


RESEARCH ARTICLE

Open Access



Choquet integral-based fuzzy molecular characterizations: when global definitions are computed from the dependency among atom/bond contributions (LOVIs/LOEIs)

César R. García-Jacas^{1*} , Lisset Cabrera-Leyva², Yovani Marrero-Ponce^{3,4}, José Suárez-Lezcano⁵, Fernando Cortés-Guzmán¹, Mario Pupo-Meriño⁶ and Ricardo Vivas-Reyes^{7,8}

Abstract

Background: Several topological (2D) and geometric (3D) molecular descriptors (MDs) are calculated from local vertex/edge invariants (LOVIs/LOEIs) by performing an aggregation process. To this end, norm-, mean- and statistic-based (non-fuzzy) operators are used, under the assumption that LOVIs/LOEIs are independent (orthogonal) values of one another. These operators are based on additive and/or linear measures and, consequently, they cannot be used to encode information from interrelated criteria. Thus, as LOVIs/LOEIs are not orthogonal values, then non-additive (fuzzy) measures can be used to encode the interrelation among them.

Results: General approaches to compute fuzzy 2D/3D-MDs from the contribution of each atom (LOVIs) or covalent bond (LOEIs) within a molecule are proposed, by using the Choquet integral as fuzzy aggregation operator. The Choquet integral-based operator is rather different from the other operators often used for the 2D/3D-MDs calculation. It performs a reordering step to fuse the LOVIs/LOEIs according to their magnitudes and, in addition, it considers the interrelation among them through a fuzzy measure. With this operator, fuzzy definitions can be derived from traditional or recent MDs; for instance, fuzzy Randic-like connectivity indices, fuzzy Balaban-like indices, fuzzy Kier–Hall connectivity indices, among others. To demonstrate the feasibility of using this operator, the QuBiLS-MIDAS 3D-MDs were used as study case and, as a result, a module was built into the corresponding software to compute them (<http://tomocomd.com/qubils-midas>). Thus, it is the only software reported in the literature that can be employed to determine Choquet integral-based fuzzy MDs. Moreover, regression models were created on eight chemical datasets. In this way, a comparison between the results achieved by the models based on the non-fuzzy QuBiLS-MIDAS 3D-MDs with regard to the ones achieved by the models based on the fuzzy QuBiLS-MIDAS 3D-MDs was made. As a result, the models built with the fuzzy QuBiLS-MIDAS 3D-MDs achieved the best performance, which was statistically corroborated through the Wilcoxon signed-rank test.

Conclusions: All in all, it can be concluded that the Choquet integral constitutes a prominent alternative to compute fuzzy 2D/3D-MDs from LOVIs/LOEIs. In this way, better characterizations of the compounds can be obtained, which will be ultimately useful in enhancing the modelling ability of existing traditional 2D/3D-MDs.

Keywords: Aggregation operators, Choquet integral, Fuzzy measures, LOVIs, LOEIs, Molecular descriptors, ToMoCoMD-CARDD software, QuBiLS-MIDAS molecular descriptors, QSAR

*Correspondence: cesarrijacas1985@gmail.com

¹ Instituto de Química, Universidad Nacional Autónoma de México (UNAM), Ciudad de México, México

Full list of author information is available at the end of the article



Introduction

In several application areas, mainly in the multi-criteria decision-making, the information aggregation process is the main step to perform [1–3]. In such a process, the individual criteria are combined into a single value (global criterion), in such a way that all properties contained in each individual criterion are included or reflected in the global criterion, by using an aggregation operator [4, 5]. Thus, several aggregation operators may be used to obtain different global criteria. In this way, decision-makers could consider diversity of criteria with the purpose of making the best final decision. Traditional aggregation operators, where individual criteria are considered as values independent of one another, are those most frequently employed (e.g. OWA-like functions [6–8]). These operators are based on linear and/or additive measures and, thus, they are not suitable to deal with the dependency among criteria.

The dependency or interaction among criteria is an intrinsic feature present in the decision-making tasks in the real world. For example, if various work teams are analyzed to select the one with the best teamwork, and with this purpose the efficiency of each worker belonging to a same team is measured, then the efficiency of each team to do teamwork is not the sum of the individual efficiencies, but the interaction among its workers to achieve the best teamwork. Therefore, it is most suitable to use non-additive measure-based operators, instead of traditional operators, for an approximate modeling of people's assessment practices. In this sense, the concept of *fuzzy (non-additive) measure*, also known as *capacity*, was introduced by Sugeno, in order to model the importance of a coalition within a set of interrelated criteria [9].

According to Lebesgue's philosophy [10], once a measure is defined, it is possible to obtain an integral with regard to that measure. Thus, associated with the concept of *fuzzy measure*, there is the concept of *fuzzy integral* [11, 12], being the Choquet integral one of the most popular [13–17]. The Choquet integral constitutes a generalization of the Lebesgue integral [10], as well as of other traditional operators (e.g. OWA-like functions), due to the fact that they coincide when the measure used is additive. The Choquet integral has been successfully used in several applications, such as: face recognition [18], rule-based systems [19], data mining [20] and decision-making [21–23]. The success of the Choquet integral as aggregation operator is due to, as already pointed out, its ability of including dependency among criteria by means of a *fuzzy measure* [14, 15].

One of the applications of the aggregation operators is in the chemical structures encoding. This process constitutes an essential step to perform several studies in the cheminformatics field, such as molecular similarity [24] and quantitative structure–activity relationships (QSAR) [25–27]. The codification of chemical structures

is performed by means of the molecular descriptors (MDs) calculation. The MDs are values computed from symbolic molecular representations, by applying different mathematical transformations [28] based on a wide variety of theories, such as quantum chemistry [29] and information theory [30]. The MDs are useful values in the sense that they can contribute to obtain a better comprehension on the interpretation of molecular properties, and/or they can integrate a model to predict biological activities in novel compounds [28].

As it can be seen in [28], several procedures to determine MDs are based on the calculation of Local Vertex Invariants (LOVIs) or Local Edge Invariants (LOEIs). These procedures perform an aggregation process on the LOVIs/LOEIs computed to determine the final value (MD) that characterizes the molecular structure (e.g. Randić–Razinger index [31] and local Balaban index [32]). The LOVIs and LOEIs depict each atom (vertex) and covalent bond (edge) of a molecule, respectively. They are computed from graph-based molecular representations without depending on any atom/bond numbering, nor on the rotation and translation of the molecules. LOVIs/LOEIs are represented into n -dimensional vectors, where n denotes the total number of atoms/bonds. The summation, summation of squares, min and max are the operators often used to obtain global MDs from LOVIs/LOEIs, being the summation operator the one most commonly used.

However, as it has already been pointed out, the use of different aggregation operators yields diversity of global criteria, that is why decision-makers consider several alternatives to make the best final possible decision. Thus, if the MDs whose calculation is based on the aggregation of atom/bond contributions (LOVIs/LOEIs) are determined using several operators, then diversity of global characterizations of the molecules can be computed. For instance, if on the well-known LOVIs called vertex degree, the kurtosis function and the traditional OWA operator are applied, then an aggregation indicating the tailedness of this LOVIs vector, and a weighted aggregation, giving more importance to the vertices (atoms) with the highest degrees, can be calculated, respectively. Both examples are quite different from the common use of the summation operator.

Inspired on this idea, recent strategies to compute topological (2D) and geometric (3D) MDs from atom/bond contributions (LOVIs/LOEIs) have been introduced [33–38]. These 2D/3D-MDs employ aggregation operators based on Minkowski norms (e.g. Euclidean norm), central tendency statistics (e.g. arithmetic mean) and dispersion statistics (e.g. kurtosis). As it has been confirmed elsewhere [39–41], the use of these operators contribute to obtain global 2D/3D-MDs with better information content (variability) and linear independence (orthogonality) than other 2D/3D-MDs reported in the literature. In addition, the diversity of the

2D/3D-MDs computed with these operators allowed to achieve successful outcomes in comparative modeling tasks [42, 43], as well as in several practical applications [44–46].

Nonetheless, up to date, the global 2D/3D-MDs computation from atom/bond contributions (LOVIs/LOEIs) is based on additive operators, under the assumption that these contributions are non-interrelated values. However, it is well-known that the biological activities or properties of the compounds do not only depend on the molecular shape, but also on the interactions that are often non-covalent in nature. Thus, non-additive (fuzzy) measure-based aggregation operators (e.g. Choquet integral) may be used, with the purpose of obtaining an approximate characterization of the interrelation that each atom (or bond) has, regarding the other ones. In this way, 2D/3D-MDs orthogonal to the other existing ones may be obtained, because of the fuzzy basis of their computations.

To the best of our knowledge, only two fuzzy MD types have been introduced: (1) by using pharmacophore-based molecular similarity [47], and (2) by using the number of interposed bonds as the measure of separation among atoms depicting pharmacophore kinds (2D-FPT MDs) [48, 49]. Therefore, fuzzy 2D/3D-MDs computed through an aggregation process on atom/bond contributions (LOVIs/LOEIs) have not been reported to date. Consequently, this work is aimed at introducing a different way for the global 2D/3D-MDs computation from LOVIs/LOEIs, by using the Choquet integral as fuzzy aggregation operator. This report is planned as follows. Second section defines some concepts regarding the fuzzy measures and the Choquet integral. Third section presents the adaptation of several procedures to compute fuzzy MDs. Fourth section presents a practical example. Fifth section studies the feasibility of using this approach. Last section describes the main findings and conclusions.

Background of fuzzy measures and Choquet integral

Definition of fuzzy measure and singleton measure.

$L_{m\delta}$ -measure: fuzzy measure composed of maximized L-measure and Delta-measure

The fuzzy measures (or capacities) are functions that determine a weight considering the interrelation (or

dependency) among criteria within a subset [11, 14]. Formally, let a universal set $X = \{x_1, x_2, \dots, x_N\}$ and $P(X)$ be the power set of X , $P(X) = 2^N$, then a fuzzy measure [9] or capacity [13] on X is a set function $\mu: P(X) \rightarrow [0, 1]$ that fulfils the following axioms:

1. $\mu(\emptyset) = 0$ (lower boundary condition).
2. $\mu(X) = 1$ (upper boundary condition).
3. If $A, B \in P(X) \wedge A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$ (monotonicity).

Therefore, for any $A \subseteq X$, $\mu(A)$ can be considered as the degree of importance (or weight) of the combination A of criteria. If $|A| = 1$, then $\mu(A) = \mu(x_i)$, and it constitutes the traditional weight when element x_i is considered separately. It is important to highlight that $\mu(x_i)$ is denominated as fuzzy density or singleton measure, denoted as $s(x_i)$, when any $A \subseteq X$ has a single element x_i . Moreover, a fuzzy measure is additive if $\mu(A \cup B) = \mu(A) + \mu(B)$, whenever $A \cap B = \emptyset$. Thus, it is enough to determine $\forall x_i \in X$ the corresponding $s(x_i)$ to define the measure completely. Other important properties of the fuzzy measures are the superadditivity and subadditivity. The former indicates high synergy or cooperative action among the criteria of a set, while the latter expresses the opposite. So, the additivity can be interpreted as the no interaction among the criteria of a set.

Several fuzzy measures have been reported in the literature, such as the Sugeno λ -measure [9] (that was the first one proposed), the P-measure [50], the Shapley values [51], the k -order fuzzy measure [52], among others [53–55]. The λ -measure and P-measure are among the most widely employed. The λ -measure is not a closed form [9], whereas the P-measure is not sensitive enough, because it only determines the max value of the input set [50] (see Additional file 1). In addition, when the number of criteria is large, then the computation of the λ parameter is quite complex in the λ -measure, because a polynomial equation of higher order must be resolved. In order to tackle these drawbacks, a fuzzy measure comprised of the Maximized L-measure (L_m -measure) [56, 57] and Delta-measure (δ -measure) [58], denoted as $L_{m\delta}$ -measure, was proposed by Liu et al. [54, 59].

Formally, a fuzzy $L_{m\delta}$ -measure, $g_{L_{m\delta}}$, on finite set $X = \{x_1, x_2, \dots, x_N\}$ is defined as follows:

$$g_{L_{m\delta}}(A) = \begin{cases} \max_{x \in A} s(x) & L = -1 \\ \frac{(1+L) \sum_{x \in A} s(x) [1 + L \max_{x \in A} s(x)]}{1 + L \sum_{x \in A} s(x)} - L \max_{x \in A} s(x) & L \in (-1, 0) \\ \frac{L(|A|-1) \sum_{x \in A} s(x) [1 - \sum_{x \in A} s(x)]}{(n-|A|) \sum_{x \in X-A} s(x) + L(|A|-1) \sum_{x \in A} s(x)} + \sum_{x \in A} s(x) & L \in (0, \infty) \end{cases} \quad (1)$$

where $A \subseteq X$, $L \in [-1, \infty)$, $s(\cdot)$ is a singleton measure for each $x_i \in X$, $\sum_{x \in X} s(x) = 1$, $g_{L_{m\delta}}(\emptyset) = 0$ and $g_{L_{m\delta}}(X) = 1$. This fuzzy measure satisfies the following properties: (1) $L_{m\delta}$ -measure is an increasing function on L ; (2) if $L = -1$, then $L_{m\delta}$ -measure is just the P-measure; (3) if $L = 0$, then $L_{m\delta}$ -measure is additive (it coincides with the λ -measure when $\lambda = 0$ —see Additional file 1); (4) if $-1 < L < 0$, then $L_{m\delta}$ -measure satisfies the subadditivity property (low synergism); and (5) if $0 < L < \infty$, then $L_{m\delta}$ -measure satisfies the superadditivity property (high synergism).

Mathematical definition of the Choquet integral

The Choquet integral was first presented in capacity theory [13]. Its use as an integral with regard to fuzzy measures was then introduced by Hohle [60] and, it was later rediscovered by Murofushi and Sugeno [61, 62]. This integral, as an n-place operator, has been used in several works [18–23], in order to fuse information when interrelated criteria are accounted for. Formally, let a finite set $X = \{x_1, x_2, \dots, x_N\} | X \in \mathfrak{N}_{\geq 0}^N$ and μ be a fuzzy measure on N , then the Choquet integral of X with respect to μ is a function $C_\mu: \mathfrak{N}_{\geq 0}^N \rightarrow \mathfrak{N}_{\geq 0}$ according to the next expression:

$$C_\mu(x_1, x_2, \dots, x_N) = \sum_{i=1}^N x_{(i)} [\mu(A_{(i)}) - \mu(A_{(i-1)})] \tag{2}$$

where (\cdot) denotes a permutation on N , so that $x_{(1)} \geq x_{(2)} \geq \dots \geq x_{(N)}$. That is, $x_{(i)}$ is the i -th largest value in the set $\{x_1, x_2, \dots, x_N\}$. Thus, $A_{(i)} = \{x_{(i)}, \dots, x_{(N)}\}$ when $i \geq 1$, and $A_0 = \emptyset$. So, for instance, if $X = \{x_1, x_2, x_3\} | x_2 \geq x_3 \geq x_1$, then the Choquet integral-based aggregation is computed as follows:

$$C_\mu(x_1, x_2, x_3) = x_2[\mu(x_2, x_3, x_1) - \mu(x_3, x_1)] + x_3[\mu(x_3, x_1) - \mu(x_1)] + x_1[\mu(x_1)]$$

As it can be seen, this operator performs a reordering of its arguments according to their magnitudes, as the OWA-like operators do [6, 7]. Indeed, as it has been demonstrated elsewhere [15, 63, 64], the Choquet integral constitutes a generalization of the latter. Moreover, it can be observed that since the values are ordered in decreasing order, then $\mu(A_{(i)}) \geq \mu(A_{(i-1)})$. Lastly, it is important to highlight that the Choquet integral fulfils some properties, such as: (1) it is a continuous function; (2) it is homogeneous of degree 1; (3) it is monotonic and idempotent, if and only if μ is a fuzzy measure; and (4) it is compensative when μ is a normalized fuzzy measure.

Extending traditional functions to derive Choquet integral-based fuzzy descriptors

Table G3 in [28] shows several traditional functions to derive classic 2D/3D-MDs from atom/bond contributions (LOVIs/LOEIs), e.g. the Zagreb indices [65], the Balaban-like indices [66], the Wiener-type indices [67], among others [28]. As it can be observed, these functions are mainly based on the summation and product aggregation operators. Consequently, the functions described in Table G3 do not consider the possible interrelation among LOVIs/LOEIs, which are only aggregated of linear and/or additive ways. Therefore, in order to consider the dependency among LOVIs/LOEIs, some of these traditional functions can be extended to compute fuzzy 2D/3D-MDs (FMDs), by using the Choquet integral (C_μ) with respect to a fuzzy measure μ as shown below:

$$MD^1(L, \alpha', \lambda') = \alpha' \cdot \sum_{i=1}^A L_i^{\lambda'} \rightarrow FMD_{C_\mu}^1(L, \alpha', \lambda') = C_\mu(\alpha' \cdot L_1^{\lambda'}, \dots, \alpha' \cdot L_A^{\lambda'}) \tag{3}$$

$$MD^4(L, \alpha', \lambda') = \alpha' \sum_{i=1}^A \sum_{j=1}^A a_{ij} (L_i \cdot L_j)^{\lambda'} \rightarrow FMD_{C_\mu}^4(L, \alpha', \lambda') = C_\mu(L_1, \dots, L_W) | L_w = a_{ij} \cdot \alpha' \cdot (L_i \cdot L_j)^{\lambda'} \quad \forall i, j \in V \tag{4}$$

$$MD^5(L, \alpha', \lambda') = \alpha' \sum_{i=1}^A \sum_{j=1}^A (L_i \cdot L_j)^{\lambda'} \rightarrow FMD_{C_\mu}^5(L, \alpha', \lambda') = C_\mu(L_1, \dots, L_W) | L_w = \alpha' \cdot (L_i \cdot L_j)^{\lambda'} \quad \forall i, j \in V, i \neq j \tag{5}$$

$$MD^6(L, \alpha', \lambda') = \alpha' \sum_{k=1}^K \left(\prod_{i=1}^{n_k} L_i \right)_k^{\lambda'} \rightarrow FMD_{C_\mu}^6(L, \alpha', \lambda') = \alpha' \sum_{k=1}^K C_\mu(L_1^{\lambda'}, \dots, L_{n_k}^{\lambda'})_k \tag{6}$$

$$MD^7(L, \alpha', \lambda', k) = \alpha' \sum_{i \in k} \sum_{j \in k} a_{ij} (L_i \cdot L_j)^{\lambda'} \rightarrow FMD_{C_\mu}^7(L, \alpha', \lambda', k) = C_\mu(L_1, \dots, L_W) \quad (7)$$

$$|L_w = a_{ij} \cdot \alpha' \cdot (L_i \cdot L_j)^{\lambda'} \quad \forall i, j \in k$$

where L_i and L_j are the LOVI values for any pair vertices v_i and v_j (atoms) of a molecular graph G , A is the number of vertices, V represents the set of vertices of G , a_{ij} denotes the coefficients of the adjacency matrix of G (1 for adjacent vertices, 0 otherwise), K is the total number of graph fragments to be considered, n_k is the number of vertices within the k th fragment, and α' and λ' are two real parameters. The superscript in notation MD represents the numbering used to identify these functions in Table G3 in [28]. Accordingly, this numbering is also used to identify the corresponding fuzzy formulations (FMD_{C_μ}). Note that these definitions can also be used to compute fuzzy 2D/3D-MDs from LOEIs in place of LOVIs.

From these fuzzy formulations, several specific descriptors can be computed, for instance: (1) from $FMD_{C_\mu}^1$ for $\alpha' = \lambda' = 1$, fuzzy DIVATI MDs [41], fuzzy GT-STAF MDs [34] and fuzzy QuBiLS-MAS MDs [37] can be obtained, when the LOVIs vector is computed with some of those families; (2) if vector L is computed with the vertex degree invariant, then from $FMD_{C_\mu}^1$ for $\alpha' = 1$ and $\lambda' = 2$, from $FMD_{C_\mu}^4$ for $\alpha' = \lambda' = 1$ and from $FMD_{C_\mu}^4$ for $\alpha' = 1$ and $\lambda' = -1/2$, the fuzzy first Zagreb index [65], the fuzzy second Zagreb index [65] and the fuzzy Randić connectivity index [68] can be obtained, respectively; (3) from $FMD_{C_\mu}^4$ for $\alpha' = \frac{B}{C+1}$ (B is the number of graph edges (covalent edges) and C is the number of rings) and $\lambda' = -1/2$, the fuzzy Balaban-like indices can be determined [32]; (4) from $FMD_{C_\mu}^6$ for $\alpha' = 1$ and $\lambda' = -1/2$, the fuzzy Kier–Hall connectivity indices can be obtained [69]; and (5) from $FMD_{C_\mu}^7$, fuzzy autocorrelation MDs can be computed. A practical example is presented below.

Practical example: Choquet integral-based fuzzy QuBiLS-MIDAS molecular descriptors

Geometric multi-linear algebraic MDs, also known as QuBiLS-MIDAS, were introduced as a novel framework to characterize molecular structures [39, 40]. These 3D-MDs are the only ones that encode structural information between two atoms of a molecule using several metrics (e.g. Soergel) [39], as well as chemical information corresponding to the relations between three and

four atoms through multi-metrics (e.g. bond and dihedral angle) [40]. QSAR studies on eight benchmark chemical datasets were carried out [43], where the QuBiLS-MIDAS 3D-MDs yielded significantly superior outcomes with respect to 12 2D/3D-QSAR methodologies established in the literature. The QuBiLS-MIDAS 3D-MDs were also applied in the prediction of inhibitory activity of bromodomain modulators (BRD2, BRD3 and BRD4) with successful results [46].

Traditional (no fuzzy) definition of the QuBiLS-MIDAS descriptors

The traditional QuBiLS-MIDAS MDs are computed from atom-level descriptors (LOVIs). Thus, the k -th atom-level two-linear $\left[{}_{ns(ss,ds,mp)}^{(*)} b_{(F)}^{a,k}(\bar{x}, \bar{y}) \right]$, three-linear $\left[{}_{ns(ss,mp)}^{(*)} tr_{(F)}^{a,k}(\bar{x}, \bar{y}, \bar{z}) \right]$ and four-linear $\left[{}_{ns(ss,mp)}^{(*)} qu_{(F)}^{a,k}(\bar{x}, \bar{y}, \bar{z}, \bar{w}) \right]$ QuBiLS-MIDAS 3D-MDs are calculated as N -linear (multi-linear) algebraic maps in \mathbb{R}^n , in a canonical basis set, when geometric coordinate-based relations among two ($N = 2$), three ($N = 3$) and four ($N = 4$) atoms are considered, respectively [39, 40]. The formulation (indicial notation) of these 3D-MDs is as follows:

$$\begin{aligned} {}_{ns(ss,ds,mp)}^{(*)} b_{(F)}^{a,k} L_a^k &= {}_{ns(ss,ds,mp)}^{(*)} b_{(F)}^{a,k}(\bar{x}, \bar{y}) \\ &= {}_{ns(ss,ds,mp)}^{(*)} G_{ij(F)}^{a,k} x^{i(*)} y^{j(*)} \end{aligned} \quad (8)$$

$$\begin{aligned} {}_{ns(ss,mp)}^{(*)} tr_{(F)}^{a,k} L_a^k &= {}_{ns(ss,mp)}^{(*)} tr_{(F)}^{a,k}(\bar{x}, \bar{y}, \bar{z}) \\ &= {}_{ns(ss,mp)}^{(*)} G_{ijl(F)}^{a,k} x^{i(*)} y^{j(*)} z^{l(*)} \end{aligned} \quad (9)$$

$$\begin{aligned} {}_{ns(ss,mp)}^{(*)} qu_{(F)}^{a,k} L_a^k &= {}_{ns(ss,mp)}^{(*)} qu_{(F)}^{a,k}(\bar{x}, \bar{y}, \bar{z}, \bar{w}) \\ &= {}_{ns(ss,mp)}^{(*)} G_{ijlh(F)}^{a,k} x^{i(*)} y^{j(*)} z^{l(*)} w^{h(*)} \end{aligned} \quad (10)$$

where n is the number of atoms, “ a ” is a particular atom ($a = 1, \dots, n$), the indices $i, j, l, h = 1 \dots n$ denote the entries of the matrices and property vectors, $k = \pm 1, \dots, \pm 12$ is the power of the matrices, and $x^{1(*)}, \dots, x^{n(*)}$, $y^{1(*)}, \dots, y^{n(*)}$, $z^{1(*)}, \dots, z^{n(*)}$ and $w^{1(*)}, \dots, w^{n(*)}$ are the coefficients of the property vectors $x^{(*)}$, $y^{(*)}$, $z^{(*)}$ and $w^{(*)}$, respectively, when central chirality aspects are codified (*) or not [70]). Moreover, ${}_{ns(ss,ds,mp)}^{(*)} G_{(F)}^{a,k}$, ${}_{ns(ss,mp)}^{(*)} G_{(F)}^{a,k}$ and

${}_{ns(ss,mp)}^{(NQ)}GQ_{(F)}^{a,k}$ denote the two-, three- and four-tuple atom-level matrices for each atom “ a ”, respectively. From these atom-level matrices, then atom-level descriptors (LOVI) are determined. Each LOVI constitutes an entry (L_a) in the corresponding vector of atom-level descriptors L^k (LOVIs vector). The notations (NQ, F , ns , ss , ds and mp) between parentheses are not mandatory during the calculation and they will be explained below.

Keep-all total matrices (G^k , GT^k and GQ^k) are the basis to compute these 3D-MDs. For $k = 1$, the entries of the matrices G^1 , GT^1 and GQ^1 denote the information encoded for the relations between two, three and four atoms of a molecule, respectively, by using several metrics and multi-metrics (see Tables 1–2 in [43]). From these matrices, neighborhood-quotient matrices (NQG^k , $NQGT^k$ and $NQGQ^k$) may be obtained, which contain information of the inter-atomic relations that satisfy certain molecular cutoffs [71]. Local-fragment matrices (G_F^k , GT_F^k and GQ_F^k) may also be computed (see Equation 13 in [39] and Equations 17–18 in [40]) to encode information of chemical fragments or atom-types (F) of interest. Normalized matrices may also be obtained using the simple-stochastic (ss —see Equation 10 in [39] and Equations 13–14 in [40]), double-stochastic (ds) [72] and mutual probability (mp —see Equation 12 in [39] and Equation 15–16 in [40]) procedures. If no normalization procedure is used, then the matrices are non-stochastic (ns).

Finally, from the keep-all (neighborhood-quotient) non-stochastic (simple-stochastic, double-stochastic or mutual-probability) total (local-fragment) matrices [${}_{ns(ss,ds,mp)}^{(NQ)}G_{(F)}^k$, ${}_{ns(ss,mp)}^{(NQ)}GT_{(F)}^k$ and ${}_{ns(ss,mp)}^{(NQ)}GQ_{(F)}^k$], the respective atom-level matrices are calculated (see Equation 9 in [39] and Equations 3–4 in [40]) with the purpose of determining the vectors of LOVIs (see Eqs. 8–10). After that, and considering the atom-level descriptors (LOVIs) as independent values of one another, then the (non-fuzzy) global k -th two-linear, three-linear and four-linear QuBiLS-MIDAS 3D-MDs are obtained using one or several (non-fuzzy) aggregation operators based on the Minkowski definition (e.g. Euclidean norm), central tendency statistics (e.g. harmonic mean) and dispersion statistics (e.g. variance) [39, 40].

Fuzzy definition of the QuBiLS-MIDAS descriptors based on the Choquet integral

So far, QuBiLS-MIDAS 3D-MDs are computed from LOVIs considered as non-interrelated values. However, as already pointed out, the biological activities or properties of the compounds do not only depend on the molecular shape, but also on the interactions that are often non-covalent in nature. Therefore, the interrelation

among atomic contributions (LOVIs) may be an aspect to consider during molecular encoding. In this way, from the corresponding LOVIs vector and considering their coefficients as interrelated values of one another, then the fuzzy global k -th two-linear, three-linear and four-linear QuBiLS-MIDAS 3D-MDs are computed using the definition of Choquet integral (see Eq. 2) as shown below:

$${}_{ns(ss,ds,mp)}^{(*)}b_{(F)}^k(\bar{x}, \bar{y}) = C_{\mu} \left({}_{ns(ss,ds,mp)}^{(*)}b_{(F)}L_1^k, \dots, {}_{ns(ss,ds,mp)}^{(*)}b_{(F)}L_n^k \right) \quad (11)$$

$${}_{ns(ss,mp)}^{(*)}tr_{(F)}^k(\bar{x}, \bar{y}, \bar{z}) = C_{\mu} \left({}_{ns(ss,mp)}^{(*)}tr_{(F)}L_1^k, \dots, {}_{ns(ss,mp)}^{(*)}tr_{(F)}L_n^k \right) \quad (12)$$

$$\begin{aligned} & {}_{ns(ss,mp)}^{(*)}qu_{(F)}^k(\bar{x}, \bar{y}, \bar{z}, \bar{w}) \\ &= C_{\mu} \left({}_{ns(ss,mp)}^{(*)}qu_{(F)}L_1^k, \dots, {}_{ns(ss,mp)}^{(*)}qu_{(F)}L_n^k \right) \end{aligned} \quad (13)$$

where $C_{\mu}(\dots)$ is the Choquet integral with respect to a fuzzy measure μ ; and ${}_{ns(ss,ds,mp)}^{(*)}b_{(F)}L_a^k$, ${}_{ns(ss,mp)}^{(*)}tr_{(F)}L_a^k$ and ${}_{ns(ss,mp)}^{(*)}qu_{(F)}L_a^k$ are the k th two-linear, three-linear and four-linear atom-level descriptors (LOVIs), respectively, determined for each atom “ a ” of a molecule, according to Eqs. 8–10. Note that these formulations coincide with the definition $FMD_{C_{\mu}}^1$ for $\alpha' = \lambda' = 1$ (see Eq. 3). Scheme 1

shows a flowchart regarding the calculation of these fuzzy 3D-MDs.

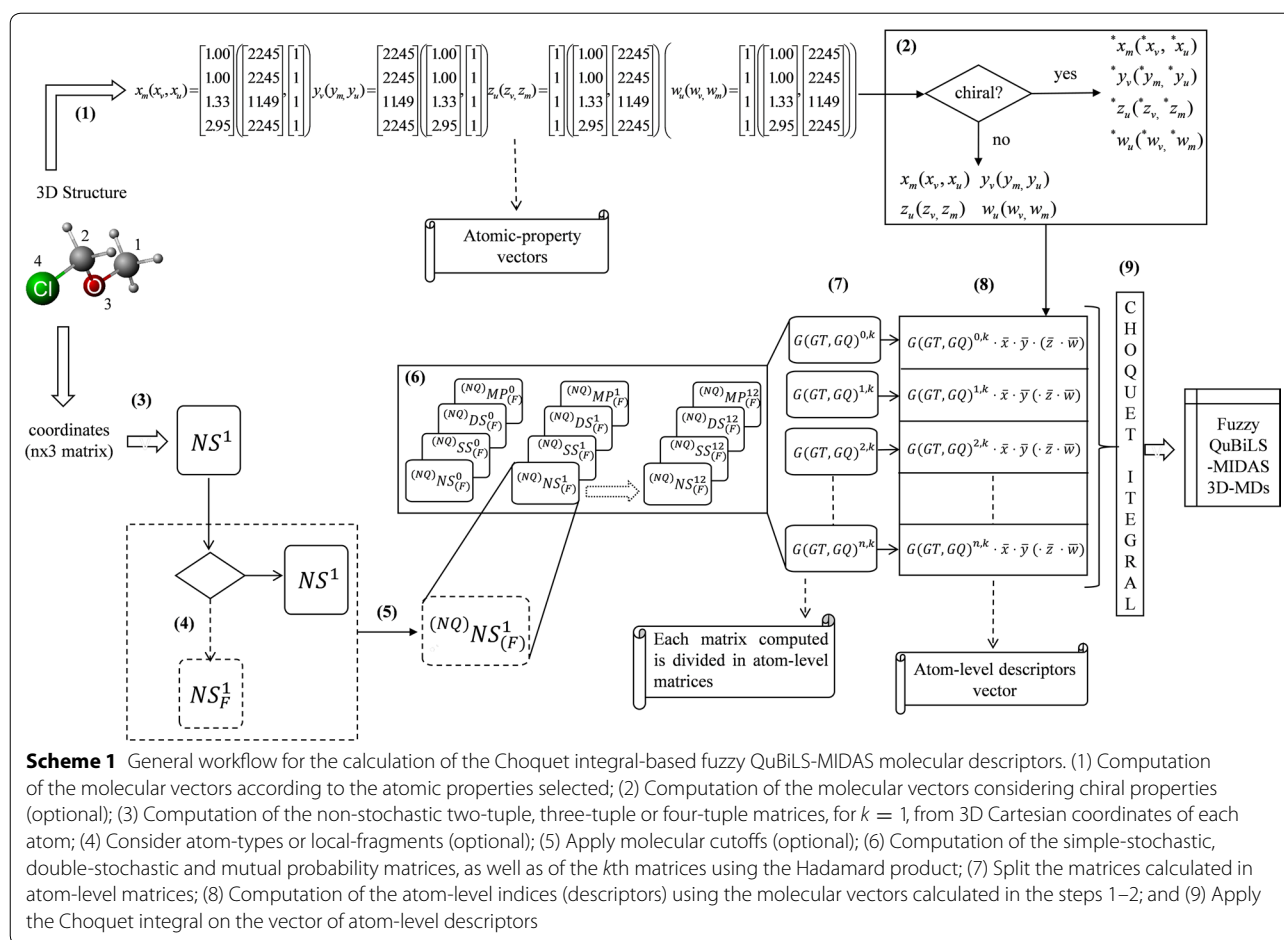
The $L_{m\delta}$ -measure [54, 59] (see Eq. 1) is used to compute the importance (weight) of the interrelation among atom-level descriptors (LOVIs) during the fuzzy QuBiLS-MIDAS 3D-MDs calculation. As it can be seen in Eq. 1, the $L_{m\delta}$ -measure depends on a singleton measure $s(x)$, so that $\sum_{x \in X} s(x) = 1$, being X a finite set of elements. In this case, set X is the LOVIs vector (L^k) computed for a compound, either the k th two-linear, three-linear or four-linear atom-level QuBiLS-MIDAS 3D-MDs. Therefore, the singleton measure $s(L_a^k)$, $\forall L_a^k \in L^k$, determines the belonging degree of the descriptor for atom “ a ” (L_a^k —see Eqs. 8–10) within the set of atom-level descriptors (L^k). To this end, the following two functions [73] were used:

Aggregated Objects Type 1 (AO1):

$$s(L_i^k) = \frac{b_i^{\alpha}}{\sum_{i=1}^n b_i^{\alpha}}, \quad i = 1, 2, \dots, n \quad (14)$$

Aggregated Objects Type 2 (AO2):

$$s(L_i^k) = \frac{(1 - b_i)^{\alpha}}{\sum_{i=1}^n (1 - b_i)^{\alpha}}, \quad i = 1, 2, \dots, n \quad (15)$$



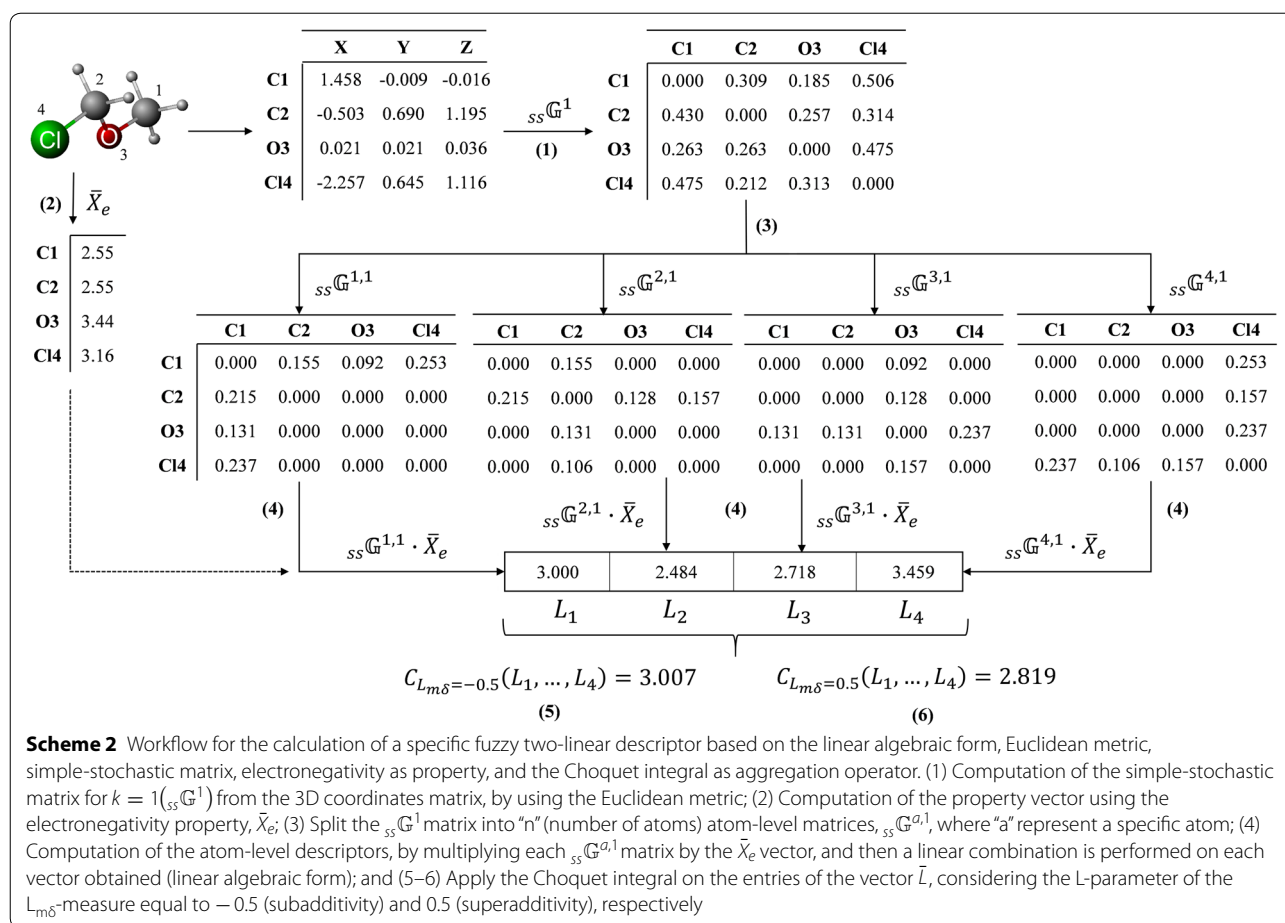
where $\alpha \in [0, 1]$; and b_i constitutes the i th largest of the atom-level descriptors $L_1^k, L_2^k, \dots, L_n^k$. As it can be analyzed, these functions calculate the belonging degree according to the magnitude of the atom-level descriptors, that is, the highest belonging degrees are assigned to the highest atom-level descriptors. The AO1 and AO2 functions are commonly used in the weightings computation for OWA-like operators [73]. However, they satisfy the same mathematical constraints as the singleton measures for the $L_{m\delta}$ -measure and, thus, they were considered to compute the fuzzy densities in this work. Scheme 2 shows an example of the calculation of these fuzzy 3D-MDs.

Lastly, the computation of these fuzzy global 3D-MDs can be performed through the module built into the QuBiLS-MIDAS software (<http://tomocomd.com/qubils-midas>) [35]. As it can be seen in Fig. 1, in this module, the value of the L-parameter corresponding to the $L_{m\delta}$ -measure (see Eq. 1), as well as the singleton measure to be used, can be customized. Several default configurations, determined according to their results in cheminformatics studies, are also provided. These

fuzzy 3D-MDs can also be obtained using the distributed computation module coupled to the heterogeneous and non-dedicated T-arenal platform (<http://tomocomd.com/t-arenal>) [74].

Performance of the Choquet integral-based fuzzy QuBiLS-MIDAS molecular descriptors

This section is dedicated to demonstrating the feasibility of using the Choquet integral in the fuzzy MDs calculation, by using the QuBiLS-MIDAS 3D-MDs as study case. To this end, models based on the multiple linear regression (MLR) technique were built using the MobyDigs software, which uses the Genetic Algorithm (GA) meta-heuristic as search method [75]. The leave-one-out cross validation (Q_{loo}^2) was used as the fitness function. The models retained for further validation were selected according to the best bootstrapping value (Q_{boot}^2). All the datasets were optimized with the CORINA software (<https://www.mn-am.com/products/corina>). From now on, the ‘atomic contributions’ term is only to refer to the atom-level QuBiLS-MIDAS 3D-MDs (LOVIs).



Configurations for the computation of fuzzy densities

This study is to determine the best configurations for the computation of fuzzy densities. To this end, two project groups with the same configuration of Choquet integral-based fuzzy QuBiLS-MIDAS 3D-MDs were built (see Additional file 2). In both project groups, the value of the α -parameter of the functions AO1 (see Eq. 14) or AO2 (see Eq. 15) was varied into the interval $[0, 1]$, with a step equal to 0.1. In this way, all the possible configurations were assessed. The value of the L-parameter of the $L_{m\delta}$ -measure (see Eq. 1) was set to -0.5 and 0.5 , in order to determine the best configurations when a low and a high synergism among the atomic contributions is accounted for. The Cramer's steroids set [76–80] is the one used to fulfill the goals of this study. This dataset is composed of 31 steroids, so that structures 1–21 and 22–31 belong to the training and test sets, respectively. Compound number 31 is left out at being outlier.

These projects on the steroid's dataset were calculated (see Additional file 2 for SDF format). For each descriptors matrix obtained (see Additional file 3), models from 1 to 4 variables were created by using the GA-MLR method, in order to predict the binding affinity to the CBG protein [81]. The statistical methods Y-scrambling, bootstrapping, external validation and Fisher function were determined for each model, in order to create a data matrix $M_{n \times k}$, where the n rows and the k columns denote the statistics computed and the configurations to be compared, respectively (see Additional file 4). In this way, the rank (first step of the Friedman test [82]) for each configuration can be computed (see Additional file 5). The configurations selected as the best ones were those with a rank lesser than the difference between the average rank and the standard deviation calculated for a same group.

As a result, Fig. 2 shows the average bootstrapping (Q_{boot}^2) and external predictive (Q_{ext}^2) accuracies

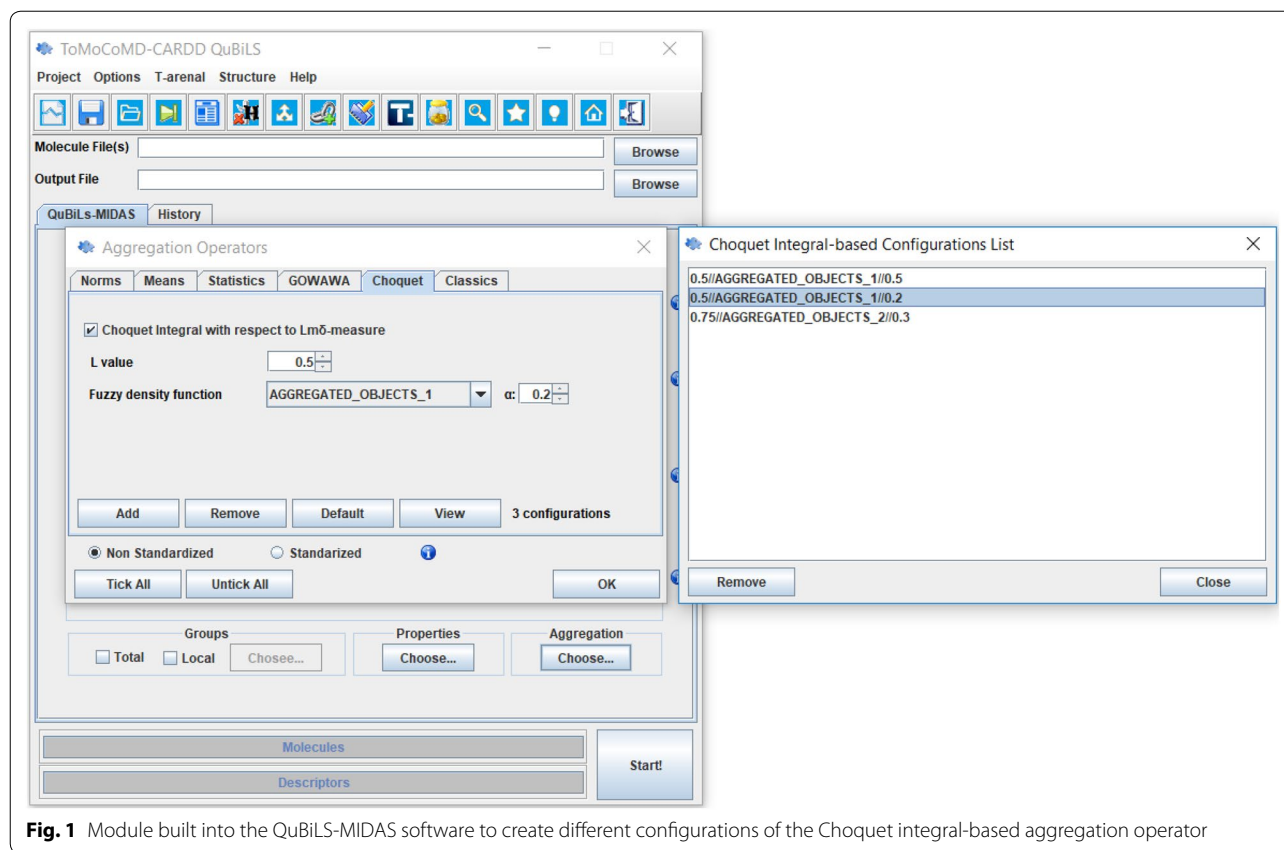


Fig. 1 Module built into the QuBiLS-MIDAS software to create different configurations of the Choquet integral-based aggregation operator

corresponding to the best configurations based on functions AO1 (see Eq. 14) and AO2 (see Eq. 15). On one hand, the configurations AO1 ($\alpha = 0.2$), AO1 ($\alpha = 0.3$), AO2 ($\alpha = 0.6$) and AO2 ($\alpha = 0.0$) are those with the best outcomes when a low synergism is considered during the fuzzy QuBiLS-MIDAS 3D-MDs calculation (see Fig. 2a). On the other hand, at considering a high synergism (see Fig. 2b), the configurations with the best behavior are AO1 ($\alpha = 0.8$), AO1 ($\alpha = 0.9$), AO1 ($\alpha = 0.2$), AO2 ($\alpha = 0.6$) and AO2 ($\alpha = 0.5$). In all cases, the average performance achieved by the models is suitable, at presenting $Q_{boot}^2 > 0.7$ and $Q_{ext}^2 > 0.6$.

Nonetheless, in a general sense, the configurations obtained for a high synergism (Fig. 2b) present a comparable-to-superior behavior with regard to the configurations obtained for a low synergism (Fig. 2a). These results suggest that the dependency among atomic contributions is an important aspect to consider during the molecular codification with the QuBiLS-MIDAS 3D-MDs. Thus, at least preliminarily, the methodological contribution of this report is justified, in which the 2D/3D-MDs fuzzy calculation from LOVIs/LOEIs is presented. However, more studies must be carried out to prove the feasibility of this fuzzy approach with respect to the traditional

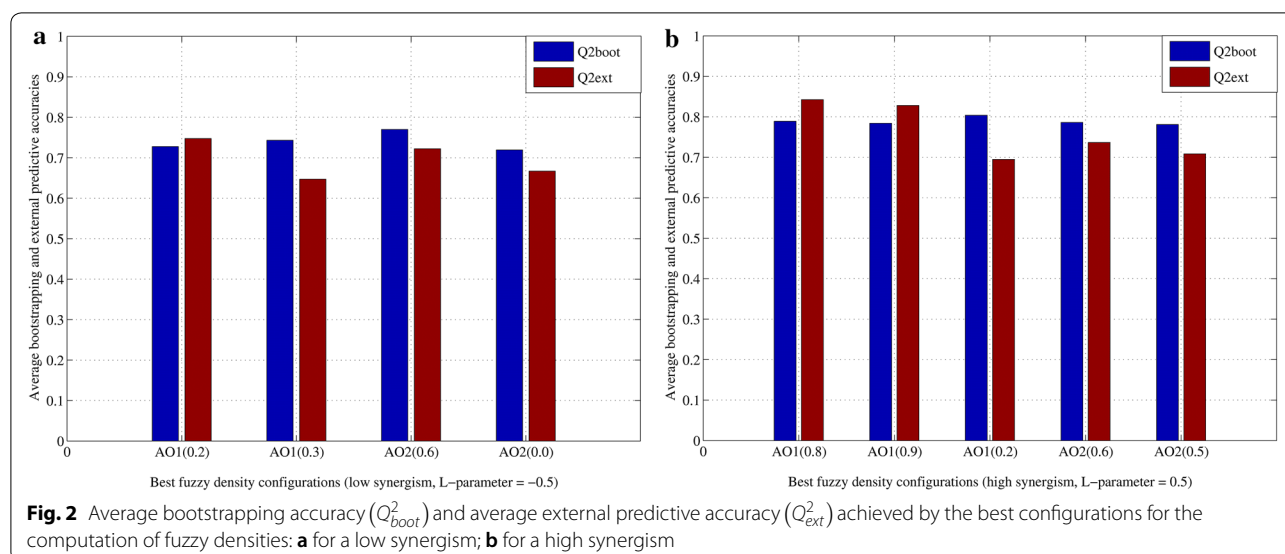
approach, where the atom/bond contributions (LOVIs/LOEIs) are considered as non-interrelated values.

Performance of the Choquet integral-based (fuzzy) QuBiLS-MIDAS descriptors versus norm-, mean- and statistic-based (non-fuzzy) QuBiLS-MIDAS descriptors

In this section, QSAR models based on the QuBiLS-MIDAS 3D-MDs were built to assess the applicability of the Choquet integral in the fuzzy MDs computation from LOVIs/LOEIs. Note that the superiority of the QuBiLS-MIDAS 3D-MDs in modeling tasks was confirmed in [43], where its performance was assessed and compared with regard to 12 methodologies reported in the literature. Therefore, the current study is only devoted to performing an internal analysis among the models built with the fuzzy 3D-MDs (based on Choquet integral) with respect to the models built with the non-fuzzy 3D-MDs (based on traditional operators).

Chemical datasets to assess the performance between fuzzy and non-fuzzy QuBiLS-MIDAS descriptors

Eight well-known benchmark datasets were used to carry out this study. These datasets have been widely employed in the literature [83–86], including the analysis to assess the performance of the QuBiLS-MIDAS MDs in QSAR



[43]. The datasets are composed of angiotensin converting enzyme (ACE) inhibitors, acetylcholinesterase inhibitors (ACHE), ligands for the benzodiazepine receptor (BZR), cyclooxygenase-2 inhibitors (COX2), dihydrofolate reductase inhibitors (DHFR), inhibitors of glycogen phosphorylase b (GPB), thermolysin inhibitors (THER) and thrombin inhibitors (THR). A description of these datasets is shown in Table 1, whereas Additional file 6 contains the corresponding SDF (Structure Data Format) files.

Methodology to assess the performance between fuzzy and non-fuzzy QuBiLS-MIDAS descriptors

Three projects with a same configuration of non-fuzzy QuBiLS-MIDAS 3D-MDs were built, each of them using the norm-, mean-, statistic-based operators, respectively. Two other projects with the same previous configuration were also created but using the Choquet integral to determine the respective fuzzy 3D-MDs. One of the Choquet integral-based projects was designed with the best fuzzy densities obtained for a low synergism among atomic contributions (see Fig. 2a), while the other project

was planned with the best fuzzy densities for a high synergism (see Fig. 2b). The L-parameter of the $L_{m\delta}$ -measure was set to -0.25 , -0.5 and -0.75 (subadditivity) in the project considering a low synergism, while the opposite values (superadditivity) were used in the other project. Additional file 6 shows the XML files of the projects described.

These projects on each chemical dataset mentioned above (see Table 1) were computed. Then, the best 1500 variables (MDs) according to the variability criterion [87] were retained by using the IMMAN software [88]. Posteriorly, the GA-MLR procedure was used to build several models for 3, 5 and 7 variables for each operator-type. The best model for each dimension on each dataset was retained (Additional file 7). A pool with the non-fuzzy 3D-MDs and other pool with the fuzzy 3D-MDs included in the best models built on each dataset were created. From these pools, non-fuzzy and fuzzy models for 7 variables were built on each dataset, and the models with the best bootstrapping value were selected as the best ones (Additional file 8). The external validation (Q_{ext}^2) statistic

Table 1 Description of the chemical datasets employed to assess the performance of the Choquet integral-based QuBiLS-MIDAS descriptors

	ACE	ACHE	BZR	COX2	DHFR	GPB	THER	THR
Total	114	111	163	322	397	66	76	88
Training	76	74	98	188	237	44	51	59
Test	38	37	49	94	124	22	25	29
Inactive			16	40	36			
Activity	pIC ₅₀	pIC ₅₀	pIC ₅₀	pIC ₅₀	pIC ₅₀	pK _i	pK _i	pK _i
Value range	2.1–9.9	4.3–9.5	5.5–8.9	4.0–9.0	3.3–9.8	1.3–6.8	0.5–10.2	4.4–8.5

parameter was computed for each model developed (see Additional file 7, Additional file 8).

The Q_{ext}^2 values (predictive abilities) obtained for each chemical dataset were used to establish a comparison and statistical assessment between the models built with the fuzzy and non-fuzzy QuBiLS-MIDAS 3D-MDs, respectively. In this sense, an analysis by means of a boxplot graphic (box-and-whisker graphic) was firstly performed, in order to examine the shape of the distributions of the results achieved. Then, a Wilcoxon signed-rank test [89] was carried out to know whether the predictive abilities achieved by the fuzzy models and the predictive abilities achieved by the non-fuzzy models differ. The SPSS software was used to perform the first analysis mentioned above, while the Keel [90] software was employed to perform the other one. A significance level $\alpha = 0.05$ was accounted for. Note that the 'fuzzy model' and 'non-fuzzy model' terms are referred to the models built with the fuzzy and non-fuzzy QuBiLS-MIDAS 3D-MDs, respectively.

Analysis of the performance achieved by the fuzzy and non-fuzzy QuBiLS-MIDAS descriptors

Figure 3 shows a comparative graphic of the average performance achieved by the models for 3, 5 and 7 variables, built with the fuzzy QuBiLS-MIDAS 3D-MDs, when a low (L-parameter < 0 in Eq. 1) and a high (L-parameter > 0 in Eq. 1) synergism among atomic contributions is accounted for. As it can be seen, the fuzzy 3D-MDs calculated for a low synergism present the best behavior on ACE ($(Q_{ext}^2)=0.5667$), BZR ($(Q_{ext}^2)=0.4052$), COX2 ($(Q_{ext}^2)=0.2930$) and GPB ($(Q_{ext}^2)=0.5931$) datasets, while the fuzzy 3D-MDs determined for a high synergism present the best behavior on ACHE ($(Q_{ext}^2)=0.4303$), DHFR ($(Q_{ext}^2)=0.3446$), THER ($(Q_{ext}^2)=0.4541$) and THR ($(Q_{ext}^2)=0.3270$) datasets. So, it is evidenced that fuzzy QuBiLS-MIDAS MDs calculated both for a low and a high synergism contribute to codify useful chemical information, and that their performances depend on the molecular structures under study. Therefore, both types of fuzzy 3D-MDs should be jointly used with the purpose of creating models with better predictive ability.

In this sense, Fig. 4 shows a plotting of the external predictive ability yielded by the models of 7 variables created

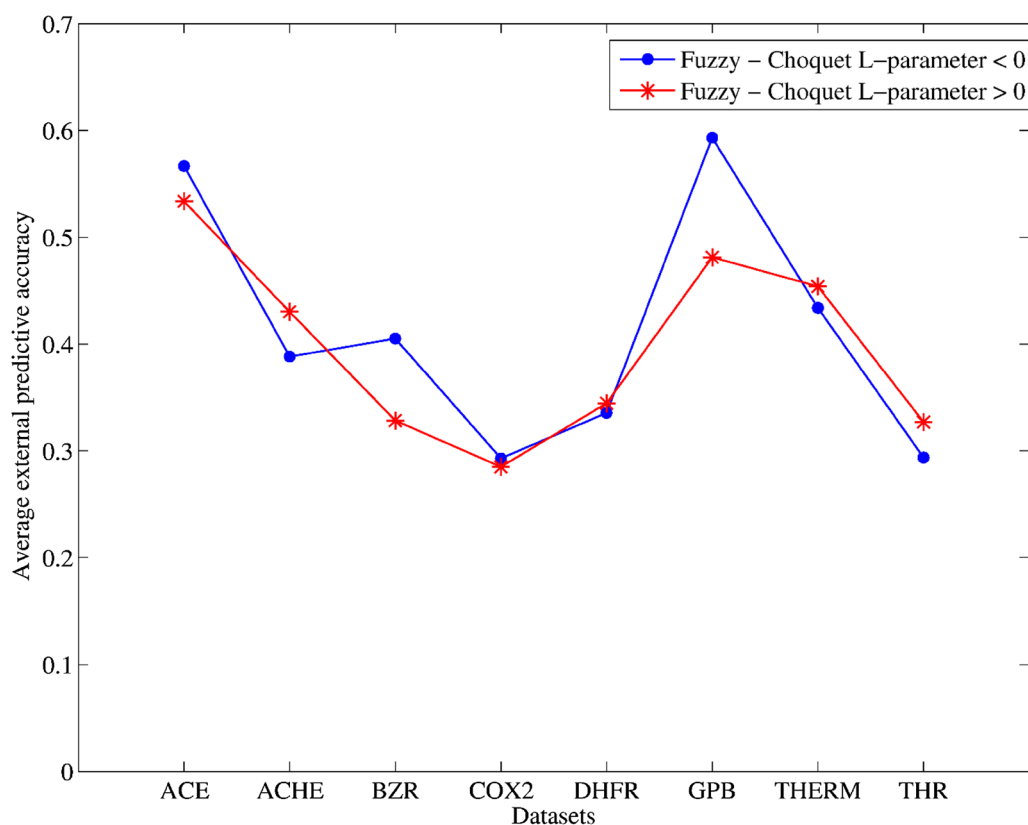


Fig. 3 Average external predictive (Q_{ext}^2) accuracies corresponding to the models based on the best configurations for the calculation of fuzzy densities, both those determined for a low (L-parameter < 0) and a high (L-parameter > 0) synergism

with the fuzzy and non-fuzzy QuBiLS-MIDAS 3D-MDs, respectively. The fuzzy MDs determined for a low and a high synergism among atomic contributions were used, while the non-fuzzy MDs used are those computed from the norm-, mean- and statistic-based operators. It can be seen that, in all the chemical datasets employed, the models built with the fuzzy MDs [(ACE, (Q_{ext}^2)=0.6103); (ACHE, (Q_{ext}^2)=0.5231); (BZR, (Q_{ext}^2)=0.5400); (COX2, (Q_{ext}^2)=0.3558); (DHFR, (Q_{ext}^2)=0.4638); (GPB, (Q_{ext}^2)=0.6447); (THER, (Q_{ext}^2)=0.4569); (THR, (Q_{ext}^2)=0.4072)] yield comparable-to-superior performances with regard to the models based on the non-fuzzy MDs [(ACE, (Q_{ext}^2)=0.5629); (ACHE, (Q_{ext}^2)=0.3887); (BZR, (Q_{ext}^2)=0.5222); (COX2, (Q_{ext}^2)=0.3387); (DHFR, (Q_{ext}^2)=0.4390); (GPB, (Q_{ext}^2)=0.6442); (THER, (Q_{ext}^2)=0.4080); (THR, (Q_{ext}^2)=0.3600)]. Thus, it can be stated that MDs with better modeling ability can be calculated using the Choquet integral-based operator, if compared with the MDs computed from the traditional (non-fuzzy) operators.

Moreover, Fig. 5a depicts the number of fuzzy QuBiLS-MIDAS 3D-MDs, both for a low and a high synergism among atomic contributions, included in the models built

on each dataset. As it can be seen, the fuzzy QuBiLS-MIDAS 3D-MDs determined for a high synergism influenced on the external predictive power of all models developed, being the models corresponding to the ACE and THR datasets exclusive of these MDs. It can also be noted that the model developed on the ACE dataset presents more QuBiLS-MIDAS 3D-MDs for a high synergism than for a low synergism. A likewise behavior is shown by the models built on the BZR, COX2, DHFR, GPB and THER datasets, but in these cases, there are more fuzzy MDs for a low synergism.

Additionally, Fig. 5b shows the L-parameter average value for the MDs included in the models. In general, it can be seen that, albeit the superadditivity is exclusive for the ACE and THR datasets, the amount of superadditivity on each dataset is moderate. This behavior can be due to the fact that the datasets used are comprised of congeneric compounds. That is, since the compounds are structurally similar, then MDs computed of additive way, or considering a low synergism among atomic contributions, may be those in achieving better correlations into a QSAR model. This assumption is supported in the average amount of low synergism obtained. As it can be seen,

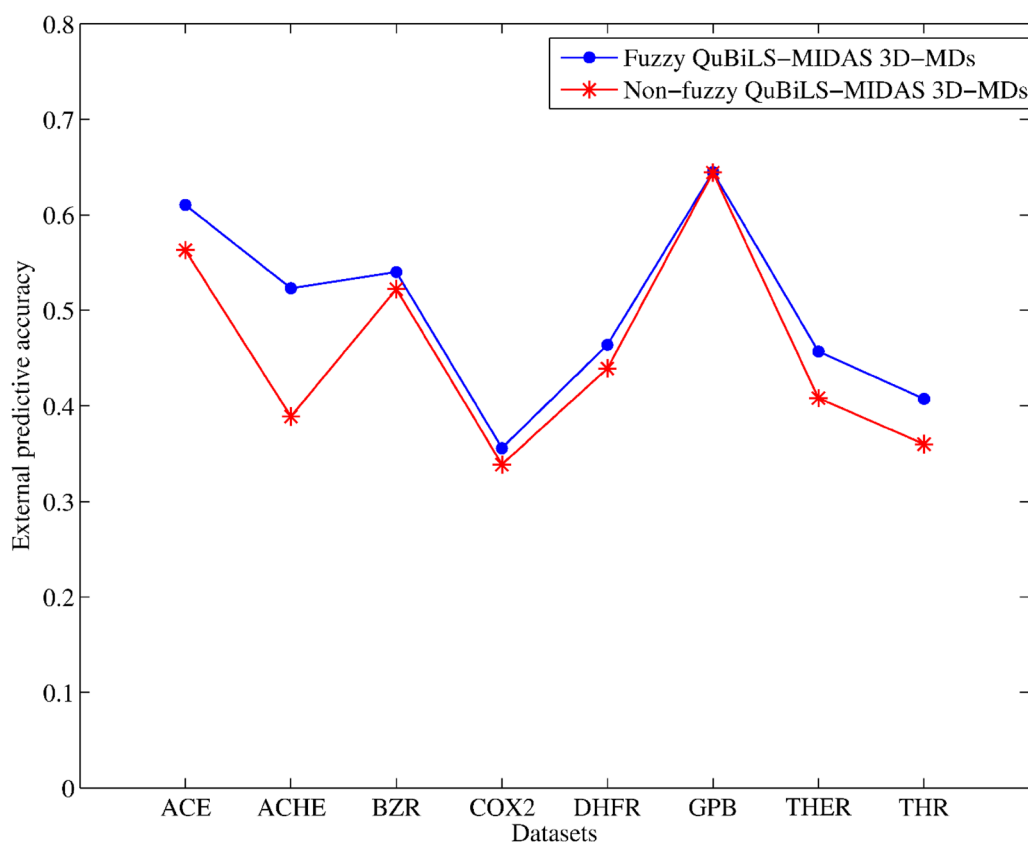
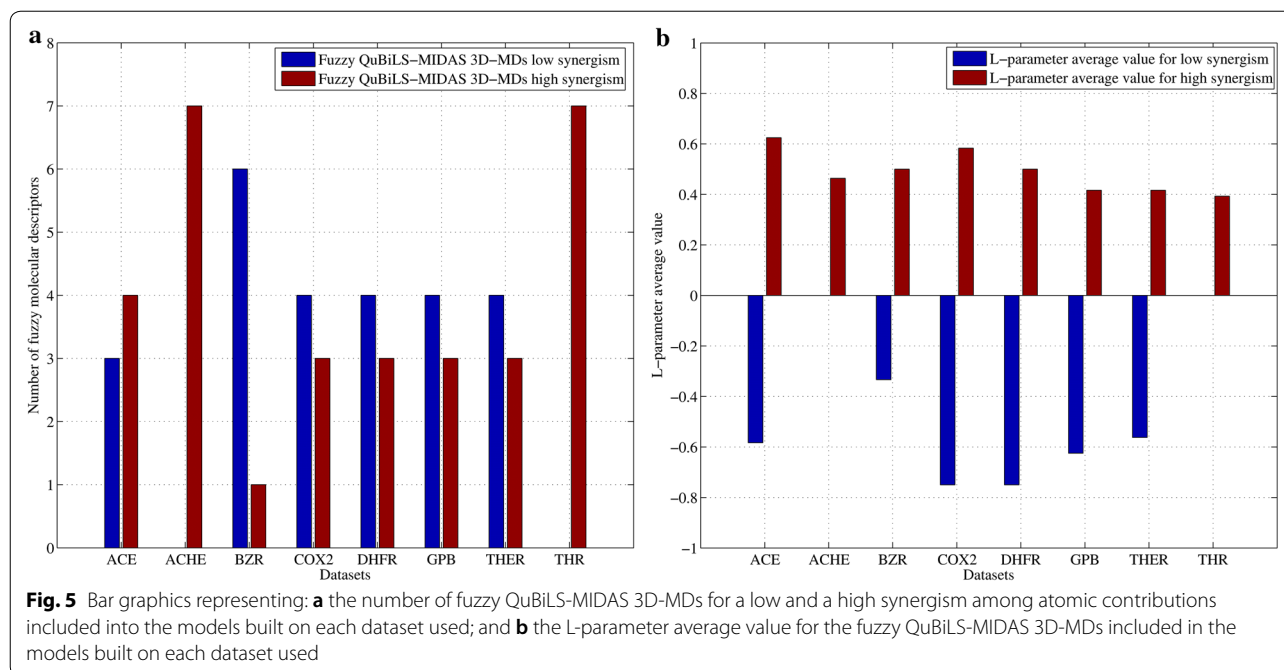


Fig. 4 External predictive accuracies (Q_{ext}^2) corresponding to the models of 7 variables built with the fuzzy and non-fuzzy molecular descriptors, respectively



the behavior for a low synergism is from moderate to high, except in the BZR dataset. Thus, at least preliminarily, it can be stated that high amounts of superadditivity will contribute to compute better MDs in non-congeneric datasets than in congeneric datasets.

Statistical analysis of the performance achieved by the fuzzy and non-fuzzy QuBiLS-MIDAS descriptors

To carry out this analysis, the predictive abilities (Q_{ext}^2) achieved by the models built with the fuzzy and non-fuzzy QuBiLS-MIDAS 3D-MDs on each dataset were accounted for. Figure 6 shows the boxplot graphic corresponding to the Q_{ext}^2 values obtained. Additional file 9 shows the descriptive statistics calculated. On one hand, it can be firstly seen that there are not outlier predictive abilities. In addition, it can be seen that the lowest Q_{ext}^2 attained by the fuzzy models is better than the lowest Q_{ext}^2 attained by the non-fuzzy models; while the highest outcomes are comparable. It can also be observed that the Q_{ext}^2 values obtained with the fuzzy models are distributed almost symmetrically (skewness = 0.095); while the Q_{ext}^2 values obtained with the non-fuzzy models are skewed to the right (skewness = 0.727). These results suggest that the models based on the fuzzy QuBiLS-MIDAS 3D-MDs tend to have a better behavior.

On the other hand, according to the results obtained from the Wilcoxon signed-rank test ($p_{value} \approx 0.008$) (Additional file 10), it can be statistically stated that the Q_{ext}^2 values achieved by the fuzzy models differ to the

ones achieved by the non-fuzzy models. In this sense, if the performances achieved by the models built on each dataset are examined (Additional file 8), it can be appreciated that for the ACHE dataset, the fuzzy model built presents the best progress of all, for a 34.58% of improvement with regard to the non-fuzzy model. Moreover, as for the THR, THERM, ACE, DHFR, COX2 and BZR datasets, the respective fuzzy models improve their predictive abilities a 13.11%, 11.99%, 8.42%, 5.65%, 5.05% and 3.41% with respect to the ones achieved by the non-fuzzy models. Only in the GPB dataset, the improvement of the fuzzy model is insignificant (0.08%). Therefore, in a general sense, it can be concluded that the Choquet integral-based fuzzy 2D/3D-MDs calculation from LOVIs/LOEIs constitutes a prominent alternative to encode relevant chemical information.

Conclusions

General approaches to compute fuzzy 2D/3D-MDs from the contribution of each atom (LOVIs) or bond (LOEIs) within a molecule were introduced, by using the Choquet integral as fuzzy aggregation operator. The Choquet integral is rather different from the other norm-, mean- and statistic-based (non-fuzzy) operators used to date. It performs a reordering step to fuse according to the magnitude of the criteria and, in addition, it considers the interrelation among criteria by using a fuzzy measure. In this work, the fuzzy $L_{m\delta}$ -measure was used to compute

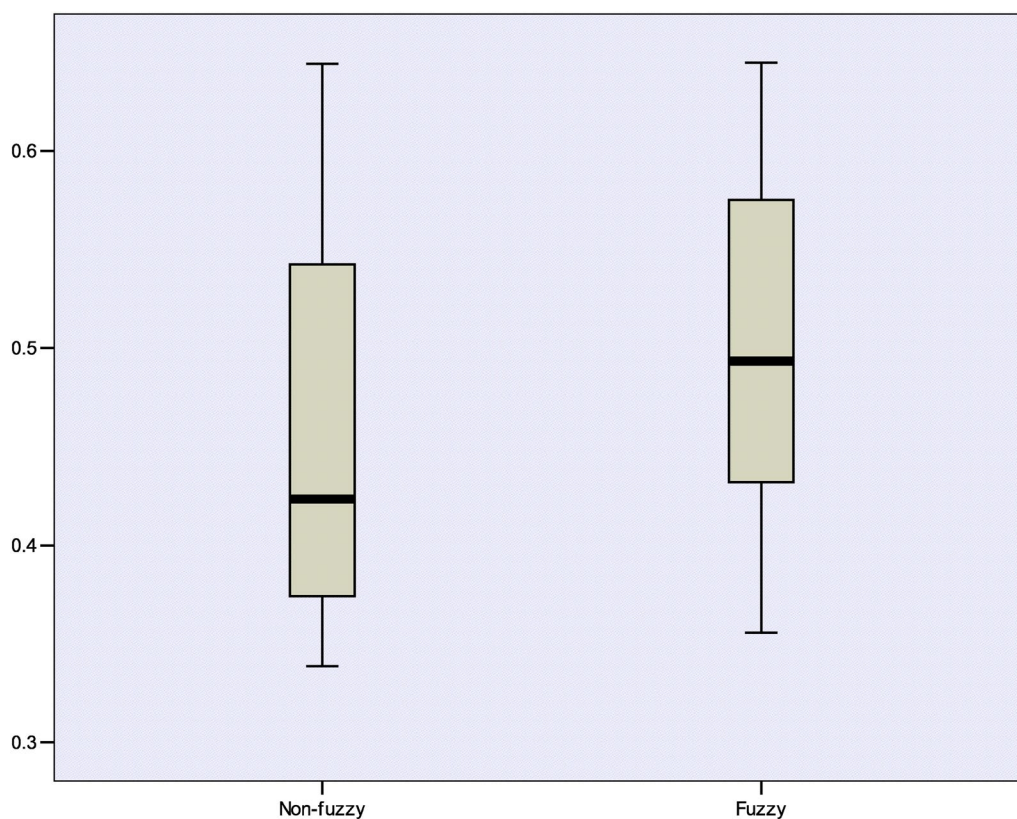


Fig. 6 Boxplot graphic corresponding to the external predictive accuracies achieved by the models built with the fuzzy and non-fuzzy QuBiLS-MIDAS 3D-MDs, respectively

the importance of the interrelation among atom/bond contributions (LOVIs/LOEIs). In this way, fuzzy descriptors can be derived from traditional or recent descriptors; e.g. fuzzy Balaban-like indices.

The feasibility of this proposal was assessed using the QuBiLS-MIDAS descriptors, by performing modeling studies on eight chemical datasets. It was demonstrated that with the Choquet integral-based descriptors, models with better predictive power can be built, if compared to the models built with the descriptors computed from the other non-fuzzy operators. These outcomes were statistically corroborated using the Wilcoxon signed-rank test. All in all, it can be concluded that the use of the Choquet integral as a fuzzy aggregation operator constitutes a prominent way to extract useful structural information of the molecules and, in this way, enhance the modeling capacity of several existing molecular descriptors in ADME-Tox and pharmacological endpoints.

Additional files

Additional file 1. Definition of the Sugeno Fuzzy λ -measure and the Fuzzy P-measure.

Additional file 2. Steroid dataset and projects used to determine the best configurations for the computation of fuzzy densities.

Additional file 3. Matrices of descriptors obtained from the calculation of the projects created in Additional file 2 on the Steroid dataset.

Additional file 4. Statistical parameters for the models built on each descriptors matrix represented in Additional file 3.

Additional file 5. Ranking of the configurations for the computation of fuzzy densities according to the results represented in Additional file 4.

Additional file 6. Chemical datasets and projects to assess the performance between fuzzy and non-fuzzy QuBiLS-MIDAS descriptors.

Additional file 7. Descriptors included in the best models for 3, 5 and 7 variables built on each chemical dataset shown in Additional file 6.

Additional file 8. Best non-fuzzy and fuzzy models of 7 variables built on each chemical dataset shown in Additional file 6.

Additional file 9. Descriptive statistics for the external predictions achieved by the best non-fuzzy and fuzzy models of 7 variables represented in Additional file 8.

Additional file 10. Results of the Wilcoxon test.

Authors' contributions

CRGJ, YMP and FCG proposed the theory of the Choquet integral-based QuBILS-MIDAS 3D-MDs, supervised the QSAR modeling, the design of the GUI and prepared the manuscript. LCL and JSL implemented the GUI designed and performed the QSAR modeling. JSL worked in the revision of the manuscript. MPM and RVY performed the statistical assessment. All authors read and approved the final manuscript.

Author details

¹ Instituto de Química, Universidad Nacional Autónoma de México (UNAM), Ciudad de México, México. ² Grupo de Investigación de Inteligencia Artificial (AIRES), Facultad de Informática, Universidad de Camagüey, Camagüey, Cuba. ³ 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), Quito, Pichincha, Ecuador. ⁴ Grupo de Investigación Ambiental (GIA), Programas Ambientales, Facultad de Ingenierías, Fundación Universitaria Tecnológico Comfenalco – Cartagena, Cr 44 DN 30 A, 91, Cartagena, Bolívar, Colombia. ⁵ Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE), Esmeraldas, Ecuador. ⁶ Grupo de Investigación de Bioinformática, Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba. ⁷ Grupo de Química Cuántica y Teórica, Facultad de Ciencias Exactas y Naturales, Programa de Química, Universidad de Cartagena, Campus de San Pablo, Cartagena, Colombia. ⁸ Grupo CipTec, Facultad de Ingenierías, Fundación Universitaria Tecnológico Comfenalco – Cartagena, Cr 44 DN 30 A, 91, Cartagena, Bolívar, Colombia.

Acknowledgements

CRGJ acknowledges the support from "Dirección General de Asuntos del Personal Académico" (DGAPA) for the postdoctoral fellowship at "Instituto de Química, Universidad Nacional Autónoma de México (UNAM)" in 2016–2018. YMP acknowledges the support from USFQ "Chancellor Grant 2016 (Project ID5454)".

Competing interests

The authors declare no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 28 June 2018 Accepted: 15 October 2018

Published online: 25 October 2018

References

- Tan C (2011) Generalized intuitionistic fuzzy geometric aggregation operator and its application to multi-criteria group decision making. *Soft Comput* 15(5):867–876
- Mokarram M, Hojati M (2017) Using ordered weight averaging (OWA) aggregation for multi-criteria soil fertility evaluation by GIS (case study: southeast Iran). *Comput Electron Agric* 132:1–13
- Marrara S, Pasi G, Viviani M (2017) Aggregation operators in information retrieval. *Fuzzy Set Syst* 324:3–19
- Calvo T et al (2002) Aggregation operators: properties, classes and construction methods. In: Calvo T, Mayor G, Mesiar R (eds) *Aggregation operators: new trends and applications*. Physica-Verlag, Heidelberg, pp 3–104
- Beliakov G, Pradera A, Calvo T (2007) Averaging functions. In: *Aggregation functions: a guide for practitioners*. Springer, Berlin, pp 39–122
- Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans Syst Man Cybern* 18(1):183–190
- Yager RR (1993) Families of OWA operators. *Fuzzy Set Syst* 59(2):125–148
- Merigó JM, Gil-Lafuente AM (2009) The induced generalized OWA operator. *Inf Sci* 179(6):729–741
- Sugeno M (1974) *Theory of fuzzy integrals and its applications*. Tokyo Institute of Technology, Tokyo
- Burkill JC (2004) *The Lebesgue integral*, vol 40. Cambridge University Press, Cambridge, p 108
- Grabisch M, Murofushi T, Sugeno M (2000) Fuzzy measures and integrals: theory and applications. In: Michio S, Toshiaki M (eds) *Studies in fuzziness and soft computing*, vol 40. Physica-Verlag, New York
- Grabisch M, Labreuche C (2016) Fuzzy measures and integrals in MCDA. In: Greco S, Ehrgott M, Figueira JR (eds) *Multiple criteria decision analysis: state of the art surveys*. Springer, New York, pp 553–603
- Choquet G (1954) Theory of capacities. *Annales de l'institut Fourier* 5:131–295
- Marichal JL (2000) An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria. *IEEE Trans Fuzzy Syst* 8(6):800–807
- Marichal J-L (2002) Aggregation of interacting criteria by means of the discrete Choquet integral. In: Calvo T, Mayor G, Mesiar R (eds) *Aggregation operators: new trends and applications*. Physica-Verlag, Heidelberg, pp 224–244
- Narukawa Y (2012) Choquet integral on the real line as a generalization of the OWA operator. In: Torra V et al (eds) *Proceedings of modeling decisions for artificial intelligence: 9th international conference, MDAI 2012, Girona, Catalonia, Spain, November 21–23, 2012*. Springer, Berlin, pp 56–65
- Chițescu I, Plăvițu A (2017) Computing Choquet integrals. *Fuzzy Set Syst* 327(Supplement C):48–68
- Karczmarek P, Kiersztyn A, Pedrycz W (2017) Generalized choquet integral for face recognition. *Int J Fuzzy Syst* 20(3):1047–1055
- Barrenechea E et al (2013) Using the Choquet integral in the fuzzy reasoning method of fuzzy rule-based classification systems. *Axioms* 2(2):208
- Wang Z, Yang R, Leung K-S (2011) Data mining with fuzzy data. In: Zadeh LA (ed) *Nonlinear integrals and their applications in data mining*. World Scientific, Singapore, pp 272–327
- Ferreira JJM, Jalali MS, Ferreira FAF (2018) Enhancing the decision-making virtuous cycle of ethical banking practices using the Choquet integral. *J Bus Res* 88:492–497
- Demirel T et al (2018) Choquet integral-based hesitant fuzzy decision-making to prevent soil erosion. *Geoderma* 313:276–289
- Liu B et al (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
- Bajorath J (2017) Molecular similarity concepts for informatics applications. In: Keith JM (ed) *Bioinformatics: volume II: structure, function, and applications*. Springer, New York, pp 231–245
- Masand VH et al (2017) QSAR modeling for anti-human African trypanosomiasis activity of substituted 2-phenylimidazopyridines. *J Mol Struct* 1130:711–718
- Amin SA et al (2017) An integrated multi-QSAR modeling approach for designing Knoevenagel-type indoles with enhancing cytotoxic profiles. *Curr Comput Aided Drug Des* 13(4):336–345
- De P, Roy K (2018) Greener chemicals for the future: QSAR modeling of the PBT index using ETA descriptors. *SAR QSAR Environ Res* 29(4):319–337
- Todeschini R, Consonni V (2009) *Molecular descriptors for chemoinformatics. Methods and principles in medicinal chemistry*, 2nd edn. Wiley-VCH, Weinheim
- Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev* 96(3):1027–1044
- Barigye SJ et al (2014) Trends in information theory-based chemical structure codification. *Mol Divers* 18(3):673–686
- Randić M, Razinger M (1995) On characterization of molecular shapes. *J Chem Inf Comput Sci* 35(3):594–606
- Balaban AT (1995) Local (atomic) and global (molecular) graph-theoretical descriptors. *SAR QSAR Environ Res* 3(2):81–95
- Marrero-Ponce Y et al (2012) Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application. *J Comput Aided Mol Des* 26(11):1229–1246
- Barigye SJ et al (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

35. García-Jacas CR et al (2014) QuBiLS-MIDAS: a parallel free-software for molecular descriptors computation based on multi-linear algebraic maps. *J Comput Chem* 35(18):1395–1409
36. Cubillán N et al (2015) 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(9):2028–2064
37. Valdés-Martini JR et al (2017) QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *J Cheminform* 9(1):35
38. García-Jacas CR et al (2018) GOWAWA aggregation operator-based global molecular characterizations: weighting atom/bond contributions (LOVIs/LOEIs) according to their influence in the molecular encoding. *Mol Inf*. <https://doi.org/10.1002/minf.201800039>
39. Marrero-Ponce Y et al (2015) Optimum search strategies or novel 3D molecular descriptors: is there a stalemate? *Curr Bioinform* 10(5):533–564
40. García-Jacas CR et al (2014) N-linear algebraic maps to codify chemical structures: is a suitable generalization to the atom-pairs approaches? *Curr Drug Metab* 15(4):441–469
41. Martínez Santiago O et al (2015) Extending graph (discrete) derivative descriptors to n-tuple atom-relations. *MATCH Commun Math Comput Chem* 73(2):397–420
42. Martínez-Santiago O et al (2017) Exploring the QSAR's predictive truthfulness of the novel N-tuple discrete derivative indices on benchmark datasets. *SAR QSAR Environ Res* 28(5):367–389
43. García-Jacas CR et al (2016) Examining the predictive accuracy of the novel 3D N-linear algebraic molecular codifications on benchmark datasets. *J Cheminform* 8(10):1–16
44. Medina Marrero R et al (2015) QuBiLS-MAS method in early drug discovery and rational drug identification of antifungal agents. *SAR QSAR Environ Res* 26(11):943–958
45. Meneses-Marcel A et al (2018) Drug repositioning for novel antitrichomonas from known antiprotozoan drugs using hierarchical screening. *Future Med Chem* 10(8):863–878
46. García-Jacas CR et al (2017) Conformation-dependent QSAR approach for the prediction of inhibitory activity of bromodomain modulators. *SAR QSAR Environ Res* 28(1):41–58
47. Horvath D, Mao B (2003) Neighborhood behavior. Fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity. *QSAR Comb Sci* 22(5):498–509
48. Bonachera F et al (2006) Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *J Chem Inf Model* 46(6):2457–2477
49. Bonachera F, Horvath D (2008) Fuzzy tricentric pharmacophore fingerprints. 2. Application of topological fuzzy pharmacophore triplets in quantitative structure–activity relationships. *J Chem Inf Model* 48(2):409–425
50. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Set Syst* 100(Supplement 1):9–34
51. Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) *Contributions to the theory of games*. Princeton University Press, Princeton, pp 307–317
52. Grabisch M (1997) k-order additive discrete fuzzy measures and their representation. *Fuzzy Set Syst* 92(2):167–189
53. Mohammed MA (2003) Q-measures: an efficient-extension of the Sugeno-measure. *IEEE Trans Fuzzy Syst* 11(3):419–426
54. Liu H-C et al (2010) Composed fuzzy measure of maximized L-measure and delta-measure. *WSEAS Trans Inf Sci Appl* 7(4):474–483
55. Ohlan A (2016) Overview on development of fuzzy information measures. *IJARES* 4(12):17–22
56. Liu HC et al (2007) A novel fuzzy measure and its choquet integral regression model. In: 2007 International conference on machine learning and cybernetics. IEEE, Hong Kong, China
57. Liu HC (2009) Maximized L-measure and its Choquet integral regression model. In: 2009 10th international symposium on pervasive systems, algorithms, and networks. IEEE, Kaohsiung, Taiwan
58. Liu H-C et al (2009) Theory of multivalent delta-fuzzy measures and its application. *WSEAS Trans Inf Sci Appl* 6(6):1061–1070
59. Mohd WRW, Abdullah L (2017) Choquet integral with respect to maximized L-measure and delta-measure. *AIP conference proceedings*, vol 1870, no 1
60. Höhle U (1982) Integration with respect to fuzzy measures. In: *Proceedings of IFAC symposium on theory and applications of digital control*. New Delhi
61. Murofushi T, Sugeno M (1989) An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure. *Fuzzy Set Syst* 29(2):201–227
62. Murofushi T, Sugeno M (1991) A theory of fuzzy measures: representations, the Choquet integral, and null sets. *J Math Anal Appl* 159(2):532–549
63. Tan C, Chen X (2010) Induced choquet ordered averaging operator and its application to group decision making. *Int J Intell Syst* 25(1):59–82
64. Llamazares B (2015) Constructing Choquet integral-based operators that generalize weighted means and OWA operators. *Inf Fusion* 23(Supplement C):131–138
65. Fath-Tabar G (2011) Old and new Zagreb indices of graphs. *MATCH Commun Math Comput Chem* 65(1):79–84
66. Deng H (2011) On the sum-Balaban index. *MATCH Commun Math Comput Chem* 66(1):273–284
67. Diudea MV, Gutman I (1998) Wiener-type topological indices. *Croat Chem Acta* 71(1):21–51
68. Randic M (1975) Characterization of molecular branching. *J Am Chem Soc* 97(23):6609–6615
69. Kier LB, Hall LH (1991) A differential molecular connectivity index. *Mol Inform* 10(2):134–140
70. García-Jacas CR et al (2017) Tensor algebra-based geometric methodology to codify central chirality on organic molecules. *SAR QSAR Environ Res* 28(6):541–556
71. García-Jacas CR et al (2016) N-tuple topological/geometric cutoffs for 3D N-linear algebraic molecular codifications: variability, linear independence and QSAR analysis. *SAR QSAR Environ Res* 27(12):949–975
72. Sinkhorn R, Knopp P (1967) Concerning nonnegative matrices and doubly stochastic matrices. *Pac J Math* 21(2):343–348
73. Xu Z (2005) An overview of methods for determining OWA weights. *Int J Intell Syst* 20(8):843–865
74. García-Jacas CR et al (2015) Multi-server approach for high-throughput molecular descriptors calculation based on multi-linear algebraic maps. *Mol Inform* 34(1):60–69
75. Todeschini R et al (2003) MobyDigs: software for regression and classification models by genetic algorithms. In: Leardi R (ed) *Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks*. Amsterdam, Elsevier, pp 141–167
76. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110(18):5959–5967
77. Schnitker J, Gopalaswamy R, Crippen GM (1997) Objective models for steroid binding sites of human globulins. *J Comput Aided Mol Des* 11(1):93–110
78. Palyulin VA, Radchenko EV, Zefirov NS (2000) Molecular field topology analysis method in QSAR studies of organic compounds. *J Chem Inf Comput Sci* 40(3):659–667
79. Tominaga Y, Fujiwara I (1997) Prediction-weighted partial least-squares regression method (PWPLS) 2: application to CoMFA. *J Chem Inf Comput Sci* 37(6):1152–1157
80. Pastor M et al (2000) GRIND-INdependent Descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* 43(17):3233–3243
81. Westphal U (1983) Corticosteroid-binding globulin. *Mol Cell Biochem* 55(2):145–157
82. Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. *Ann Math Stat* 11(1):86–92
83. 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
84. Klamt A et al (2012) COSMOsar3D: molecular field analysis based on local COSMO σ -profiles. *J Chem Inf Model* 52(8):2157–2164

85. Hinselmann G et al (2011) jCompoundMapper: an open source Java library and command-line tool for chemical fingerprints. *J Cheminform* 3(1):3
86. Bruce CL et al (2007) Contemporary QSAR classifiers compared. *J Chem Inf Model* 47(1):219–227
87. Godden JW, Stahura FL, Bajorath J (2000) Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J Chem Inf Comput Sci* 40(3):796–800
88. Urias RWP et al (2015) IMMAN: free software for information theory-based chemometric analysis. *Mol Divers* 19(2):305–319
89. Wilcoxon F (1945) Individual comparisons by ranking methods. *Biom Bull* 1(6):80–83
90. Alcalá-Fdez J et al (2009) KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Comput* 13(3):307–318

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

