Examining the predictive accuracy of the novel 3D N-linear algebraic molecular codifications on benchmark datasets

César R. García-Jacas\textsuperscript{1,2,3,4*}, Ernesto Contreras-Torres\textsuperscript{8}, Yovani Marrero-Ponce\textsuperscript{3,4,5}, Mario Pupo-Meriño\textsuperscript{2}, Stephen J. Barigye\textsuperscript{3,7} and Lisset Cabrera-Leyva\textsuperscript{1,6}

Abstract

Background: Recently, novel 3D alignment-free molecular descriptors (also known as QuBiLS-MIDAS) based on two-linear, three-linear and four-linear algebraic forms have been introduced. These descriptors codify chemical information for relations between two, three and four atoms by using several (dis-)similarity metrics and multi-metrics. Several studies aimed at assessing the quality of these novel descriptors have been performed. However, a deeper analysis of their performance is necessary. Therefore, in the present manuscript an assessment and statistical validation of the performance of these novel descriptors in QSAR studies is performed.

Results: To this end, eight molecular datasets (angiotensin converting enzyme, acetylcholinesterase inhibitors, benzodiazepine receptor, cyclooxygenase-2 inhibitors, dihydrofolate reductase inhibitors, glycogen phosphorylase b, thermolysin inhibitors, thrombin inhibitors) widely used as benchmarks in the evaluation of several procedures are utilized. Three to nine variable QSAR models based on Multiple Linear Regression are built for each chemical dataset according to the original division into training/test sets. Comparisons with respect to leave-one-out cross-validation correlation coefficients ($Q_{\text{loo}}^2$) reveal that the models based on QuBiLS-MIDAS indices possess superior predictive ability in 7 of the 8 datasets analyzed, outperforming methodologies based on similar or more complex techniques such as: Partial Least Square, Neural Networks, Support Vector Machine and others. On the other hand, superior external correlation coefficients ($Q_{\text{ext}}^2$) are attained in 6 of the 8 test sets considered, confirming the good predictive power of the obtained models. For the $Q_{\text{ext}}^2$ values non-parametric statistic tests were performed, which demonstrated that the models based on QuBiLS-MIDAS indices have the best global performance and yield significantly better predictions in 11 of the 12 QSAR procedures used in the comparison. Lastly, a study concerning to the performance of the indices according to several conformer generation methods was performed. This demonstrated that the quality of predictions of the QSAR models based on QuBiLS-MIDAS indices depend on 3D structure generation method considered, although in this preliminary study the results achieved do not present significant statistical differences among them.

Conclusions: As conclusions it can be stated that the QuBiLS-MIDAS indices are suitable for extracting structural information of the molecules and thus, constitute a promissory alternative to build models that contribute to the prediction of pharmacokinetic, pharmacodynamics and toxicological properties on novel compounds.

Keywords: Multiple Linear Regression, QuBiLS-MIDAS, 3D-QSAR, TOMOCOMD-CARDD

*Correspondence: cesarjacas1985@gmail.com

\textsuperscript{1} Escuela de Sistemas y Computación, Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE), Esmeraldas, Ecuador

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

© 2016 García-Jacas et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

Computational methods that employ statistical and/or artificial intelligence procedures are widely used in the drug discovery process, where the Quantitative Structure–Activity Relationship (QSAR) studies have an important role [1–4]. These studies are based on the principle that the biological activity (or property) of compounds depends on their structural and physicochemical features and thus, are primarily aimed at finding good correlations among molecular features and specific biological activities [5]. In this way, models with high external predictive ability in novel compounds could be built.

Right from the works developed by Hansch and Fujita in 1960s [6, 7], considered as the origins of the modern QSAR studies [8], several approaches have been reported in the literature with most of these being 2D-QSAR methods, that is, they only consider the topological structural features of molecules often using matrix representations such as the connectivity and distance matrices [8]. However, with the introduction of the CoMFA [9] methodology in 1988, the 3D-QSAR approaches become popular. These take into account the geometric (3D) features of molecules, which can be computed either from the information represented in a grid through an alignment process with respect to a reference compound or a pharmacophore [2, 10, 11], or using procedures based on Cartesian coordinates [8, 12, 13], molecular spectra [14, 15] and molecular transforms [16], or by the adaptation of 2D methods to take into account three-dimensional (3D) aspects [17–21].

However, despite the number and variety of procedures defined up to date, there exists continued interest in creating or extending the current approaches to more generalized forms in order to codify more relevant chemical information with the aim of yielding QSAR models with better predictive ability. This assertion is in accordance with the Non Free Lunch Theorem [22], which could be interpreted as no single QSAR procedure yields superior predictions than all the others when its performance is averaged over all possible compound datasets. This can be confirmed in a report performed by Sutherland et al. [23], where it is observed how well-established procedures, assessed in eight diverse chemical datasets, present moderate predictions and without significant differences among them (see Additional file 1: Table S1 for a statistical analysis). The justification for this observation is that one family of molecular descriptors (MDs) may not suffice to codify all chemical information and/or molecular properties for different chemical datasets. In other words, the relevance of MDs depends on the nature of the compounds under study. It is therefore necessary to search for alternative methods/approaches to codify novel and orthogonal chemical information.

Inspired by the previous idea, recently the 3D N-linear algebraic molecular descriptors have been introduced as a novel mathematical procedure for computing the structural features of chemical compounds [24–26]. These MDs employ the bilinear, quadratic and linear algebraic maps [27] to codify information between atom-pairs by using several (dis-)similarity metrics [25]. Also, the N-linear algebraic forms [28] were used as generalized expressions of the bilinear, quadratic and linear algebraic maps, when relations among three and four atoms are studied [26]. In this way, the geometric matrix [8] was extended to consider for the first time relations for more than two atoms.

Several studies aimed at assessing the quality of this novel descriptor family, also called QuBiLS-MIDAS (acronym of Quadratic, Bilinear and N-Linear Maps based on N-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings), were performed and these included an evaluation of the information content (variability) and linear independence using Shannon’s entropy based variability analysis [29] (using IMMAN software [30]) and the principal component analysis (PCA) technique [31], respectively. Also, comparisons with other MDs reported in the literature were performed [25, 26]. In general sense, the results demonstrated that the novel MDs have superior variability than 3D DRAGON indices and another approaches implemented in several software [32–35]. Furthermore, the results revealed that the novel 3D N-linear indices not only do they codify all information contained in the 3D DRAGON MDs, but capture information orthogonal to the latter. Lastly, the QuBiLS-MIDAS MDs were used for modeling the binding affinity to the corticosteroid-binding globulin (CBG), achieving superior results with respect to other QSAR methodologies (see Tables 8–9 in Ref. [25] and Tables 9–10 in Ref. [26]).

However, although the initial results with QuBiLS-MIDAS MDs are promissory, it cannot be stated that these are most suitable for building QSAR models for all chemical datasets. It is thus necessary to evaluate the performance of the 3D N-linear algebraic MDs in QSAR modeling with different molecular sets. Therefore, this paper is dedicated to the assessment of the utility of the QuBiLS-MIDAS approach in the prediction of the biological activity in several compound datasets and the comparison of the obtained results with those of other QSAR procedures reported in the literature.

Mathematical overview of the 3D N-linear algebraic molecular descriptors

In this report, the total and local-fragment 3D N-linear Algebraic indices [25, 26] (also known as QuBiLS-MIDAS) are employed to assess the predictive
accuracy of this approach in QSAR studies. These molecular descriptors (MDs) are calculated from the contribution of each atom in a molecule. That is, if a molecule is comprised of n atoms then the kth two-linear, three-linear and four-linear algebraic indices for each atom “a” are computed as N-linear (Multi-linear) algebraic forms (maps) in $\mathbb{R}^n$, in a canonical basis set, when relations among two ($N = 2$), three ($N = 3$) and four ($N = 4$) atoms are considered, respectively. These descriptors are mathematically expressed as follows:

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

$$u_{(F)L_a} = u^{a,k}_{(F)}(\bar{x}, \bar{y}, \bar{z}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} g^{a,k}_{ijlh(F)} x^i y^j z^l = G^{a,k}_{(F)} \cdot \bar{x} \cdot \bar{y} \cdot \bar{z}$$  \hspace{1cm} (2)

$$q_{(F)L_a} = q^{a,k}_{(F)}(\bar{x}, \bar{y}, \bar{z}, \bar{w}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{h=1}^{n} g^{a,k}_{ijlh(F)} x^i y^j z^l w^h = G^{a,k}_{(F)} \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \cdot \bar{w}$$  \hspace{1cm} (3)

where, “a” is a specific atom (a = 1, 2,...,n), n is the number of atoms in a molecule, $b_{(F)L_a}$ is the entry corresponding to the contribution of the atom “a” in the vector of atom-level indices $(F)L_a$, $F$ is a local-fragment (group or atom-type) that may or not be considered in the index computation, and $x^i, x^j, y^l, y^h, z^l, z^h, w^i, w^j$ are the values (coordinates or components) of the molecular vectors $\bar{x}, \bar{y}, \bar{z}$ and $\bar{w}$, respectively. In addition, the coefficients $g^{a,k}_{ij(F)}$ and $g^{a,k}_{ijlh(F)}$ are the elements of the kth two-tuple, three-tuple and four-tuple atom-level total (or local-fragment) spatial-(dis)similarity matrices $G^{a,k}_{(F)}, G^{a,k}_{(F)}$, and $G^{a,k}_{(F)}$, which are obtained from the corresponding kth two-tuple, three-tuple and four-tuple total (or local-fragment) spatial-(dis)similarity matrices $G^{k}_{(F)}, G^{k}_{(F)}$, and $G^{k}_{(F)}$. Lastly, k ($\pm 1, \pm 2,...,\pm 12$) is the power to which the matrix approaches are raised through the Hadamard product.

The molecular vectors (or property vectors) $\bar{x}, \bar{y}, \bar{z}$ and $\bar{w}$ are calculated by using the Chemistry Development Kit (CDK) library [36] considering the following fragment- and atom-based properties: atomic mass (m), the van der Waals volume (v), the atomic polarizability (p), atomic electronegativity in Pauling scale (e), atomic Ghose-Crippen LogP (a), Gasteiger-Marsili atomic charge (c), atomic polar surface area (psa), atomic refractivity (r), atomic hardness (h) and atomic softness (s).

The total matrix approaches $G^k, G^k_{(F)}$ and $G^k_{(F)}$ constitute the basis for the calculation of the two-linear, three-linear and four-linear indices and these are employed to represent the chemical information codified on interactions among “N” atoms of a molecule. Specifically, for $k = 1$ (matrix of order 1) the coefficients $g^1_{ij}$ and $g^1_{ijlh}$ corresponding to the matrices $G^1, G^1_{(F)}$ and $G^1_{(F)}$ can be calculated by using several (dis)-similarity metrics and multi-metrics to capture the information on the relations between two, three and four atoms, respectively [25, 26]. To compute the atom-pair relations, metrics (see Table 1) derived from the general Minkowski definition (e.g. Manhattan, Euclidean) as well as others that have been successfully used in machine learning algorithms and similarity/dissimilarity studies (e.g. Canberra, Soergel, Clark) are employed. On the other hand, different multi-metrics (see Table 2) to calculate the ternary (three) and quaternary (four) relations among atoms of a molecule can be utilized, such as: bond angle for relations among three atoms and dihedral angle for relations among four atoms. Table 3 shows examples of two-tuple and three-tuple total spatial-(dis)similarity matrices calculated with some previously mentioned metrics and multi-metrics.

From these total matrix approaches ($G^k, G^k_{(F)}$ and $G^k_{(F)}$), local-fragments matrices may be computed in order to consider atom-types or chemical regions of interest and thus yielding the kth two-tuple, three-tuple and four-tuple local-fragment spatial-(dis)similarity matrices, denoted by $G^k_{F}, G^k_{(F)}$ and $G^k_{(F)}$, respectively (see Eq. 13 in Ref. [25] and Eqs. 17–18 in Ref. [26]). Specifically, the local-fragments (or atom-types), $F$, that could be taken into account to compute these indices include: hydrogen bond acceptors (A), carbon atoms in aliphatic chains (C), hydrogen bond donors (D), halogens (G), terminal methyl groups (M), carbon atoms in aromatic portion (P) and heteroatoms (X) (see Table 4 for examples).

These total (or local-fragment) matrix approaches ($G^k_{F}, G^k_{(F)}$ and $G^k_{(F)}$) are also known as kth non-stochastic two-tuple, three-tuple and four-tuple total (or local-fragment) spatial-(dis)similarity matrices denoted by $mG^k_{F}, mG^k_{(F)}$ and $mG^k_{(F)}$ respectively, because no normalizing procedure is used in their computation. Nonetheless, with the purpose of obtaining normalized matrix representations three probabilistic schemes may be employed to compute the QuBLS-MIDAS MDs. In this way, the
following normalized matrix representations are obtained from the corresponding non-stochastic matrices: the \(k\)th simple-stochastic two-tuple, three-tuple and four-tuple total (or local-fraction) spatial-(dis)similarity matrices \([\alpha G^k_F, \alpha GT^k_F, \alpha GQ^k_F]\) (see Eq. 10 in Ref. [25] and Eqs. 13–14 in Ref. [26]), the \(k\)th double-stochastic two-tuple (or local-fraction) spatial-(dis)similarity matrix \([dG^k_F]\) (see Sinkhorn–Knopp algorithm in Ref. [37]) and the \(k\)th mutual probability two-tuple, three-tuple and four-tuple total (or local-fraction) spatial-(dis)similarity matrices \([mp G^k_F, mp GT^k_F, mp GQ^k_F]\) (see Eq. 12 in Ref. [25] and Eqs. 15–16 in Ref. [26]). Table 5 shows the results obtained with the three probabilistic transformations on a two-tuple spatial-(dis)similarity matrix.

Finally, from the non-stochastic (simple-stochastic, double-stochastic or mutual-probability) total (or local-fraction) matrices \([\alpha G^k_F, \alpha GT^k_F, \alpha GQ^k_F]\), the corresponding atom-level matrices (denoted as \(\alpha G_{\text{atom}}^k, \alpha GT_{\text{atom}}^k, \alpha GQ_{\text{atom}}^k\) respectively) are calculated and their coefficients are used in the descriptors calculation (see Eqs. 1–3). Each atom-level matrix determines an atom-level index for atom “a” of a molecule and this value constitutes a component (entry) of the vector \(\varphi L\). Once the vector \(\varphi L\) is computed then the global definition of the \(k\)th two-linear, three-linear and four-linear algebraic indices is obtained by applying over the entries of \(\varphi L\) one or several aggregation operators (see Additional file 1: Table S2 for mathematical definition) [25, 26], which have been successfully employed in other reports [38–40]. In the Scheme 1 a general flowchart regarding the calculation process of the QuBiLS-MIDAS MDs detailed in this section may be observed, while Scheme 2 is a graphic representation of each step employed in the computation of a specific two-linear algebraic index.

In order to automatize the calculation of the 3D N-linear algebraic indices used in the present manuscript the QuBiLS-MIDAS software has been developed [41]. This software has as one of its main features the multi-core processing of the MDs, as well as the option to carry out the distributed calculation of the indices by using the Multi-Server Distributed Computing Platform known as T-arenal [42]. The latter is particularly useful for high-throughput calculation tasks. Both software are freely available via internet at: http://tomocomd.com/.

| Metrics | Formula\(^a\) | Range\(^b\) | Average Range |
|---------|----------------|-------------|---------------|
| Minkowski (M1–M7) | \(d_{XY} = \left( \sum_{j=1}^{n} |x_j - y_j|^p \right)^{1/p}\) | \([0, \infty)\) | \(\bar{d} = \frac{d_{\infty}}{p}\) | \([0, \infty)\) |
| Chebyshev/Lagrange (M8) (Minkowski formula when \(p = \infty\)) | \(d_{XY} = \max\{ |x_j - y_j| \} \) | \([0, n]\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, 1]\) |
| Canberra (M10) | \(d_{XY} = \sum_{j=1}^{n} \frac{|x_j - y_j|}{|x_j| + |y_j|} \) | \([0, 1]\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, 1]\) |
| Lance-Williams/Bray-Curtis (M11) | \(d_{XY} = \frac{\sum_{j=1}^{n} |x_j - y_j|}{\sum_{j=1}^{n} (|x_j| + |y_j|)} \) | \([0, 1]\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, 1]\) |
| Clark/coefficient of divergence (M12) | \(d_{XY} = \sqrt{\sum_{j=1}^{n} \left( \frac{x_j - y_j}{|x_j| + |y_j|} \right)^2} \) | \([0, \infty)\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, \infty)\) |
| Soergel (M13) | \(d_{XY} = \sqrt{\sum_{j=1}^{n} \left( \frac{x_j - y_j}{\max(|x_j|, |y_j|)} \right)^2} \) | \([0, 1]\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, 1]\) |
| Bhattacharyya (M14) | \(d_{XY} = \sqrt{\sum_{j=1}^{n} \left( \sqrt{x_j} - \sqrt{y_j} \right)^2} \) | \([0, \infty)\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, \infty)\) |
| Wave-Edges (M15) | \(d_{XY} = \sqrt{\sum_{j=1}^{n} \left( \frac{1 - \max(x_j, y_j)}{x_j + y_j} \right)^2} \) | \([0, n]\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, 1]\) |
| Angular separation/(1 − Cosine [Ochiai]) (M16) | \(d_{XY} = 1 - \cos_{XY}\) where, \(\cos_{XY} = \frac{x_y}{x_f} = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}}\) | \([0, 2]\) | \(\bar{d} = \frac{d_{\infty}}{n}\) | \([0, 2]\) |

\(^a\) The variables \(x_j\) and \(y_j\) are the values of the coordinate \(j\) of the atoms \(X\) and \(Y\) of a molecule, respectively. The \(h\) value is equal to 3 and corresponds to the 3D Cartesian coordinates \((x, y, z)\) of an atom. The \(p\) values in Minkowski metric are 0.25, 0.5, 1 (Manhattan), 1.5, 2 (Euclidean), 2.5 and 3 (Minkowski)

\(^b\) “Range” refers to “range” and not to “rank” and is defined as \(\text{Range} = \max(x_j) - \min(x_j)\).
Table 2 Measures used to compute the ternary (A) and quaternary (B) relations (multi-metrics) among atoms of a molecule

| Measure | Formula |
|---------|---------|
| **(A) Ternary measures**<br>(T_{XYZ})<br>Perimeter (M19–M20)<br>Triangle area (M21–M22)<br>Sides summation (M25–M26)<br>Bond angle (angle between sides) (m27–m28) | \( T_{XYZ} = d_{xy} + d_{yz} + d_{zx} \)<br>\( T_{XYZ} = \sqrt{(s - d_{xy})(s - d_{yz})(s - d_{zx})} \)<br>\( s = \frac{d_{xy} + d_{yz} + d_{zx}}{2} \)<br>\( T_{XIZ} = d_{xy} + d_{yz} \)<br>\( AX, AY, AZ \) coordinates of three atoms of a molecule<br>\( U = A_{X} - A_{Y} \), \( V = A_{Z} - A_{Y} \)<br>\( T_{XIZ} = \alpha = \arccos \left( \frac{U \cdot V}{|U| \cdot |V|} \right) \) |
| **(B) Quaternary measures**<br>(T_{XYZW})<br>Perimeter (M19–M20)<br>Volume (M23–M24)<br>Sides summation (M25–M26)<br>Dihedral angle (M29–M30) | \( Q_{XYZW} = d_{xy} + d_{yz} + d_{zw} + d_{wx} \)<br>\( Q_{XZWY} = \frac{1}{6} \left( \begin{array}{c} AX - AY \\ AX - AZ \\ AX - AW \\ AX - X \end{array} \right) \times \left( \begin{array}{c} AY - AY \\ AY - AZ \\ AY - AW \\ AY - X \end{array} \right) \)<br>\( Q_{XZWY} = d_{xy} + d_{yz} + d_{zw} \)<br>\( AX, AY, AZ \) coordinates of four atoms of a molecule<br>\( BW, BY, BZ \) coordinates of three atoms of a molecule in the plane A<br>\( U_{a} = (AX - AY) \times (A2 - AY) \)<br>\( U_{b} = (BW - AY) \times (B2 - AY) \)<br>\( Q_{XZWY} = \alpha = \arccos \left( \frac{U_{a} \cdot U_{b}}{|U_{a}| \cdot |U_{b}|} \right) \) |

**Methods**

In order to assess the correlation ability of the QuBiLS-MIDAS MDs for different biological activities eight well-known chemical datasets were used. These were previously employed by Sutherland et al. in a comparative study of QSAR methods commonly used in chemoinformatics analysis [23] and since then, these have been utilized as “benchmarks” for comparing results obtained in other approaches [43–47]. These datasets are comprised by angiotensin converting enzyme (ACE) inhibitors, acetylcholinesterase (AchE) inhibitors, ligands for the benzodiazepine receptor (BZR), cyclooxygenase-2 (COX2) inhibitors, dihydrofolate reductase inhibitors (DHFR), inhibitors of glycosyl phosphorylase b (GPB), thermolysin inhibitors (THER) and thrombin inhibitors (THR). In this study the 3D coordinates were generated using CORINA software, and the same partitioning into training and test sets used in the initial study was considered in order to guarantee comparability of results.

For these datasets, several configurations based on 3D two-linear, three-linear and four-linear algebraic indices were computed (see Additional file 1: Table S3) using the QuBiLS MIDAS software [41]. Due to the fact that numerous MDs are computed with this program yielding a high-dimensional space, then strategies for data reduction are necessary. In this sense, the following workflow for each set of indices calculated for each chemical dataset was performed only considering the training set compounds:

- The 1000 MDs with best variability behavior according to their Shannon’s Entropy values [29] were retained by using the IMMAN software [30].
- The MDs with values represented as power of 10 (scientific notation) and whose exponents are greater or lesser than ±5 were removed.
- Filters for removing the MDs with correlation equal or greater than 0.95 and standardized entropy lesser than 0.3 were applied.
- The statistical method Multiple Linear Regression (MLR) implemented in the STATISTICA software was employed in order to select the MDs included in the model by using Forward Stepwise and Backward Stepwise selection procedures.
- The MDs retained after applying the previous steps and computed for the same compounds were merged into a single dataset.
With the reduced data matrices for each chemical datasets, QSAR models were built with the MLR technique to determine the relationship between the response (activity) and predictor variables (MDs). The MLR technique is coupled with the Genetic Algorithm (GA) meta-heuristic as the variable selection method [48]. This strategy (MLR + GA) is implemented in the MobyDigs software (version 1.0) which was utilized to carry out this study [49]. In this sense, to perform the search process several populations with 100 3D N-linear MDs each were created, while the following configurations were used for the GA procedure: Number of iterations equal to 500,000, Reproduction/mutation trade-off equal to 0.5, Selection bias was initially set to 0 (indicative of random selection) until achieving the 80% of the maximum number of iterations and was later set to 1 (indicates tournament selection) in order to increase the selection pressure. The values of the previous parameters were selected according to the study performed by Todeschini et al. in Ref. [49].

The search process was carried out by using the $Q^2_{loo}$ ("leave-one-out" cross validation) statistical parameter as the fitness function. Once the exploration in each population was completed, then the MDs included in the built 9-variable models were retained with the purpose of creating new populations until 100 MDs. This process is repeated until achieving an only one population with 100 MDs as maximum. Finally, from the final population and for each compound dataset, 3–9 variable regression models were built for the corresponding biological activity. However, as the MobyDigs software generates a set of MLR models then the choice of the most suitable model was performed according to the following steps:

| C1 | C2 | O3 | Cl4 | C1 | C2 | O3 | Cl4 |
|----|----|----|-----|----|----|----|-----|
| 0.000 | 2.408 | 1.439 | 3.939 | 0.000 | 1.000 | 0.973 | 1.000 |
| 2.408 | 0.000 | 1.438 | 1.757 | 1.000 | 0.000 | 0.954 | 0.293 |
| 1.439 | 1.438 | 0.000 | 2.598 | 0.973 | 0.954 | 0.000 | 0.973 |
| 3.939 | 1.757 | 2.598 | 0.000 | 1.000 | 0.293 | 0.973 | 0.000 |

| C1 | C2 | O3 | Cl4 | C1 | C2 | O3 | Cl4 |
|----|----|----|-----|----|----|----|-----|
| 0.000 | 1.158 | 1.003 | 1.709 | 0.000 | 1.354 | 0.558 | 1.875 |
| 1.158 | 0.000 | 1.234 | 1.359 | 1.354 | 0.000 | 0.318 | 0.237 |
| 1.003 | 1.234 | 0.000 | 2.235 | 0.558 | 0.318 | 0.000 | 0.952 |
| 1.709 | 1.359 | 2.235 | 0.000 | 1.875 | 0.237 | 0.952 | 0.000 |

| C1 | C2 | O3 | Cl4 | C1 | C2 | O3 | Cl4 |
|----|----|----|-----|----|----|----|-----|
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.578 |
| 0.000 | 0.000 | 0.578 | 2.470 | 0.000 | 0.000 | 0.000 | 0.281 |
| 0.000 | 1.985 | 0.000 | 2.682 | 1.985 | 0.000 | 0.000 | 0.697 |
| 0.000 | 0.390 | 0.163 | 0.000 | 0.390 | 0.000 | 0.553 | 0.000 |

| C1 | C2 | O3 | Cl4 | C1 | C2 | O3 | Cl4 |
|----|----|----|-----|----|----|----|-----|
| 0.000 | 0.578 | 0.000 | 0.297 | 0.000 | 0.281 | 0.297 | 0.000 |
| 0.578 | 0.000 | 0.000 | 1.892 | 2.470 | 0.000 | 1.892 | 0.000 |
| 0.000 | 0.000 | 0.000 | 0.000 | 2.682 | 0.697 | 0.000 | 0.000 |
| 0.163 | 0.553 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 3 (A) Chemical structure of Chloro(methoxy)methane and its labeled molecular scaffold, (B) examples of two-tuple total spatial-(dis)similarity matrices for $k = 1$ (order) calculated from different (dis-)similarity metrics, (C) example of three-tuple total spatial-(dis)similarity matrix for $k = 1$ (order) calculated from bond angle ternary measure.
Table 4 (A) Two-tuple total spatial-(dis)similarity matrix for \( k = 1 \), \( G^1 \), computed from 3D coordinates of the molecule Chloro(methoxy)methane (see Table 1A), (B) examples of two-tuple local-fragment spatial-(dis)similarity matrices, \( G^2 \), obtained with different chemical fragments.

|       | C1   | C2   | O3   | C14  |
|-------|------|------|------|------|
| C1    | 0.000| 2.408| 1.439| 3.939|
| C2    | 2.408| 0.000| 1.438| 1.757|
| O3    | 1.439| 1.438| 0.000| 2.598|
| C14   | 3.939| 1.757| 2.598| 0.000|

(B) two-tuple local-fragment spatial-(dis)similarity matrices, \( G^2 \):

|       | C1   | C2   | O3   | C14  |
|-------|------|------|------|------|
| C1    | 0.000| 0.000| 0.000| 1.969|
| C2    | 1.204| 0.000| 0.000| 0.000|
| O3    | 0.719| 0.000| 0.000| 0.000|
| C14   | 1.969| 0.878| 1.299| 0.000|

Table 5 Example of probabilistic transformations on the non-stochastic two-tuple total spatial-(dis)similarity matrix for \( k = 1 \), \( m_0 G^1 \), computed from 3D coordinates of the Chloro(methoxy)methane compound (see Table 1A) by using the Euclidean metric.

|       | C1   | C2   | O3   | C14  |
|-------|------|------|------|------|
| C1    | 0.000| 0.387| 0.246| 0.368|
| C2    | 0.387| 0.000| 0.368| 0.246|
| O3    | 0.246| 0.368| 0.000| 0.387|
| C14   | 0.368| 0.246| 0.387| 0.000|

The “best” 50 models according to the \( Q^2_{\text{loo}} \) parameter were retained.

To each model retained the validation methods “bootstrapping” [50] \( (Q^2_{\text{boot}}) \) and “Y-scrambling” [51] \( (\alpha(Q^2)) \) were applied in order to assess the predictive power and the possible chance correlation with respect to the modeled biological activity, respectively. The former randomly creates training sets (with repeated objects) of the same size as the original and the objects left out constitute the test set, while the latter randomly changes the true response variables to determine the quality of the model. Both procedures were repeated 5000 and 300 times, respectively. These methods were applied due to the fact that \( Q^2_{\text{loo}} \) procedure does not suffice to validate the stability of a predictive model [52].

- For each model the function \( f(x) = (1 - Q^2_{\text{boot}}) + |\alpha(Q^2)| \) was computed, which takes into account the results obtained with the two validation procedures employed and the model with the smallest \( f(x) \) value constitutes the “best” regression model.
- The “best” regression model was assessed by using “external validation” \( (Q^2_{\text{ext}}) \) procedure in the corresponding test set in order to measure its generalization ability.

Results and discussion

Assessment of the QuBiLS-MIDAS models versus other approaches

In this section the performance of the QuBiLS-MIDAS models for the chemical datasets described in section “Methods” is compared with respect to 16 QSAR methodologies (or descriptor sets) reported in the literature. The Table 6 shows the statistical parameters and equations of the best regression model based on total and local-fragment QuBiLS-MIDAS indices corresponding to each chemical dataset used in this report. In general sense, it can be observed that the bootstrapping validation coefficient \( (Q^2_{\text{boot}}) \) calculated for each model presents a value greater than 0.6, indicative of the good predictive power of the built models. Also, the coefficients...
computed from scrambling tests ($a(Q^2)$) have in all cases values inferior to 0.4, indicating reduced propensity to chance correlation. Lastly, the values achieved in the external prediction ($Q^2_{\text{ext}}$) suggest that the models based on QuBiLS-MIDAS MDs have appropriate generalization ability, given that all $Q^2_{\text{ext}}$ parameters present values superior to 49% of the total variance even when outlier compounds are retained in the validation set.

On the other hand, the Tables 7 and 8 show the comparisons with respect to other approaches reported in the literature, as well as the results obtained by the models based on total QuBiLS-MIDAS MDs exclusively (see Additional file 1: Table S4 for information related with the best models from 3 to 9 variables). In this manner, the importance of considering local-fragments (atom-types or group) in the calculation of the QuBiLS-MIDAS MDs and subsequently in the building of QSAR models can be analyzed. As can be observed in both tables, the performance of the QuBiLS-MIDAS models is superior when local-fragments are considered with respect to those QuBiLS-MIDAS models that do not use them. Particularly, it can be noted that in 6 of the 8 datasets studied the $Q^2_{\text{loo}}$ parameter is rather comparable, while better performances are attained according to $Q^2_{\text{ext}}$. Both parameters for the COX2 dataset present the best improvements, achieving in the external prediction a value greater than 49% of the total variance, while no other QSAR procedure outperforms this threshold. On the other hand, only in the DHFR and GPB datasets does the utilization of the local-fragment QuBiLS-MIDAS MDs not influence the performance of the developed QSAR models. It can thus be stated that considering a mixture of total and local-fragment QuBiLS-MIDAS MDs in building of QSAR models contributes to the improvement of the predictive ability.

Also, it can be observed from Table 7 that the cross-validation performances achieved by the QuBiLS-MIDAS models have comparable-to-superior behavior with
respect to the approaches reported in the literature. Until now, the best $Q^2_{\text{loo}}$ value for the datasets ACE, ACHE, BZR, COX2, GPB, THER and THR had been attained by the procedures HQSAR (and 2.5D) $Q^2_{\text{loo}} = 0.72$, SAMFA-RF ($Q^2_{\text{loo}} = 0.58$), All-Shortest Path [ASP] Fingerprint ($Q^2_{\text{loo}} = 0.56$), COMSIA extra ($Q^2_{\text{loo}} = 0.57$), HQSAR (and SAMFA-RF) $Q^2_{\text{loo}} = 0.66$, O3A/O3Q ($Q^2_{\text{loo}} = 0.67$) and COMSIA extra ($Q^2_{\text{loo}} = 0.72$), respectively, by using PLS, Random Forest (RF) or Support Vector Machine (SVM) techniques. However, all these previous results are clearly outperformed by the QuBiLS-MIDAS models [(ACE, $Q^2_{\text{ext}} = 0.7422$), (ACHE, $Q^2_{\text{ext}} = 0.5692$), (COX2, $Q^2_{\text{ext}} = 0.4932$), (GPB, $Q^2_{\text{ext}} = 0.8283$), (THER, $Q^2_{\text{ext}} = 0.7248$) and (THR, $Q^2_{\text{ext}} = 0.7674$) test sets outperform the best results reported up to date for each dataset previously mentioned, which correspond to COSMOs-ar3D ($Q^2_{\text{ext}} = 0.43$) in COX2 and to the 2D-FPT methodology in the other datasets [(ACE, $Q^2_{\text{ext}} = 0.713$), (BZR, $Q^2_{\text{ext}} = 0.378$), (GPB, $Q^2_{\text{ext}} = 0.667$), (THER, $Q^2_{\text{ext}} = 0.649$) and (THR, $Q^2_{\text{ext}} = 0.737$)]. The 2D-FPT models were developed by using SQS framework that determines linear and non-linear models (see Table 8), while the model corresponding to COSMOs-ar3D is based on the PLS technique. Even so, the obtained MLR models have better predictive accuracy, even when these are compared with respect to more complex or similar procedures.
### Table 6  Statistical parameters and equations of the best models developed for each chemical dataset analyzed

| Size | $R^2$ | ($Q^2_{	ext{top}}$) | ($Q^2_{	ext{bottom}}$) | a($Q^2$) | ($Q^2_{	ext{ext}}$) | SDEP$_{ext}$ | Models* |
|------|-------|---------------------|----------------------|----------|-------------------|-------------|--------|
| ACE dataset | 6 | 0.814 | 0.7756 | 0.765 | -0.169 | 0.7422 | 1.078 |
| Act | 1.576 ($\pm$1.283) + 0.132 ($\pm$0.018) | $^{SD}_{SS}$, $T_p$ | $^{M20(M1)}$ | -17.977 ($\pm$3.649) | $^{RA}_{SS}$, $B_{a-c}$ | + 2.135 ($\pm$0.071) | $^{RA}_{MP}$, $QuQd_{ext}^{0.257}$ |
| ACH dataset | 8 | 0.738 | 0.6574 | 0.626 | -0.213 | 0.6309 | 0.784 |
| Act | 7.622 ($\pm$0.564) - 0.010 ($\pm$0.004) | $^{SD}_{SS}$, $TrQBM_{ext}^{2.257} | $^{M20(M1)}$ | -0.204 ($\pm$0.046) | $^{AC}_{SS}$, $K_{SS}$ | + 0.417 ($\pm$0.002) | $^{AC}_{SS}$, $TrBM_{ext}^{2.257}$ |
| BZR dataset | 9 | 0.754 | 0.6931 | 0.669 | -0.170 | 0.5692 | 0.631 |
| Act | 8.589 ($\pm$0.592) + 0.160 ($\pm$0.024) | $^{TS}_{SS}$, $K_{SS}$ | $^{M20(M1)}$ | -0.416 ($\pm$0.076) | $^{RA}_{SS}$, $B_{a-c}$ | + 0.018 ($\pm$0.006) | $^{AC}_{SS}$, $TrBM_{ext}^{2.257}$ |
| COX2 dataset | 9 | 0.670 | 0.6313 | 0.615 | -0.091 | 0.4932 | 1.038 |
| Act | -94.390 ($\pm$8.607) + 1.759 ($\pm$0.150) | $^{ES}_{SS}$, $K_{SS}$ | $^{M20(M1)}$ | -0.032 ($\pm$0.009) | $^{AC}_{SS}$, $K_{SS}$ | + 0.317 ($\pm$0.070) | $^{AC}_{SS}$, $TrBM_{ext}^{2.257}$ |
| DHFR dataset | 9 | 0.732 | 0.7055 | 0.697 | -0.077 | 0.6405 | 0.826 |
| Act | 3.127 ($\pm$0.519) + 0.019 ($\pm$0.005) | $^{GS}_{SS}$, $B_{a-c}$ | $^{M20(M1)}$ | -0.500 ($\pm$0.007) | $^{GS}_{SS}$, $B_{a-c}$ | + 0.540 ($\pm$0.004) | $^{GS}_{SS}$, $K_{SS}$ |
| GPB dataset | 8 | 0.893 | 0.8124 | 0.774 | -0.394 | 0.8283 | 0.499 |
| Act | 2.073 ($\pm$0.351) + 0.334 ($\pm$0.078) | $^{AC}_{SS}$, $K_{SS}$ | $^{M20(M1)}$ | -0.147 ($\pm$0.051) | $^{AC}_{SS}$, $K_{SS}$ | + 0.147 ($\pm$0.006) | $^{AC}_{SS}$, $TrBM_{ext}^{2.257}$ |
| THER dataset | 7 | 0.815 | 0.7530 | 0.723 | -0.260 | 0.7248 | 1.197 |
| Act | -11.296 ($\pm$3.480) + 126.508 ($\pm$41.628) | $^{GS}_{SS}$, $B_{a-c}$ | $^{M20(M1)}$ | + 0.016 ($\pm$0.003) | $^{GS}_{SS}$, $B_{a-c}$ | + 0.016 ($\pm$0.006) | $^{AC}_{SS}$, $TrBM_{ext}^{2.257}$ |
| THR dataset | 9 | 0.866 | 0.8149 | 0.789 | -0.286 | 0.7674 | 0.540 |
| Act | 5.251 ($\pm$0.605) - 2120.900 ($\pm$253.086) | $^{TS}_{SS}$, $K_{SS}$ | $^{M20(M1)}$ | -0.000 ($\pm$0.001) | $^{AC}_{SS}$, $K_{SS}$ | + 0.016 ($\pm$0.006) | $^{AC}_{SS}$, $TrBM_{ext}^{2.257}$ |

* See Additional file 1: Table S7 for nomenclature of the QuBILS-MIDAS descriptors

QuBILS-MIDAS approach is inferior to the best results reported so far in the literature. In the former dataset, the methods 2D-FPT ($Q^2_{ext} = 0.714$), O3Q ($Q^2_{ext} = 0.67$) and O3A/O3Q ($Q^2_{ext} = 0.65$) offer better predictions than the proposed model ($Q^2_{ext} = 0.6309$), albeit this can be considered as suitable (explains 63% of total variance). Additionally, when the DHFR test set is taken into account the 2D-FPT approach ($Q^2_{ext} = 0.683$) has more predictive
Table 7 Comparison of the cross-validation statistic parameter ($Q^2_{loo}$) obtained from the QuBiLS-MIDAS models with respect to the performance achieved by 15 QSAR procedures

|          | ACE   | ACHE  | BZR   | COX2  | DHFR  | GPB   | THER  | THR   |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|
| QuBiLS-MIDAS$^a$ | 0.7756 | 0.6574 | 0.6931 | 0.6313 | 0.7055 | 0.8124 | 0.7530 | 0.8149 |
| QuBiLS-MIDAS$^b$ | 0.7773 | 0.6521 | 0.6886 | 0.6664 | 0.7055 | 0.8124 | 0.7495 | 0.8047 |
| CoMFA [23] | 0.68   | 0.52   | 0.32   | 0.49   | 0.65   | 0.42   | 0.52   | 0.59   |
| COMSIA basic [23] | 0.65   | 0.48   | 0.41   | 0.43   | 0.63   | 0.43   | 0.54   | 0.62   |
| COMSIA extra [23] | 0.66   | 0.49   | 0.45   | 0.57   | 0.65   | 0.61   | 0.51   | 0.72   |
| EVA [23] | 0.70   | 0.42   | 0.40   | 0.45   | 0.64   | 0.58   | 0.48   | 0.47   |
| HQSAR [23] | 0.72   | 0.34   | 0.42   | 0.50   | 0.69   | 0.66   | 0.49   | 0.50   |
| 2D [23] | 0.68   | 0.32   | 0.36   | 0.49   | 0.51   | 0.31   | 0.62   | 0.62   |
| 2.5D [23] | 0.72   | 0.31   | 0.35   | 0.55   | 0.53   | 0.46   | 0.66   | 0.52   |
| SAMFA-RF [43] | 0.69   | 0.58   | 0.43   | 0.38   | 0.70   | 0.66   | 0.52   | 0.53   |
| SAMFA-SVM [43] | 0.52   | 0.29   | 0.38   | 0.39   | 0.57   | 0.53   | 0.18   | 0.39   |
| SAMFA-PLS [43] | 0.65   | 0.54   | 0.49   | 0.40   | 0.68   | 0.61   | 0.60   | 0.56   |
| Fingerprints Library [44] | 0.69   | 0.57   | 0.56   | 0.55   | 0.76   | 0.53   | 0.53   | 0.58   |
| O3Q [45] | 0.69   | 0.52   | 0.42   | 0.48   | 0.70   | 0.55   | 0.48   | 0.59   |
| O3QMFA [46] | 0.65   | 0.41   | 0.41   | 0.43   | 0.69   | 0.30   | 0.47   | 0.65   |
| O3A/O3Q [45] | 0.71   | 0.55   | 0.46   | 0.46   | 0.66   | 0.50   | 0.67   | 0.68   |
| COSMOsasr3D [46] | 0.71   | 0.53   | 0.45   | 0.54   | 0.69   | 0.61   | 0.58   | 0.74   |

$^a$ $Q^2_{loo}$ values corresponding to the best model reported considering total and local-fragment QuBiLS-MIDAS indices (see Table 6)

$^b$ $Q^2_{loo}$ values corresponding to the best model reported considering only total QuBiLS-MIDAS indices (see Additional file 1: Table S4)

Italic values correspond to the best results reported in the literature and those obtained by the QuBiLS-MIDAS 3D-MDs

Table 8 Comparison of the external predictive accuracy ($Q^2_{ext}$) attained by the QuBiLS-MIDAS models with respect to the generalization ability achieved with 12 QSAR procedures

|          | ACE   | ACHE  | BZR   | COX2  | DHFR  | GPB   | THER  | THR   |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|
| QuBiLS-MIDAS$^a$ | 0.7422 | 0.6309 | 0.5692 | 0.4932 | 0.6405 | 0.8283 | 0.7248 | 0.7674 |
| QuBiLS-MIDAS$^b$ | 0.7255 | 0.5989 | 0.5459 | 0.4660 | 0.6405 | 0.8283 | 0.7061 | 0.7498 |
| CoMFA [23] | 0.49   | 0.47   | 0.00   | 0.29   | 0.59   | 0.42   | 0.54   | 0.63   |
| COMSIA basic [23] | 0.52   | 0.44   | 0.08   | 0.03   | 0.52   | 0.46   | 0.36   | 0.55   |
| COMSIA extra [23] | 0.49   | 0.44   | 0.12   | 0.37   | 0.53   | 0.59   | 0.53   | 0.63   |
| EVA [23] | 0.36   | 0.28   | 0.16   | 0.17   | 0.57   | 0.49   | 0.36   | 0.11   |
| HQSAR [23] | 0.30   | 0.37   | 0.17   | 0.27   | 0.63   | 0.58   | 0.53   | −0.25  |
| 2D [23] | 0.47   | 0.16   | 0.14   | 0.25   | 0.47   | −0.06  | 0.14   | 0.04   |
| 2.5D [23] | 0.51   | 0.16   | 0.20   | 0.27   | 0.49   | 0.04   | 0.07   | 0.28   |
| O3Q [45] | 0.69   | 0.67   | 0.17   | 0.32   | 0.60   | 0.50   | 0.51   | 0.67   |
| O3QMFA [46] | 0.45   | 0.61   | 0.13   | 0.37   | 0.59   | 0.29   | 0.49   | 0.60   |
| O3A/O3Q [45] | 0.54   | 0.65   | 0.24   | 0.28   | 0.53   | 0.41   | −0.18  | 0.30   |
| COSMOsasr3D [46] | 0.62   | 0.61   | 0.13   | 0.43   | 0.58   | 0.63   | 0.69   | 0.66   |
| 2D-FPT [47] | 0.71$^L$ | 0.71$^N$ | 0.37$^N$ | 0.32$^N$ | 0.68$^N$ | 0.66$^N$ | 0.64$^N$ | 0.73$^N$ |

$^a$ $Q^2_{ext}$ values corresponding to the best model reported considering total and local-fragment QuBiLS-MIDAS indices (see Table 6)

$^b$ $Q^2_{ext}$ values corresponding to the best model reported considering only total QuBiLS-MIDAS indices (see Additional file 1: Table S4)

$^L$ 2D-FPT-based linear models

$^N$ 2D-FPT-based non-linear models

Italic values correspond to the best results reported in the literature and those obtained by the QuBiLS-MIDAS 3D-MDs
ability than the corresponding QuBILS-MIDAS model ($Q^2_{\text{ext}} = 0.6405$), but the latter is superior to the remaining methodologies. Nonetheless, it is important to highlight that the procedures O3Q and O3A/O3Q are alignment dependent and thus their use is generally restricted to congeneric datasets [45]. In the specific case of the 2D-FPT methodology for ACHE and DHFR datasets, the achieved results are based on non-linear models while the proposed outcomes are determined with linear models.

The obtained results evidence that the QuBILS-MIDAS MDs properly codify structural information of the molecules considering interactions among $N$ ($N = 2, 3, 4$) atoms and thus are suitable for developing QSAR models that contribute to the prediction of biological activity in novel structures. However, notwithstanding the comparable-to-superior predictions achieved by the proposed models, it is important to statistically validate these results.

**Statistical analysis of the external predictive accuracy**

To perform this analysis the values corresponding to the external predictions ($Q^2_{\text{ext}}$) obtained by the QuBILS-MIDAS models were taken into consideration as well as the ones reported in the literature over the external compounds belonging to each dataset (see Table 8). Firstly, a descriptive analysis through boxplot graphics was performed (with SPSS software) and the obtained results are represented in Fig. 1. As can be observed, the QuBILS-MIDAS and 2D-FPT models tend to have a similar behavior and superior to the remaining procedures. Also, it can be noted that the highest prediction among all procedures analyzed is achieved by the QuBILS-MIDAS models. In addition, taking into account the graphics corresponding to the QuBILS-MIDAS and 2D-FPT approaches, it can be concluded that the predictions obtained by the former are less scattered than those attained by the latter and thus, the QuBILS-MIDAS models have a more suitable external predictive ability irrespective of the chemical dataset analyzed. However, these results are not enough to state that the models based on QuBILS-MIDAS MDs are statistically the best.

Therefore, an exploratory study was performed to analyze the normality of the data by using Kolmogorov–Smirnov (K–S) test corrected by Lilliefors [53] and the Shapiro–Wilk test [54]. This was done in order to guarantee that the variable $Q^2_{\text{ext}}$ is not normally distributed, at least for one model, and so to ensure that the non-parametric tests are the proper choice. As can be observed in Additional file 1: Table S5, the null hypotheses of normality can only be rejected with a high certainty for $Q^2_{\text{ext}}$ values in the 2D-FTP and COSMOsar3D models, although with Shapiro–Wilk test the rejection of the null
hypothesis is achieved for COMSIA basic as well. Therefore the non-parametric tests may be considered as suitable for this statistical analysis.

Subsequently, a Friedman test [55] for multiple comparisons was performed taking into consideration the results of all QSAR procedures. As can be seen in Additional file 1: Table S6A, there are global differences among the considered methods, with the QuBiLS-MIDAS models being those with the best performance followed by the 2D-FPT, O3Q and COSMOsar3D approaches, respectively, with a Kendall’s W [56] concordance level of 0.607 (see Additional file 1: Table S6B). In order to determine the specific statistical differences a Wilcoxon signed-ranks test [57] was carried out (with R software) by using Benjamini and Hochberg [58] (BH) as the adjustment method (one-tailed p values calculation) for controlling the false discovery rate (FDR). The results of this analysis are shown in Table 9, where a significant p value (p value <0.05) means that the row approach is superior to the corresponding column. So, it can be noted that the QuBiLS-MIDAS models yield statistically better predictions than the other methodologies considered, with the exception of the 2D-FPT approach.

Table 10 External predictive accuracy achieved by QSAR models developed from 3D molecular structures generated with six different programs

|        | ACE   | ACHE  | BZR   | COX2  | DHFR  | GPB   | THER  | THR   | Rank average |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|
| BALLOON| 0.3296| 0.1943| 0.3949| 0.2451| 0.3758| 0.0000| 0.0000| 0.0000| 4.5          |
| CHEMAXON| 0.5504| 0.1343| 0.4163| 0.3361| 0.2978| 0.1687| 0.0000| 0.1386| 3.375        |
| CORINA | 0.4133| 0.0556| 0.3628| 0.2865| 0.4288| 0.2767| 0.1915| 0.2334| 3.25         |
| FROG2  | 0.4832| 0.3535| 0.3635| 0.3393| 0.3786| 0.2712| 0.3264| 0.1457| 2.125        |
| OPENBABEL| 0.3993| 0.1306| 0.1715| 0.2775| 0.3460| 0.4742| 0.2806| 0.0803| 4            |
| RDKit  | 0.4181| 0.1770| 0.3024| 0.2189| 0.5008| 0.4511| 0.0000| 0.0710| 3.75         |

Analysis of the predictive ability according to conformer generation methods

The conformer generation constitutes an important step when chemoinformatics tasks are performed, particularly in the computer-aided drug design, where the outcomes of a virtual screening process may depend on 3D structures employed to build the procedure to be used, e.g. a QSAR model [59]. Therefore, in this section an evaluation of the sensibility of the QuBiLS-MIDAS MDs to the different conformer generation methods is performed in order to comprehend how these could affect in the performance of the indices. To this end, the software FROG2 [60], RDKit [61], BALLOON [62], OpenBabel [63] and Standardizer ChemAxon [64] were employed to generate the 3D structures, taking as starting point the SMILES representations corresponding to the eight compound datasets considered in this report.

Firstly, a study with the purpose of knowing if the models developed using the training structures generated with CORINA (see Table 6) are applicable to the test structures generated with the previously mentioned programs was performed. The external predictive abilities obtained after performing this study are graphically represented in Fig. 2. These results are significantly inferior to those achieved with the test sets based on CORINA (see Additional file 1: Table S8), with the exception of RDKit. This demonstrates that QSAR models based on QuBiLS-MIDAS MDs are not suitable to predict biological activity into compounds optimized with other procedure different from than used for the training structures. Thus, it can be stated that the performance of the QuBiLS-MIDAS MDs depend on 3D conformations from which are computed.

It is important to highlight that the previous results do not mean that CORINA software is the most suitable to generate the 3D structures to be used in the development of the QSAR models based on QuBiLS-MIDAS MDs. In this sense, in order to prove this assertion the following simple workflow was carried out considering the

![Boxplot graphic for the external predictive accuracy achieved by the QSAR models reported in this manuscript (see Table 6) and fitted using structures generated by CORINA software, over the corresponding test sets optimized by five different toolkits](image)
conformations generated by each previously mentioned program (including CORINA) for each chemical dataset:

- 8640 two-linear algebraic indices (Additional file 1: Table S9) were computed.
- CfsSubsetEval feature selection procedure, implemented in WEKA software, was applied in order to retain those MDs with high correlation according to dependent-variable and with low intercorrelation among them.
- The MLR-GA procedure implemented in MobyDigs software was employed to build 9-variable models performing 100,000 iterations and considering the tabu list options of removing MDs with correlation equal or greater than 0.95, fourth order moment greater than 8 and standardized entropy lesser than 0.3. The fitness function used was the statistical parameter $Q^2_{loo}$.
- The model with the highest $Q^2_{loo}$ value was selected as the best model, to which the external predictive ability was determined.

Table 10 shows the external predictive power of the models developed from different 3D conformations, as well as the average of the rankings corresponding to the conformer generation methods considered in this study. As can be observed, the best predictions are achieved by the models built from 3D molecular structures generated by FROG2 procedure, followed by the results obtained from the methods CORINA, CHEMAXON, RDKIT, OPENBABEL and BALLOON, respectively. However, in Additional file 1: Table S10 is demonstrated through a Friedman test that there exists no global statistic differences among previous predictions, which proves, at least for this preliminary study that with QuBiLS-MIDAS MDs can be developed QSAR models with good predictive accuracy irrespective of the procedure used to obtain optimized structures.

Note that for the forthcoming version of QuBiLS-MIDAS software, RDKIT program will be incorporated in the QuBiLS-MIDAS software as a built-in option for conformer generation. This is due to the fact that FROG2 procedure can only be accessed using a web browser, while CORINA and CHEMAXON software are not freely available for use. In addition, according to a study performed in Ref. [65] in order to assess the quality of the conformations generated by several free methods, RDKIT tends to generate the most similar conformations to the experimental structures, in addition to being the second fastest among all toolkits analyzed.

Conclusions
In this report the predictive accuracy of the novel alignment-free geometric molecular descriptors based on N-linear algebraic maps (so called QuBiLS-MIDAS) has been examined. To this end, QSAR models for predicting the biological activity in eight molecular datasets were developed by using MLR as statistical technique. The results obtained with the QuBiLS-MIDAS models were compared with respect to several QSAR procedures reported in the literature according to the correlation coefficients achieved with the leave-one-out cross-validation ($Q^2_{loo}$) and external prediction ($Q^2_{ext}$) methods, and generally superior performances were observed with this QuBiLS-MIDAS framework.

A few exceptions were observed: for the $Q^2_{loo}$ parameter, the QuBiLS-MIDAS approach is exclusively outperformed by the ASP-based (fingerprint) method in the DHFR dataset, while for the $Q^2_{ext}$ parameter, the QuBiLS-MIDAS method yields inferior results with respect to the 2D-FPT methodology in the DHFR and ACHE test set, respectively. Also, inferior $Q^2_{ext}$ values are yielded by the QuBiLS-MIDAS approach with respect to the O3Q and O3A/O3Q procedures in the ACHE test set. However, these previous methodologies are based on techniques more complex than MLR and/or cannot be used in non-congeneric datasets because are alignment-depend. Thus, considering the maximum parsimony principle (Ockham's razor), the QuBiLS-MIDAS approach seems to be more suitable than the other QSAR methods.

Additionally, several steps for statistically validating the obtained results are detailed. In this sense, the external predictive ability of the developed models was compared with respect to other methodologies by means of the multiple comparison tests. It was demonstrated that the QuBiLS-MIDAS models yield the best predictions, and that these are significantly superior in 11 of the 12 methodologies compared. Therefore, it can be suggested that the 3D Algebraic N-linear molecular descriptors (also known as QuBiLS-MIDAS) are suitable for extracting structural information of the molecules and thus, constitute a promissory alternative to build models that contribute to the prediction of pharmacokinetic, pharmacodynamics and toxicological properties of novel compounds.

Additional file

Additional file 1. The statistical parameters and equations of the 3–9 variable QSAR models developed and the corresponding outputs for the statistical analysis performed.

Authors' contributions
CRGJ proposed the theory of the QuBiLS-MIDAS indices, supervised the QSAR modeling on the eight chemical datasets, performed the study about the performance of the indices according to several structure generation methods and prepared the manuscript. ECT worked in the QSAR modeling on the eight chemical datasets. YMP leaded the research related with this manuscript. MPM performed the statistical analysis. SJB worked in the definition of
the QuBILS-MIDAS indices and prepared the manuscript. LCL worked in the QSAR modeling on the eight datasets. All authors read and approved the final manuscript.

Author details
1 Escuela de Sistemas y Computación, Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE), Esmeraldas, Ecuador. 2 Grupo de Investigación de Bioinformática, Centro de Estudio de Matemática Computacional (CEMC), Escuela de Ciencias Informáticas, La Habana, Cuba. 3 Computer-Aided Molecular Biosilico Discovery and Bioinformatics Research International Network (CAMAD-BIR IN), Cumbayá, Quito, Ecuador. 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 Escuela de Medicina, Colegio de Ciencias de la Salud, Edificio de Especialidades Médicas, Hospital de los Valles, Universidad San Francisco de Quito (USFQ), Av. Interoceánica Km 12½ - Cumbayá, Quito, Ecuador. 6 Grupo de Investigación de Inteligencia Artificial (AIRES), Facultad de Informática, Universidad de Caramayue, Caramayue, Cuba. 7 Departamento de Química, Universidad Federal de Lavras, UFLA, Caixa Postal 3037, Lavras, MG 37200-000, Brazil. 8 Departamento de Técnicas de Programación, Facultad 6, Universidad de las Ciencias Informáticas, La Habana, Cuba.

Acknowledgements
García-Jacas, CR thanks the program “International Professor” for a fellowship to work at “Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE)” in 2015–2016. Barigye, SJ acknowledges support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq): M-P Y acknowledge also the partial financial support from Colegio de Medicina, USFQ. Finally, but not least, this work was supported in part by ISC-USFQ.

Competing interests
The authors declare that they have no competing interests.

Received: 14 October 2015 Accepted: 9 February 2016 Published online: 25 February 2016

References
1. Norinder U (1996) Single and domain mode variable selection in 3D QSAR applications. J Chemom 10(2):95–105
2. Sung-Sau S, Karplus M (1997) Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. J Med Chem 40(26):4360–4371
3. Aires-de-Sousa J, Gasteiger J (2002) Prediction of enantiomeric selectivity in chromatography: application of conformation-dependent and conformation-independent descriptors of molecular chirality. J Mol Graph Model 20(5):373–388
4. Chen H, Zhou J, Xie G (1998) A new substituent constant, π, derived from local vertex invariants. Curr Comput Aided Drug Des 1704–1709
5. Kubinyi H (1997) QSAR and 3D QSAR in drug design: 1. Methodology. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in drug design. Kluwer/ESCOM, Dordrecht, pp 339–353
6. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evolut Comput 1(1):67–82
7. Sutherland JJ, O’Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure–activity relationships. J Med Chem 47(2):5541–5554
8. Balaban AT (1997) From chemical topology to three-dimensional geometry. Springer, New York
9. Todeschini R, Consonni V (2000) Molecular descriptors for the prediction of biological activities. J Chem Inf Comput Sci 40(3):796–800
10. Steinbeck C et al (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43(2):493–500
11. Consonni V, Todeschini R, Pavani M (2002) Structure/response correlation and similarity/diversity analysis by GETAWAY descriptors. Part 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42(3):682–692
12. Todeschini R, Lasagni M, Marengo E (1994) New molecular descriptors for 2D and 3D structures. Theory. J Chemom 8(4):263–272
13. Consonni V, Todeschini R, Pavani M (2002) Structure/response correlation and similarity/diversity analysis by GETAWAY descriptors. Part 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42(3):682–692
14. Buri R et al (1999) Comparative spectra analysis (CoSA): spectra as three-dimensional molecular descriptors for the prediction of biological activities. J Chem Inf Comput Sci 39(5):861–867
15. Turner DB et al (1999) Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies. 2. Model validation using a benchmark steroid dataset. J Comput Aided Mol Des 13(3):271–296
16. Gasteiger G et al (1996) Chemical information in 3D space. J Chem Inf Comput Sci 36(5):1030–1037
17. Balaban AT (1997) From chemical topology to three-dimensional geometry. Springer, New York
18. Bogdanov B, Nikolić S, Trinajstić N (1998) On the three-dimensional Wiener number: a comment. J Math Chem 5(3):305–306
19. Mekenyan O et al (1986) Modelling the interaction of small organic molecules with biomacromolecules. I. Interaction of substituted pyridines with anti-3-azopyridine antibody. Arzneim Forsch 36(2):176–183
20. Randić M (1995) Molecular profiles novel geometry-dependent molecular descriptors. New J Chem 19:781–791
21. Pearlman RS, Smith KM (1998) Novel software tools for chemical diversity. In: Kubinyi H, Folgers G, Martin YC (eds) 3D QSAR in drug design. Kluwer/ESCOM, Dordrecht, pp 339–353
22. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evolut Comput 1(1):67–82
23. Sutherland JJ, O’Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure–activity relationships. J Med Chem 47(2):5541–5554
24. Cubillán N et al (2015) Novel global and local 3D atom-based linear descriptors of the Minkowski distance matrix: theory, diversity–variability analysis and QSAR applications. J Math Chem 53(9):2028–2064
25. Mancero-Ponce Y et al (2015) Optimum search strategies or novel 3D molecular descriptors: is there a stalemate? Curr Bioinf 10(5):533–564
26. García-Jacas CR et al (2014) N-linear algebraic maps to codify chemical structures: is a suitable generalization to the atom-pairs approachs? Curr Drug Metab 15(4):441–469
27. Edwards CH, Penney DE (1988) Elementary linear algebra. Prentice Hall, Englewoods Cliffs
28. Johnson JW, Huang CH, Johnson JR (1991) Multilinear algebra and parallel programming. J Supercomput 5(2–3):189–217
29. 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
30. Uras RW et al (2015) IMMAN: free software for information theory-based chemometric analysis. Mol Divers 19(2):305–319
31. Somorjai RL (2010) Multivariate statistical methods. In: John L (ed) Encyclopedia of spectroscopy and spectrometry. Academic Press, Oxford, pp 1704–1709
32. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Chemom 32(7):1466–1474
33. Georg H (2008) BlueDesc-molecular descriptor calculator. University of Tubingen, Tübingen
34. Hong H et al (2008) MolsEd, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Comput Sci 48(7):1337–1344
35. Mauri A et al (2006) DRAGON software: an easy approach to molecular descriptor calculations. Match 56(2):237–248
36. Steinbeck C et al (2003) The Chemistry-Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43(2):493–500
37. Sinkhorn R, Knopp P (1967) Concerning nonnegative matrices and doubly stochastic matrices. Pacific J Math 21(2):343–348
38. 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
39. Barigye SJ et al (2013) Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices. J Comput Chem 34(4):259–274
40. 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
41. 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
42. 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
43. Manchester J, Czerminski R (2008) SAMFA: simplifying molecular description for 3D-QSAR. J Chem Inf Model 48(6):1167–1173
44. Hinselmann G et al (2011) JCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. J Cheminform 3(1):3
45. Tosco P, Balle T (2011) A 3D-QSAR-driven approach to binding mode and affinity prediction. J Chem Inf Model 52(2):302–307
46. Klamt A et al (2012) COSMOsor3D: molecular field analysis based on local COSMO ω-profiles. J Chem Inf Model 52(8):2157–2164
47. Bonachéra 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
48. Leardi R, Boggia R, Terriile M (1992) Genetic algorithms as a strategy for feature selection. J Chemom 6(5):267–281
49. 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. Elsevier, Amsterdam, pp 141–167
50. Wu CFJ (1986) Jackknife, bootstrap and other resampling methods in regression analysis. Ann Stat 14(4):1261–1295
51. Lindgren F et al (1996) Model validation by permutation tests: applications to variable selection. J Chemom 10(5–6):521–532
52. Eilisseff A, Pontil M (2003) Leave-one-out error and stability of learning algorithms with applications. NATO science series sub series III computer and systems sciences, vol 190, pp 111–130
53. Lilliefors HW (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 62(318):399–402
54. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591–611
55. Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 11(1):86–92
56. Siegel S (1957) Nonparametric statistics. Am Stat 11(3):13–19
57. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1(6):80–83
58. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc A 57(1):289–300
59. Hechinger M, Leonhard K, Marquardt W (2012) What is wrong with quantitative structure-property relations models based on three-dimensional descriptors? J Chem Inf Model 52(8):1984–1993
60. Mteva MA, Gyanon F, Tuffery P (2010) Frog2: efficient 3D conformation ensemble generator for small compounds. Nucleic Acids Res 38(suppl 2):W622–W627
61. RDKit: cheminformatics and machine learning software. February 2, 2016; http://www.rdkit.org/
62. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47(6):2462–2474
63. O’Boyle N et al (2011) Open Babel: an open chemical toolbox. Jcheminform 3(1):33
64. Standardizer ChemAxon 5.9.0. February 2, 2016. https://www.chemaxon.com/products/standardizer/
65. Ebejer JP, Morris GM, Deane CM (2012) Freely available conformer generation methods: how good are they? J Chem Inf Model 52(5):1146–1158

Submit your manuscript to a SpringerOpen journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at springeropen.com