielding superior statistical parameters compared to other strategies reported in the literature and demonstrating the usefulness of this approach in codifying relevant chemical structural information, despite its simplicity and low computational cost. Moreover, the QSPR studies demonstrate that better predictors may be obtained with other AOs other than the summation operator (N1 in our scheme). Finally, it is important to note that the proposed generalization scheme may be applied to any family of MDs defined at the atomic or bond level and obtained diverse MDs.

Altogether, these results suggest that **MD-LOVIs** software is a promissory tool for use in cheminformatics studies. This software, as well as predefined configurations of descriptors (in form of batch projects) considered to be orthogonal and the most informative, is freely available online at http://tomocomd.com/md-lovis. All studies performed herein demonstrated that the **MD-LOVIs** software generates indices as simple as possible, but not simpler; by using the AOs, the diversity of the codified chemical information is enhanced and the performance of the obtained indices in QSA(P)R modeling is improved. This fact reminds us of the phrase stated by Albert Einstein that "Everything Should Be Made as Simple as Possible, But Not Simpler."

Supplementary information available

The value of multiple linear regression models and the **MD-LOVIs** molecular descriptor sets used in all studies, PCA results, speed test data and Shannon entropy-based variability analysis are freely available via the Internet at http://static-content.springer.com/esm/.

Acknowledgements Yoan Martínez-López thanks the program International Investigator Invited for a postdoctoral fellowship to work at USFQ in 2019. Yovani Marrero-Ponce acknowledges the support from USFQ "Chancellor Grant 2018 (Project ID13525)."

Funding This work was partially supported from the USFQ (Project ID13525 "Chancellor Grant 2018").

Availability of Software and Material The MD-LOVIs software and the respective user manual are freely available online at http://tomocomd.com/md-lovis.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

1. Todeschini R, Consoni V (2009) Handbook of molecular descriptors. Wiley VCH, Weinheim
2. Mani-Varnosfaderani A, Neiband MS, Benvidi A (2019) Identification of molecular features necessary for selective inhibition of B cell lymphoma proteins using machine learning techniques. Mol Divers 23(1):55–73
3. DRAGON for Windows (software for molecular descriptor calculations) (2005)
4. CODESSA 2.13. Semichem edn, 7204 Mullen, Shawnee, KS 66216, USA
5. Yap CW (2010) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474. https://doi.org/10.1002/jcc.21707
6. García-Jacas CR, Marrero-Ponce Y, Acevedo-Martínez L, Barigye SJ, Valdés-Martiní JR, Contreras-Torres E (2014) QuBiLS-MIDAS: a parallel free-software for molecular descriptors computation based on multilinear algebraic maps. J Comput Chem 35(18):1395–1409
7. Valdés-Martiní JR, Marrero-Ponce Y, García-Jacas CR, Martinez-Mayorga K, Barigye SJ, d'Almeida YSV YSV, Pérez-Giménez F, Morell CA (2017) QuBiLS-MAS, open source multi-platform software for atom-and bond-based topological (2D) and chiral (2.5 D) algebraic molecular descriptors computations. J Cheminform 9(1):35
8. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 48(7):1337–1344
9. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics. Curr Pharm Des 12(17):2111–2120
10. Dong J, Cao D-S, Miao H-Y, Liu S, Deng B-C, Yun Y-H, Wang N-N, Lu A-P, Zeng W-B, Chen AF (2015) ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform 7(1):60
11. Gutman I, Das KC (2004) The first Zagreb indices 30 years after. MATCH Commun Math Comput Chem 50:83–92
12. Randic M (1975) Characterization of molecular branching. J Am Chem Soc 97(23):6609–6615
13. Broto P, Moreau G, Vandycke C (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies, autocorrelation descriptor. Eur J Med Chem 19:66–70
14. Katritzky AR, Lobanov VS, Karelson M, Murugan R, Grendze MP, Toomey JEJ (1996) Comprehensive descriptors for structural and statistical analysis. 1. Correlations between structure and physical properties of substituted pyridines. Rev Roum Chim 41(85):81–867
15. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth

16. Zhao YH, Abraham MH, Zissimos AM (2003) Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. J Org Chem 68(19):7368–7373
17. Wolpert D, Macready W (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82
18. Barigye SJ, Marrero-Ponce Y, Martínez Santiago O, Martínez López Y, Torrens F (2013) Shannon's, mutual, conditional and joint entropy-based information indices. Generalization of global indices defined from local vertex invariants. Curr Comput-Aided Drug Des 9(2):164–183
19. García-Jacas CR, Cabrera-Leyva L, Marrero-Ponce Y, Suárez-Lezcano J, Cortés-Guzmán F, García-González LA (2018) GOWAWA aggregation operator-based global molecular characterizations: weighting atom/bond contributions (LOVIs/LOEIs) according to their influence in the molecular encoding. Mol Inform 37(12):1800039
20. Martínez-Santiago O, Millán-Cabrera R, Marrero-Ponce Y, Barigye SJ, Martínez-López Y, Torrens F, Pérez-Giménez F (2014) Discrete derivatives for atom-pairs as a novel graph-theoretical invariant for generating new molecular descriptors: orthogonality, interpretation and QSARs/QSPRs on benchmark databases. Mol Inform 33(5):343–368
21. Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57(3):519–530
22. Fleming PJ, Wallace JJ (1986) How not to lie with statistics: the correct way to summarize benchmark results. Commun ACM 29(3):218–221
23. Calvo T, Mayor G, Mesiar R (2012) Aggregation operators: new trends and applications, vol 97. Physica, Heidelberg
24. Merigó JM, Palacios-Marqués D, Soto-Acosta P (2017) Distance measures, weighted averages, OWA operators and Bonferroni means. Appl Soft Comput 50:356–366
25. Karczmarek P, Kiersztyn A, Pedrycz W (2018) Generalized Choquet integral for face recognition. Int J Fuzzy Syst 20(3):1047–1055
26. Wang Z, Yang R, Leung K (2010) Nonlinear integrals and their applications in data mining. In: Advances in fuzzy systems—applications and theory, vol 24. https://doi.org/10.1142/9789812814685_0001
27. Liu B, Fu M, Zhang S, Xue B, Zhou Q, Zhang S (2018) An interval-valued 2-tuple linguistic group decision-making model based on the Choquet integral operator. Int J Inf Sci 49(2):407–424
28. Fontaine F, Pastor M, Gutiérrez-de-Terán H, Lozano JJ, Sanz F (2003) Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries. Mol Divers 6(2):135–147
29. Maldonado AG, Doucet JP, Petitjean M, Fan BT (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10(1):39–79
30. Bajorath J (2017) Molecular similarity concepts for informatics applications. In: Keith J (ed) Bioinformatics. Springer, Berlin, pp 231–245
31. Marrero-Ponce Y (2004) Linear Indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. J Chem Inf Comput Sci 44(6):2010–2026. https://doi.org/10.1021/ci049950k
32. Basak S, Gute B (1997) Characterization of molecular structures using topological indices. SAR QSAR Environ Res 7(1–4):1–21
33. Merigó JM, Gil-Lafuente AM (2010) New decision-making techniques and their application in the selection of financial products. Inf Sci 180(11):2085–2094
34. Xu ZS (2012) Fuzzy ordered weighted distances. Fuzzy Optim Decis Making 11:73–97
35. García-Jacas CR, Cabrera-Leyva L, Marrero-Ponce Y, Suárez-Lezcano J, Cortés-Guzmán F, Pupo-Meriño M, Vivas-Reyes R (2018) Choquet integral-based fuzzy molecular characterizations: when global definitions are computed from the dependency among atom/bond contributions (LOVIs/LOEIs). J Cheminform 10(1):51
36. Bolton J, Gader P, Wilson JN (2008) Discrete Choquet integral as a distance metric. IEEE Trans Fuzzy Syst 16(4):1107–1110
37. Merigó JM (2011) A unified model between the weighted average and the induced OWA operator. Expert Syst Appl 38(9):11560–11572
38. Ertl P, Rohde B, Selzer P (2000) 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 43(20):3714–3717
39. Ghose AK, Crippen GM (1987) Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci 27(1):21–35
40. Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen EL (2003) The chemistry development kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43(2):493–500
41. Kier LB, Hall LH (1999) Molecular structure description. The electrotopological state. Academic Press, New York
42. Kier LB, Hall LH (1990) An electrotopological-state index for atoms in molecules. Pharm Res 7(8):801–807
43. Harary F, Palmer E, Robinson R, Read R (1976) In: Balaban AT (ed) Chemical applications of graph theory. Academic Press, London, p 25
44. Kupchik EJ (1988) Structure—molar refraction relationships of alkylgermanes using molecular connectivity. Quant Struct-Act Relat 7(2):57–59
45. Hu Q-N, Liang Y-Z, Yin H, Peng X-L, Fang K-T (2004) Structural interpretation of the topological index. 2. The molecular connectivity index, the kappa index, and the atom-type E-State index. J Chem Inf Comput Sci 44:1193–1201
46. Beliakov G (2003) How to build aggregation operators from data. Int J Intell Syst 18:903–923
47. Alikhanidi S, Takahash Y (2006) New molecular fragmental descriptors and their application to the prediction of fish toxicity. MATCH Commun Math Comput Chem 55:205–232
48. Ivanciuc O (1989) Design on topological indices. 1. Definition of a vertex topological index in the case of 4-trees. Revue Roumaine de Chimie 34(6):1361–1368
49. Visual Paradigm 8.0 for UML Enterprise (2010). 8.0 edn
50. (MDL Information Systems). http://en.wikipedia.org/wiki/MDL_Information_Systems. Accessed Jan 2019
51. Holmes G, Donkin A (1994) Witten IH Weka: a machine learning workbench. In: 2nd Australian and New Zealand conference on intelligent information systems, Brisbane, Australia, vol 357–361
52. OTAVA L (2019) OTAVA chemicals. https://www.otavachemicals.com/products/compound-libraries-for-hts/diversity-sets. Accessed Jan 2019
53. Mangal M, Sagar P, Singh H, Raghava GP, Agarwal SM (2013) NPACT: naturally occurring plant-based anti-cancer compound-activity-target database. Nucleic Acids Res 41(D1):D1124–D1129. https://doi.org/10.1093/nar/gks1047
54. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46(D1):D1074–D1082
55. Georg H (2008) BlueDesc-molecular descriptor calculator. University of Tübingen, Tübingen

56. Urias RWP, Barigye SJ, Marrero-Ponce Y, García-Jacas CR, Valdes-Martiní JR, Perez-Gimenez F (2015) IMMAN: free software for information theory-based chemometric analysis. Mol Divers 19(2):305–319
57. Liu K, Feng J, Young SS (2005) PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J Chem Inf Model 45(2):515–522
58. STATISTICA version. 6.0 (2001). Statsoft, I., Tulsa
59. Todeschini R, Consonni V, Mauri A, Pavan M (2003) MobyDigs: software for regression and classification models by genetic algorithms. In: Leardi R (ed) Data handling in science and technology, vol 23. Elsevier, Amsterdam, pp 141–167
60. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. JACS 110(18):5959–5967
61. Tuppurainen K, Viisas M, Peräkylä M, Laatikainen R (2004) Ligand intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset. JCAMD 18:175–187
62. Coats EA (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. Perspect Drug Discov Des 12–14:199–213
63. Hodge VJ, Austin J (2004) A Survey of outlier detection methodologies. Artif Intell Rev 22:85–126
64. Moldovan CD, Diudea MV, Costescu A, Katona G (2008) Application to QSAR studies of 2-furylethylene derivatives. J Math Chem 45(2):442
65. Estrada E, Molina E (2001) Novel local (fragment-based) topological molecular descriptors for QSPR/QSAR and molecular design. J Mol Graphics Model 20(1):54–64
66. Aires-de-Sousa J, Gasteiger J, Gutman I, Vidovic D (2004) Chirality codes and molecular structure. J Chem Inf Comput Sci 44:831–836
67. Damale MG, Harke SN, Kalam Khan FA, Shinde DB, Sangshetti JN (2014) Recent advances in multidimensional QSAR (4D-6D): a critical review. Mini Rev Med Chem 14(1):35–55
68. Abraham B (ed) (1998) Quality improvement through statistical methods. Statistics for industry and technology. Birkhäuser, Boston
69. MACCS Drug Data Report (2000). MDL Information Systems, Inc. 14600 Catalina Street, San Leandro, CA 94577
70. Cosentino U, Moro G, Bonalumi D, Bonati L, Lasagni M, Todeschini R, Pitea D (2000) A combined use of global and local approaches in 3D-QSAR. Chemom Intell Lab Syst 52:183–194
71. Alcalá-Fdez J, Sánchez L, García S, Jesus MJd, Ventura S, Garrell JM, Otero J, Romero C, Bacardit J, Rivas VM, Fernández JC, Herrera F (2009) KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput 13:307–318
72. Barigye SJ, Marrero-Ponce Y, Martínez López Y, Martínez Santiago O, Torrens F, García Domenech R, Galvez J (2013) Event-based criteria in GT-STAF information indices: theory, exploratory diversity analysis and QSPR applications. SAR QSAR Environ Res 24:3–34
73. Estrada E, Molina E (2001) 3D conectivity indices in QSPR/QSAR studies. J Chem Inf Comput Sci 41:791–797
74. Martinez-Lopez Y, Caballero Y, Barigye SJ, Marrero-Ponce Y, Millan-Cabrera R, Madera J, Castillo-Garit JA (2017) State of the art review and report of new tool for drug discovery. Curr Top Med Chem 17(26):2957–2976
75. Martínez-López Y, Barigye SJ, Martínez-Santiago O, Marrero-Ponce Y, Green J, Castillo-Garit JA (2017) Prediction of aquatic toxicity of benzene derivatives using molecular descriptor from atomic weighted vectors. Environ Toxicol Pharmacol 56:314–321
76. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
77. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure—activity relationships. J Med Chem 47(22):5541–5554
78. Salahinejad M, Ghasemi JB (2014) 3D-QSAR studies on the toxicity of substituted benzenes to Tetrahymena pyriformis: coMFA, CoMSIA and VolSurf approaches. Ecotoxicol Environ Safety 105:128–134