Title: Prediction of aquatic toxicity of benzene derivatives using molecular descriptor from atomic weighted vectors.

Authors: Yoan Martínez-López, Stephen J. Barigye, Oscar Martínez-Santiago, Yovani Marrero-Ponce, James Green, Juan A. Castillo-Garit

PII: S1382-6689(17)30292-2
DOI: https://doi.org/10.1016/j.etap.2017.10.006
Reference: ENVTOX 2898

To appear in: *Environmental Toxicology and Pharmacology*

Received date: 26-5-2017
Revised date: 9-10-2017
Accepted date: 11-10-2017

Please cite this article as: Martínez-López, Yoan, Barigye, Stephen J., Martínez-Santiago, Oscar, Marrero-Ponce, Yovani, Green, James, Castillo-Garit, Juan A., Prediction of aquatic toxicity of benzene derivatives using molecular descriptor from atomic weighted vectors.Environmental Toxicology and Pharmacology https://doi.org/10.1016/j.etap.2017.10.006

Prediction of aquatic toxicity of benzene derivatives using molecular descriptor from atomic weighted vectors.

Yoan Martínez-López[a,b], Stephen J. Barigye[c], Oscar Martínez-Santiago[b], Yovani Marrero-Ponce[d], James Green[e] and Juan A. Castillo-Garit[b,e,f,*]

[a]*Department of Computer Sciences, Faculty of Informatics, Camaguey University, Camaguey City, 74650, Camaguey Cuba.*

[b]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Universidad Central "Martha Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*

[c]*Departamento de Química, Universidade Federal de Lavras, CP 3037, 37200-000, Lavras, MG, Brazil.*

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

[e]*Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada.*

[f]*Unidad de Toxicologia Experimental, Universidad de Ciencias Médicas de Villa Clara Santa Clara, 50200, Villa Clara, Cuba.*

***To whom correspondence should be addressed:**
Unidad de Toxicología Experimental, Universidad de Ciencias Médicas de Villa Clara, Carretera a acueducto y Circunvalación, Santa Clara, Villa Clara, Cuba. CP: 50200, Cuba. Tel: +53-42273236,
e-mail: jacgarit@yahoo.es

Highlights

- 1-MD-AWV are calculated using MD-LOVIs software and various Aggregation Operators
- 2-We develop several QSAR models to predict aquatic toxicity of benzenes derivatives

- 3-Our models compares favorably with other previously published with the same dataset
- 4-These descriptors provide an effective alternative for determining aquatic toxicity

Abstract

Several descriptors from atom weighted vectors are used in the prediction of aquatic toxicity of set of organic compounds of 392 benzene derivatives to the protozoo ciliate *Tetrahymena pyriformis* ($\log(IGC50)^{-1}$). These descriptors are calculated using the MD-LOVIs software and various Aggregation Operators are examined with the aim comparing their performances in predicting aquatic toxicity. Variability analysis is used to quantify the information content of these molecular descriptors by means of an information theory-based algorithm. Multiple Linear Regression with Genetic Algorithms is used to obtain models of the structure–toxicity relationships; the best model shows values of Q^2= 0.830 and R^2=0.837 using six variables. Our models compare favorably with other previously published models that use the same data set. The obtained results suggest that these descriptors provide an effective alternative for determining aquatic toxicity of benzene derivatives.

Keywords: Aggregation Operator, Aquatic Toxicity, Atom Weighted Vector, Molecular Descriptor, Multiple Linear Regression, Variability Analysis

1. Introduction

Benzene derivatives are widely used in industry as pesticides, insecticides, herbicides, lubricants, detergents, polymers, solvents, and in the manufacturing of plastics, resins and nylon [1, 2]. Likewise, in the pharmaceutical industry, they are often used as drugs.

Many of these derivatives can cause damage to the environment, humans, animals, and plants due to their toxicity. Environmental contamination, in addition to the possible accumulation of substituted benzenes in water and soil, make them potentially damaging chemicals [1, 3]. There is increasing interest in the toxicity of environmental chemicals. However, experimental measurements are costly and time-consuming. Therefore, studying quantitative structure-activity/toxicity relationships (QSAR/QSTR) provide an invaluable tool in the prediction of aquatic toxicity directly from the molecular structure of compounds [3].

QSARs are employed as scientifically credible tools for predicting the acute toxicity of chemicals when few empirical data are available. Consistent with the development and application of QSARs for the design of more efficacious pharmaceuticals and pesticides, there has been increasing acceptance of using structure-activity relationships to predict adverse effects of xenobiotics in risk assessment [4-8].

The development of an ecotoxicity-based QSAR requires three components: 1) a dataset of chemicals, most often organic, defined by some selection criteria; 2) a measure of toxicity for each chemical; 3) molecular structure and/or property data (i.e. the descriptors, variables, or predictors) for each chemical. The molecular descriptors must then be related to toxicity, normally via statistical methods [9].

The inhibition of growth database [10] of the ciliated protozoan *Tetrahymena pyriformis* is considered to be a high quality dataset [11]. It has been developed in a single laboratory over more than two decades. While slight variations in the static protocol and nominal concentrations exist within the data, the dataset remains an excellent primary source of information; it is also unique in terms of its size, molecular diversity, and quality [5]. Therefore, several studies have attempted to predict the toxicity of benzene derivatives toward *T. pyriformis* using the known toxicity database of 392 benzene

derivatives. For example, Bordbar *et al* built a model using DRAGON software to extract molecular descriptors based on constitutional, topological, molecular walk counts, aromaticity indices, geometrical, Weighted Holistic Invariant Molecular (WHIM), functional group, empirical and other properties [12]. Salahinegad and Ghasemi conducted 3D-QSAR studies following the comparative molecular field analysis Comparative (CoMFA), Molecular Similarity Index Analysis (CoMSIA) and VolSurf approaches [3] Several other applications of QSAR in toxicology are presented in comprehensive reviews [13-19].

Furthermore, our research group have obtained several remarkable results in the identification/selection of new molecular entities with potential action against different targets [20-22] with several families of molecular descriptors (MD) [23-25]. Therefore, the aim of this paper is to compare the performance of a new family of MD, derived from Atomic Weighted Vectors (AWV), [26] in predicting the aquatic toxicity of a large group of substituted benzenes. These MD are derived from AWV using Aggregation Operators (AOs) to convert the AWV into scalar quantities describing aspects of the molecule. Furthermore, different "atomic groupings" are applied such that the computed MD reflects various atomic subsets of the structure under study. These MD-AWVs are then evaluated for their applicability in environmental toxicology and pharmacology.

2. Materials and Methods

2.1 Theoretical Background. An AWV is a representation, in $\mathbb{R}^n$ space, of the weights of all atoms within a structure. Here, $\mathbb{R}$ represents the real number space and *n* represents the number of atoms within the molecular structure (MS) represented by the AWV.

$$X = [X_1, X_2, \ldots, X_N] \in \mathbb{R}^n \quad (1)$$

On the other hand, an aggregation operator (AO) is a function which gives a real number y to an n-dimensional vector of real numbers [X_1, X_2,…, X_N] [26, 27]:

$$y = AO([X_1, X_2, \ldots, X_N]) \quad (2)$$

The information codified in a weighted MS may be estimated in an $\mathbb{R}^n$ space, where $\mathbb{R}$ a set of real numbers and n is the number of atoms in a molecule structure. If the molecular vector W in the $\mathbb{R}^n$ space is considered to start from the origin, then the vector components represent given atom, atom-type or group properties[26]. Subsequently, the AOs may be applied to W to yield a set of total and local MDs. The AOs, as molecular structure characterizing parameters, have a significant property of being invariant, in that they yield the similar result nonetheless of the arrangement (numbering) of the vector components [26, 27].

Taking as an example the following MS:

Figure 1 comes about here (caption at the end of document)

This MS could be represented in the $\mathbb{R}^8$ space by an eight dimensional vector, with components [X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8]. Consequently, the property for each atom is divided by vertex degree [i.e. $Wi = P(X_i)/\delta(X_i)$], obtaining the following vector of atomic weights [26]:

W=[E(C1)/δ(C1) , E(C2)/δ(C2), E(N3)/δ(N3), E(C4)/δ(C4), E(C5)/δ(C5), E(C6)/δ(C6), E(Br7)/δ(Br6), E(S8)/δ(CS8)]

2.2 Atomic Weight of the Molecular Structure. In addition to molecular weight, an AWV can represent other physico-chemical properties of a MS. In such a way, MD can be derived from the AWV to represent various chemical or biological properties of the molecule [26, 27]. To distinguish each atom with these AWV we can use: 1) **chemical properties**, such as Atomic Number (Z), Atomic Mass (A), Volume Van der Waals area (VW), Covalent radius (R), Polarizability (P), Pauling Electronegativity (E); 2)

physical properties, such as Topological Surface Area (T), Ghose-Crippen A LogP (L), Ghose-Crippen Molar Refractivity (M), Charge(C); or 3) **vertex degree properties** [26]**,** such as Valence Degree (**N**), Intrinsic State (**I**) is a modification of the Valence Degree (**N**), also proposed by Kier and Hall [28], Electrotopological State (**ES**) or also called the E-state [29], Kupchik's Vertex Degree (**KU**) [30], Bond's Vertex Degree (**BD**): Accounts for connectedness as well as the bond multiplicity, Hu-Xu's Vertex Degree (HX) is defined by Hu-Xu [31], Li's Vertex Degree (LI) was defined by Li *et. al* [32], Alikhanidi's Vertex Degree (Alk) proposed by Alikhanidi *et. al.* [33], as a modification of the Hu–Xu vertex degree, Ivanciuc's Vertex Degree (IN) proposed by Ivanciuc [34] as combinations of topological distances, Distance Count (DC) and Eccentric Connectivity (Y).

2.3 Aggregation Operator or Invariants Applied to Atomic Weight's Vector. An Aggregation Operator (AO) takes the form of a mathematical function, F, applied to the elements x_i (i=1: n) of the AWV, w, in order to obtain a scalar quantity from the vector. The AO are defined by following form:

$$F(x_1, x_2, \ldots, x_n) = AO(x_1, x_2, \ldots, x_n) \quad (3)$$

AO represent a generalization of the traditional method of obtaining global (or local) MDs strictly by summation. These AOs are classified into four groups such as, 1) **Norms (or Metrics)**: Minkowski's norms (N1, N2, N3), and Penrose's size (PN); 2) **Mean Invariants (first statistical moment):** Geometric Mean (G), Arithmetic Mean (M), Quadratic Mean (P2), Potential Mean (P3) and Harmonic Mean (A); 3) **Statistical Invariants (highest statistical moments):** Variance (V), Skewness (S), Kurtosis (K), Standard Deviation (DE), Variation Coefficient (CV), Range (R), Percentile 25 (Q1), Percentile 50 (Q2), Percentile 75 (Q3), Inter-quartile Range (I50), X max (MX) and X min (MN); 4) **"Classical algorithm" Invariants:** Autocorrelations AC(i), Gravitational

(GI(i)), Total sum at *k* lags (TSk(i)), Ivanciuc-Balaban operator (IB(i)) and Electrotopological state (ES(i)). For the latter class of invariants, *i* represents a tunable parameter (e.g. the offset of the autocorrelation). These AOs have been applied to various indices by different researchers [26, 35-38]. The following table shows the formula of the AOs:

2.4 Local and Total Molecular Descriptors obtained from Atomic Weights Vector. Molecular Descriptor from Atomic Weighted Vector (MD-AWV) can be obtained using all elements of the AWV or using only a subset of the vector corresponding to particular atom-type formalism. For example, by applying different atomic groupings to the AWV, MD may be computed from the subset of atoms which are heteroatoms (HT), halogens (HL), carbon atoms (Cb), hydrogen bond acceptors (HA), hydrogen bond donors (HD), carbon atoms in aliphatic chains (LA), carbon atoms in aromatic portion (RA), terminal methyl groups (MC), unsaturation bonds(IS), and group at lag k (GL) [26].

2.5 Methodology for obtaining the MD-AWV. In general, MD-AWV can be defined such as:

$$Y = F_L(w), \quad (4)$$

Where, F_L is the AO, *w* represents the atomic weight vector, and *L* is the particular selection criterion (e.g. atom type) applied to select the subset of the AWV used in the calculation of the final MD-AWV. This is summarized in the following algorithm:

Step 1: Obtain the AWV (w) from the MS.

Step 2: Calculate the local atomic weight vector (w^L) from w: where w^L is a subset of w, which are obtained to apply atom-type formalism.

Step 3: Calculate the **MD-AWV** by applying the **AO** to w^L: MD:= AO(w^L).

Note: The label of a **MD** is set to [**Aggregation Operator**] $^{[\textbf{Property}]}_{[\textbf{atom-type formalism}].}$, for example, $N2^{VW}HT$ corresponds to a MD computed from N2-norm of the Volume Van der Waals area values for all heteroatoms within the structure.

This algorithm is implemented in a software package called MD-LOVIs (**M**olecular **D**escriptors from **LO**cal **V**ertex **I**nvariant**s** and Related Maps) [39], which is available in http://www.tomocomd.com.

2.6 Chemical Data Set. A chemical dataset of 392 benzene derivatives tested in *T. pyriformis* was used to create and assess models of aquatic toxicity. The data comprise diverse structural substituted benzene molecules containing nitrobenzenes, aminobenzenes, phenols, and benzenitriles [40]. These diverse molecules represent several mechanisms of toxic action. The toxicology was quantified using the concentration of chemical required to reduce growth of the *T. pyriformis* pathogen by 50% after 40 hours ($pIGC_{50}$) [10, 40]. For predictive model development, we make use of the same training set and test set previously used by Castillo-Garit *et. al.* [5]. The training set is composed of 313 benzene derivatives and the validation set (to assess the predictive capability of the QSAR models) contains an additional 79 compounds.

2.7 Preprocessing of the MDs. The IMMAN software package [41] was used to preprocess the calculated MDs, including extraction, feature selection, cleaning, and deletion of MDs from the final predictive model. Prior to building the QSAR models, feature selection of MDs was carried out with the aim of eliminating irrelevant MDs. This selection was performed with IMMAN software using Shannon Entropy (SE) as measure of ranking.

3. Results and Discussion

3.1 Variability Analysis.

The evaluation of the discriminative power of MD generated using different AO was conducted using the IMMAN software package. As above, SE quantifies MD relevance when examining the MD values computed for a set of different MS. An elevated SE indicates that the MD tends to vary with variations in the molecular structure, while a small SE indicates a non-informative MD.

3.1.1 Comparison between AO Groups. The principal aim of this study is to evaluate the contribution of the MDs from different AO groups in terms of an increase in the variability. A total of 120 MDs were computed for each AO group in this study using the 392 benzene derivatives. When using a discretization scheme of 392 bins, the maximum achievable entropy (Hartley's entropy) is $\log_2 392 = 8.644$ bits.

Figure 2 plots the number of MDs generated by each class of AO that have an SE greater than the value indicated on the x-axis. It can be observed that Autocorrelation (AC), Gravitational (GI), and Total Sum (TS) present the best entropy distribution of the variables, with some variables presenting entropy near 7.5 bits. These MDs with superior variability should yield good correlations with chemical and biological properties of chemical structures, as greater sensitivity to progressive structural changes is achieved. On the other hand, the *statistical*, *means* and *norms* AO produced the worst entropy distribution with a low maximal SE value near 6.5 bits. Moreover, Ivanciuc-Balaban (IB) and Electrotopological State (ES) had a moderate contribution. As can be observed, as the required entropy increases, it decreases the quantity of variables, visualizing its variability for this dataset. MD with higher SE values are more likely to be informative by encoding useful structural information and therefore to be useful in building models of aquatic toxicity.

Figure 2 comes about here (caption at the end of document)

3.2 Prediction of Aquatic Toxicity.

For the prediction of aquatic toxicity, 392 benzene derivatives with quantified toxicity to the ciliate *Tetrahymena pyriformis* were used. The MDs were computed by AO groups. Taking into account the variability analysis described above, the MD set was selected based on their Shannon Entropy (SE). In the prediction of aquatic toxicity of benzene derivatives, Multiple Linear Regression (MLR) was selected as technique to develop the models and Genetic Algorithms (GA) was the strategy used for selecting variables. This experiment was performed using the MOBYDIGs software (version 1.0) [42], which implements this method. The method was configured with a *population size* of 100 and a *reproduction/mutation trade-off* of 0.5. For each AO class analyzed, several MLR models (from 3-6 variables) for the corresponding aquatic toxicity were developed, using as fitness function the statistical parameter Q^2_{loo} ("leave-one-out" cross validation). The *"bootstrapping"* (Q^2_{boot}) statistical technique was used to evaluate the predictive power and *"Y-scrambling"* $(a\ (Q^2))$ was used to evaluate possible random correlation with respect to the modeled aquatic toxicity [43, 44]. Therefore, taking into consideration the previous parameters (including Q^2_{loo}), a multi-criteria evaluation was adopted to select the *"best"* model in each case. Next, the best model in each case was evaluated for its *generalization ability* using the *"external validation"* (Q^2_{ext}) procedure.

3.2.1 Comparison between AO groups. In this study, the AO groups were compared in the task of predicting aquatic toxicity. The results obtained by using different AO groups are shown in Table 2. In that table and hereafter, *N* is the size of the data set, R^2 is the determination coefficient, *s* is the standard deviation of the regression, *F* is the Fisher ratio, *and SDEC* is the square correlation coefficient (standard deviation) of the cross-validation performed with the LOO procedure. As can be observed, the *GI* AO

showed the best performance with the statistical parameters of Q^2_{loo} = 0.82 and R^2 = 0.83.

Next, AO are used for obtaining the model of aquatic toxicity, in order to evaluate the applicability of MD-AWV. The models obtained by using AO groups for two to six variable models are shown in Table 3 and as can be seen, the best obtained model (Eq. **5**) use six MD-AWV to explain 84% of the experimental variance of the aquatic toxicity with an adequate standard deviation value of 0.31. Likewise, the obtained models for five, four, three and two MD-AWV also had good behavior, explaining 82%, 81%, 78% and 71% of the experimental variance of the aquatic toxicity, respectively. Moreover, these models showed good values for the *LOO-CV* square correlation coefficient Q^2_{loo}= (0.83, 0.81, 0.80, 0.77 and 0.70) for six, five, four, three and two variables, respectively. Since these values are significantly higher than 0.5, they can be considered as proof of the high-predictive ability of the model. However, external validation is the only way to establish the real predictive power of the models. Here, the models again showed good values, with Q^2_{ext} = (0.79, 0.74, 0.72, 0.71 and 0.61) for six, five, four, three and two variable models, respectively. These results suggest that these descriptors can form the basis for accurate predictive models of aquatic toxicity. We therefore believe that they will be useful in the other cheminformatics studies to predict properties and activities of interest.

In this sense, the variables are essential component for building any mathematical predicting models, where its relationships are modeled using linear predictor functions. The general model for predicting aquatic toxicity, given n observations, is:

$$Log(1/IGC_{50}) = B_0 + B_1x_{i1} + B_2x_{i2} + ... + B_px_{ip} + E_i \quad (\mathbf{10})$$

As can be observed, the models that shown the best performances (Eqs. **5-9**) were built with several AWVs; they mostly use invariants were: AC, TS and some Norms and

Statistics. Among the most frequent norms that are used we can see: Manhattan norm and Minkowsky norm. Moreover, were selected for building the models some weights such as Polarizability, ALogP, Ivanciuc's Vertex Degree, Eccentric Connectivity; and were used several locals atom-types such as carbon atoms in aromatic portion, heteroatoms, unsaturation bonds and carbon atoms. Table 4 give detailed information of the names of all variables used to develop the best models by using AWVs (two to six variable models) presented in Table 3.

Many QSAR models that estimate several toxicity endpoints (including aquatic toxicity) are available. For example, DEMETRA was developed for the toxicity prediction of pesticide; allowing prediction of pesticide toxicity in trout, daphnia, bee and quail to facilitate the Dossier preparation for pesticide registration. TOPKAT employs robust and cross-validated QSAR models for assessing various measures of toxicity (more than ten) and utilizing the patented Optimal Predictive Space validation method to assist in interpreting the results. ECOSAR is a computerized predictive system that estimates a chemical's acute (short-term) toxicity and chronic (long-term or delayed) toxicity to aquatic organisms, such as fish, daphnia, green algae, and Mysid. However, none of them include models to predict the aquatic toxicity in *T. pyriformis*. Our predictive models had been obtained by using MD-LOVIs software; taking into account several properties, such as, chemical and physical properties, and different local fragments (as we pointed out before), which is a different way to obtain prediction models from molecular descriptors. These models were built from MD-AWV, specifically to predict the toxicity in *T. pyriformis*, taking into consideration different characteristic and information of the molecular structure of previously assayed compounds.

3.2.2 Comparison between MD-AWV and others approaches from the literature. In this study, different MD-AWV models are compared with other approaches from the literature in the task of predicting aquatic toxicity in *T. pyriformis*. The results obtained by using MD-AWV groups are shown in the Table 5 were shown the models obtained with other approaches and the same dataset. The comparison was mainly based on the quality of the statistical parameters of the regression models, the number of compounds and the number of variables used to develop the model. First, we will develop the comparison with those approaches that also use the entire dataset of 313 compounds. Particularly, the results of the present approach (the two better models) showed good performance R^2 = 0.820 and 0.837, for five and six variables, correspondingly. Although COMSIA achieved higher R^2 value (0.844) they use seven attributes, more than we use in our models (built with five and six attributes); their model shows the lowest value of SDEC with 0.295 and our models achieve values of 0.321 and 0.305, respectively. Moreover, the COMISA model shows a value of Q^2_{loo}=0.514, value far lower than those observed for our method, Q^2_{loo}=0.813 (Eq. 6) and 0.830 (Eq. 5). So, we can say that their model has less stability to perturbation in opposite to our models that were robust, not only in the LOO experiment, but also in the *bootstrapping* experiment (Q^2_{boot} of 0.811 and 0.827, respectively) unfortunately the other approaches do not developed this experiment to validate their models. Other approaches like CoMFA, VolSurf and ERM-VolSurf (all developed also with seven variables) showed values of R^2 between 0.797 and 0.816, as well as values of Q^2_{loo} among 0.611 and 0.783, lower than those achieved by our models. The last five rows of Table 5 report on QSAR models developed using 2D DRAGON´s molecular descriptors and with six variables but their performances were limited. The model obtained with molecular walk count descriptors was not considered in the comparison because of the poor behavior showed.

The other DRAGON-based models achieved R^2 values between 0.516 and 0.716, Q^2_{loo} values between 0.478 and 0.682, the values of SDEC were between 0.423 and 0.545; see Table 5 for other details.

The two models developed with stochastic and non-stochastic linear indices, and with six variables, also shown values of R^2 and Q^2_{loo} below 0.735 as well as values of SDEC greater than 0.394, far of the result of our model with five variables (0.324). After remove several outliers, the performance of the stochastic (five outlier) and non-stochastic (six outliers) models shown an improvement with values of R^2 of 0.799 and 0.791, respectively; as well as values of Q^2_{loo} of 0.786 and 0.781, correspondingly. For these models also the test set was evaluated showing values of Q^2_{ext} of 0.797 and 0.762, respectively; these values are very similar to the obtained with our model of six variables (Q^2_{ext} =0.789) and better than the result achieved with our model developed with five variables (Q^2_{ext} =0.742). Notice that, the best value obtained for the test set with the stochastic model (Q^2_{ext} =0.797) was obtained after remove an outlier compound from the test set; while our result for Q^2_{ext} (0.789) was obtained using the entire test set of 79 compounds. Another model was developed with Quantum chemical descriptors, four outliers were detected and removed, the result achieved with 16 variables were R^2=0.820, SDEC =0.29 (similar to our models but with about three times more variables). For the test set, this models shows a Q^2_{ext} value of 0.81 which is the best, but with the same handicap of use 16 variables in opposite of our models developed with only five and six variables. These results suggest that the correlations exhibited by MD-AWV-based models for the prediction of aquatic toxicity can be considered to be statistically significant similar to better in comparison with others strategies reported in the literature, despite their simplicity.

Conclusions

In this article, a set of MD-AWV was tested using MLR models for the prediction of aquatic toxicity of benzene derivatives against *T. pyriformis*. Furthermore, the obtained MD-AWV-based models were demonstrated to be effective in terms of the R^2 and Q^2_{loo} values. The model developed with MD-AWV for six variables showed particularly good values of Q^2_{loo} = 0.830 and R^2 = 0.837. The results obtained with the MD-AWV models were compared 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 *square coefficient relation (R^2)* methods. Generally, performance was observed to be highly competitive with the state of the art. It can be suggested that the MD-AWV are suitable for codifying important structural information of the molecules and, thus, constitute an interesting alternative to building effective models for the prediction of aquatic toxicity of benzenes against *T. pyriformis.*

ASSOCIATED CONTENT

The value of experimental results in excel files are available via the Internet at http://

References

1. Huff, J., *Benzene-induced cancers: abridged history and occupational health impact.* Int. J. Occup. Environ. Health, 2007. **13**: p. 213–221.
2. Tranfo, G., *Benzene and its Derivatives: New Uses and Impacts on Environment and Human Health.* 2011, Rome: Nova Science Pub Incorporated.
3. Salahinejad, M. and J.B. Ghasemi, *3D-QSAR studies on the toxicity of substituted benzenes to Tetrahymena pyriformis: CoMFA, CoMSIA and VolSurf approaches.* Ecotoxicol. Environ. Safety, 2014. **105**: p. 128–134.
4. Bradbury, S.P., *Quantitative structure–activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research.* Toxicol. Lett., 1995. **79**: p. 229–237.
5. Castillo-Garit, J.A., *et al.*, *A novel approach to predict aquatic toxicity from molecular structure.* Chemosphere, 2008. **73**: p. 415-427.

6. Brito-Sánchez, Y., *et al., Comparative study to predict toxic modes of action of phenols from molecular structures.* SAR QSAR Environ Res, 2013. **24**(3): p. 235-251.
7. Kleandrova, V.V., *et al., Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions.* Environ. Sci. Tech., 2014. **48**(24): p. 14686-14694.
8. Castillo-Garit, J.A., *et al., Machine learning-based models to predict modes of toxic action of phenols to Tetrahymena pyriformis.* SAR QSAR Environ. Res., 2017. **28**(9): p. 735-747.
9. Schultz, T.W., *et al., Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective.* J. Mol. Struct. (Theochem.), 2003. **622**: p. 1–22.
10. Schultz, T.W., *TERATOX: Tetrahymena pyriformis population grow impairment endpoint- A surrogate for fish lethality.* Toxicol. Methods, 1997. **7**: p. 289-309.
11. Bradbury, S.P., *et al., Overview of data and conceptual approaches for derivation of quantitative structure-activity relationships for ecotoxicological effects of organic chemicals.* Environ. Toxicol. Chem., 2003. **22**(8): p. 1789-1798.
12. Bordbar, M., *et al., Chemometric Modeling to Predict Aquatic Toxicity of Benzene Derivatives Using Stepwise-Multi Linear Regression and Partial Least Square.* Asian J. Chem., 2013. **25**(1): p. 331-342.
13. Cronin, M.T., *et al., Quantitative structure-activity study of the toxicity of benzonitriles to the ciliate Tetrahymena pyriformis.* SAR QSAR Environ Res, 1995. **3**(1): p. 1-13.
14. Castillo-Garit, J.A., *et al., Prediction of Aquatic Toxicity of Benzene Derivatives to Tetrahymena pyriformis According to OECD Principles.* Curr. Pharm. Des. , 2016. **22**(33): p. 5085-5094.
15. Gleeson, M.P.e.a., *The challenges involved in modeling toxicity data insilico: a review.* Curr. Pharm. Des., 2012. **18**: p. 1266–1291.
16. Schultz, T.W. and V.A. Tucker, *Structure-toxicity relationships for the effects of N- and N,N'-alkyl thioureas to Tetrahymena pyriformis.* Bull Environ Contam Toxicol, 2003. **70**(6): p. 1251-8.
17. Casanola-Martin, G., *et al., Tyrosinase enzyme: 1. An overview on a pharmacological target.* Curr. Top. Med. Chem., 2014. **14**(12): p. 1494-1501.
18. Tropsha, A., *Recent trends in statistical QSAR modeling of environmental chemical toxicity.* Molecular, Clinical and Environmental Toxicology., ed. A.E. Luch. 2012, Basel: Springer.
19. Zhu, H., *From QSAR to QSIIR: searching for enhanced computational toxicology models.* Computational Toxicology., ed. B. Reisfeld, Mayeno, A.N. . Vol. Volume II. 2013, GmbH , Berlin Heidelberg: Springer-Verlag.
20. Castillo-Garit, J.A., *et al., Identification In Silico and In Vitro of Novel Trypanosomicidal Drug-Like Compounds.* Chem. Biol. Drug Des., 2012. **80**(1): p. 38-45.
21. Castillo-Garit, J.A., *et al., In silico Antibacterial Activity Modeling Based on the TOMOCOMD-CARDD Approach.* J. Braz. Chem. Soc., 2015. **26**: p. 1218-1226.
22. Cañizares-Carmenate, Y., *et al., An approach to identify new antihypertensive agents using Thermolysin as model: In silico study based on QSARINS and docking.* Arab. J. Chem.
23. Marrero-Ponce, Y., *et al., Bond-based linear indices of the non-stochastic and stochastic edge-adjacency matrix. 1. Theory and modeling of ChemPhysical properties of organic molecules.* Mol. Divers., 2010. **14**(4): p. 731-753.
24. Castillo-Garit, J.A., *et al., Applications of Bond-Based 3D-Chiral Quadratic Indices in QSAR Studies Related to Central Chirality Codification.* QSAR Comb. Sci., 2009. **28**: p. 1465-1477.
25. Castillo-Garit, J.A., *et al., Bond-based bilinear indices for computational discovery of novel trypanosomicidal drug-like compounds through virtual screening.* Eur. J. Med. Chem., 2015. **96**(0): p. 238-244.

26. Martínez-López, Y., *et al.*, *The Summation of Atomic Contributions is an Overly Simplified Characterization of the Holistic Molecular Behavior*. Lett. Drug Des. Discov., 2016. **13**: p. 12.
27. Martínez-López, Y., *et al.*, *State of the Art Review and Report of New Tool for Drug Discovery*. Curr. Top. Med. Chem., 2017. **17**: p. http://dx.doi.org/10.2174/1568026617666170821123856.
28. Kier, L.B. and L.H. Hall, *Molecular Structure Description. The Electrotopological State*. 1999, New York: Academic Press.
29. Kier, L.B. and L.H. Hall, *An electrotopological-state index for atoms in molecules.* Pharm Res, 1990. **7**(8): p. 801-7.
30. Kupchik, E.J., *Structure—Molar Refraction Relationships of Alkylgermanes Using Molecular Connectivity.* Quantitative Structure-Activity Relationships, 1988. **7**(2): p. 57-59.
31. Hu, Q.-N., *et al.*, *Structural Interpretation of the Topological Index. 2. The Molecular Connectivity Index, the Kappa Index, and the Atom-type E-State Index.* J. Chem. Inf. Comput. Sci., 2004. **44**: p. 1193-1201.
32. Li, X., A. Jalbout, and M. Solimannejad, *Definition and application of a novel valence molecular connectivity index.* J. Mol. Struc-THEOCHEM, 2003. **663**(1): p. 81-85.
33. Alikhanidi, S. and Y. Takahash, *New Molecular Fragmental Descriptors and Their Application to the Prediction of Fish Toxicity*. MATCH Commun. Math. Comput. Chem, 2006. **55**: p. 205-232.
34. Ivanciuc, O., *Design on topological indices. 1. Definition of a vertex topological index in the case of 4-trees.* Revue Roumaine de Chimie, 1989. **34**(6): p. 1361-1368.
35. Barigye, S.J., *et al.*, *Relations Frequency Hypermatrices in Mutual, Conditional and Joint Entropy-Based Information Indices.* J. Comput. Chem., 2013. **34**: p. 259-274.
36. Barigye, S.J., *et al.*, *Event-Based Criteria in GT-STAF Information Indices: Theory, Exploratory Diversity Analysis and QSPR Applications.* SAR & QSAR Environ. Res., 2013. **24**: p. 3-34.
37. Barigye, S.J., *et al.*, *Shannon's, Mutual, Conditional and Joint Entropy-Based Information Indices. Generalization of Global Indices Defined from Local Vertex Invariants* Curr. Comput.-Aided Drug Des., 2013. **9**.
38. Marrero-Ponce, Y., *et al.*, *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, 2013.
39. Martínez-López, Y., Y. Marrero-Ponce, and S. Barigye, *MD-LOVIs for windows: Software for molecular de scriptors calculator from LOVIs. Version 1.0*. 2010. p. CAMD-BIR Unit.
40. Schultz, T.W. and T.I. Netzeva, *Development and evaluation of QSARs for ecotoxic endpoints: the benzene response-surface model for Tetrahymena toxicity.*, in *Modelling Enviromental Fate and Toxicity*, M.T. Cronin and D. Livingstone, Editors. 2004, CRC Press: Boca Raton, FL. p. 265-284.
41. Barigye, S.J., *et al.*, *IMMAN (Information Theory based Chemometric Analysis)*. 2013, CAMD-BIR Unit: Santa Clara
42. Todeschini, R., *et al.*, *MOBYDIGS version 1.0*. 2005: Milano.
43. Golbraikh, A. and A. Tropsha, *Beware of q2!* J. Mol. Graphics Model., 2002. **20**(4): p. 269-76.
44. Golbraikh, A. and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection.* J. Comput-Aided Mol. Des., 2002. **16**(5): p. 357-369.
45. OECD, *Guidance document on the validation of quantitative structure-activity relationship [QSAR] models.*, in *OECD Environment Health and Safety Publication, Series on testing and assessment no. 69*. 2007: Paris.

Figures Captions

Figure 1 Representation of Molecular Structure of a molecule.

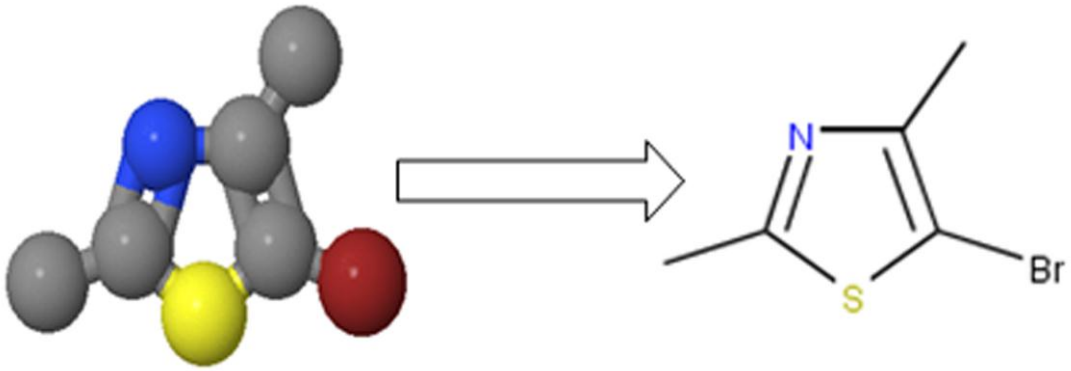

Figure 2. Shannon distributions of the AO groups.

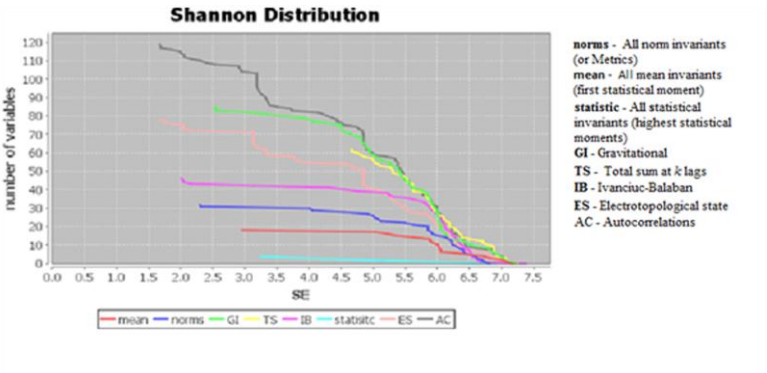

Table 1. Norms, Means and Statistical AOs and Classical algorithms that generalize the first three groups.

No.	Group[a]	Name	ID	Formula[b]

1	Norms (Metrics)	Minkowsky norm (p = 1) Manhattan norm	N1	$N1 = \sum_{a=1}^{n} \lvert L_a \rvert$
2		Minkowsky norm (p = 2) Euclidean norm	N2	$N2 = \sqrt{\sum_{a=1}^{n} \lvert L_a \rvert^2}$
3		Minkowsky norm (p = 3)	N3	$N3 = \sqrt[3]{\sum_{a=1}^{n} \lvert L_a \rvert^3}$
4		Penrosesize	PN	$PN = \sqrt{\frac{1}{n^2}\left[\sum_{a=1}^{n} (L_a)\right]^2}$
5	Mean (firststatisticalmoment)	Geometric Mean	GM	$G = \sqrt[n]{\prod_{a=1}^{n} L_a}$
6		Arithmetic Mean (Power mean of degree β = 1)	M	$M_\beta = \left(\frac{L_1^\beta + L_2^\beta + \ldots + L_n^\beta}{n}\right)^{\frac{1}{\beta}}$
7		Quadratic Mean (Power mean of degree β = 2)	P2	
8		Power mean of degree β = 3	P3	
9		Harmonic Mean (Power mean of degree β = -1)	A	
10	Statistical (higheststatistical moments):	Variance	V	$V = \frac{\sum_{a=1}^{n} (L_a - M)^2}{n - 1}$
11		Skewness	S	$S = \frac{n * (X_3)}{(n-1)(n-2)(DE)^3}$ $X_3 = \sum_{a=1}^{n} (L_a - M)^3$ M, arithmetic mean DE, standard deviation
12		Kurtosis	K	$k = \frac{n(n+1)X_4 - 3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(DE)^4}$ $X_j = \sum_{a=1}^{n} (L_a - M)^j$ M, arithmetic mean DE, standard deviation
13		Standard Deviation	DE	$DE = \sqrt{\frac{(\sum L_a - M)^2}{n-1}}$
14		VariationCoefficient	CV	$CV = DE / M$
15		Range	R	$R = L_{max} - L_{min}$
16		Percentile 25	Q1	$Q1 = \left\lfloor \frac{N}{4} + \frac{1}{2} \right\rfloor$ N, La number
17		Percentile 50	Q2	$Q2 = \left\lfloor \frac{N}{2} + \frac{1}{2} \right\rfloor$ N, La number
18		Percentile 75	Q3	$Q3 = \left\lfloor \frac{3N}{4} + \frac{1}{2} \right\rfloor$ N, La number
19		Inter-quartileRange	I50	$I50 = Q3 - Q1$
20		Maximumvalue	MX	MX = La max

21		Minimumvalue	MN	MN = La min
22	Classical	Autocorrelation	ACk	$AC_k = \sum_{i=1}^{n} \sum_{j\geq 1}^{n} L_i \times L_j \bullet (\delta(d_{ij}, k))$ $k = 1,2,..7$
23		Gravitational	GIk	$GI_k = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{L_i L_j}{{}_k d_{ij}} \bullet \delta(d_{ij}, k))$ $k = 1,2,..7$
24		Total sum at lag k	TSk	$TS_k = \sum_{i=1}^{n} \sum_{j=1}^{n} L_{ij} \bullet \delta(d_{ij}, k))$ $k = 1,2,\ldots,7$
25		Kier-Hall connectivity	CNm	${}^m KH_t = \sum_{i=1}^{K} (\prod_{i=1}^{n_k} L_i, w)_k^{\lambda}$ where, K is the number of sub-graphs, nk is the number of atoms in a fragment, λ is equal to ½, m and t are the sub-graph order and type, respectively
26		Mean Information Content	MI	$MI = -\sum_{i=1}^{A} \frac{N_g}{N_o} \cdot \log_2 \frac{N_g}{N_o}$ where, Ng is the number of atoms with the same LOVI value. No is the number of atoms in a molecule
27		Total Information Content	TI	$TI = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g$
28		StandardizedInformation Content	SI	$SI = \frac{IT}{N_0 \cdot \log_2 N_0}$
29		Electrotopological state (E-state index)	ES	$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^{n} \frac{I_i - I_j}{(d_{ij}+1)^2}$ where, Ii is the intrinsic state of the ith atom and ΔIi is the field effect on the ith atom calculated as perturbation of the Ii of ith atom by all other atoms in the molecule, dij is the topological distance between the ith and the jth atoms, and n is the number of atoms. The exponent k is 2.
30		Ivanciuc-Balaban Type-Indices	IB	$J_k = \frac{n^2 \cdot B}{n + C + 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \left[L_i \times L_j\right]^{-\frac{1}{2}}$ where, the summation goes over all pairs of atoms but only pairs of adjacent atoms are accounted for by means of the elements aij of the adjacency matrix. The n, B, and C are the number of atoms, bonds, and rings (cyclomatic number), respectively.

[a]The second group (invariants 5-9) could be re-named as "location statistics" if percentiles and maximum (minimum) are taken into consideration in this group. In this case, the third group (invariants 10-21) could be re-named as "spread and shape statistics",[b]LOVIs for "a" atoms in molecule.

Table 2. Results of statistical parameter of AO groups.

AO	N	No. MD	r^2	Q^2_{loo}	Q^2_{boot}	Q^2_{ext}	$a(Q^2)$	SDEC	F	s
AC	313	6	0.80	0.79	0.78	0.75	-0.06	0.34	206.86	0.34
ES	313	6	0.82	0.81	0.79	0.77	-0.07	0.32	226.99	0.33
GI	313	6	**0.83**	**0.82**	**0.81**	**0.77**	**-0.06**	**0.31**	**244.72**	**0.32**
IB	313	6	0.81	0.81	0.80	0.76	-0.06	0.33	221.36	0.33
TS	313	6	0.82	0.81	0.80	0.72	-0.06	0.32	234.02	0.32
means	313	6	0.79	0.77	0.77	0.66	-0.06	0.32	227.6	0.33
statistic	313	6	0.73	0.71	0.70	0.62	-0.06	0.35	187.14	0.35
norms	313	6	0.82	0.81	0.81	0.76	-0.05	0.40	134.49	0.40

Table 3. Best models obtained from MD-AWV (from two to six variables).

N	Size	Model	R^2	Q^2_{boot}	Q^2_{ext}	$a(Q^2)$	SDEC	F	s	Eq.
313	6	**Log(1/IGC$_{50}$)**= -1.917(±0.096) +0.179(±0.009)M^{Y}_{RA} +0.867(±0.109)R^{P}_{RA}+0.451(±0.019)$N1^{L}_{HT}$ -0.010(±7.4 x 10^4)$N3^{IN11}_{HT}$ +1x10^4(±1x10^4) $AC^{1}(PN)^{IN14}_{HT}$ +1.483 x10^2(±1.65 x10^3)$TS^{2}(N2)^{IN7}_{IS}$	0.84	0.83	0.79	-0.06	0.31	262.63	0.31	5
313	5	**Log(1/IGC$_{50}$)**= -1.536(±0.102) +0.087(±0.011)$P3^{Y}_{RA}$ +0.537(±0.02)$N1^{L}_{HT}$ +0.10(±0.019) $AC^{1}(N1)^{L}_{CB}$ -7 x10^4(±6 x10^5)$AC^{T}(N3)^{IN2}_{HA}$ +0.025(±1.99 x10^3)$TS^{2}(N2)^{IN7}_{IS}$	0.82	0.81	0.74	-0.05	0.32	279.49	0.32	6
313	4	**Log(1/IGC$_{50}$)**= -1.735(±0.098) +0.127(±8.59 x10^3)$P3^{Y}_{RA}$ +0.525(±0.02)$N1^{L}_{HT}$ -7.4 x10^4(±6 x10^5)AC^{T} $(N3)^{IN2}_{HA}$ + 0.023(±2.04 x10^3)$TS^{2}(N2)^{IN7}_{IS}$	0.81	0.80	0.72	-0.05	0.33	317.68	0.34	7
313	3	**Log(1/IGC$_{50}$)**= -0.899(±0.067) +0.533(±0.021)$N1^{L}_{HT}$ +0.05(±2.81 x10^3)PN^{IN15}_{CB} -8.8 x10^4(±6x10^5)$AC^{T}(N3)^{IN2}_{HA}$	0.78	0.77	0.71	-0.04	0.36	358.38	0.36	8
313	2	**Log(1/IGC$_{50}$)**= -1.062(±0.096) +0.109(±7.3 x10^3)$Q3^{Y}_{IS}$ +0.49(±0.022)$N1^{L}_{HT}$	0.71	0.70	0.61	-0.04	0.41	380.29	0.41	9

Table 4. Name of the variables used in the best models

Variable	Name
IGC_{50}	impairment growth concentration for 50% of the *T. pyriformis* population
M^{Y}_{RA}	Arithmetic Mean- Eccentric Connectivity- carbon atoms in aromatic portion
R^{P}_{RA}	Range- Polarizability- carbon atoms in aromatic portion

$N1^{L}{}_{HT}$	Manhattan norm- A LogP- heteroatoms
$N3^{IN11}{}_{HT}$	Minkowsky norm{3}- Ivanciuc's Vertex Degree{11}- heteroatoms
$AC^{1}(PN)^{IN14}{}_{HT}$	Autocorrelation[1](Penrose)- Ivanciuc's Vertex Degree{14}- heteroatoms
$TS^{2}(N2)^{IN7}{}_{IS}$	Total sum[45](Euclidean norm)- Ivanciuc's Vertex Degree[45]- unsaturation bonds
$P3^{Y}{}_{RA}$	Power mean{3}- Eccentric Connectivity- carbon atoms in aromatic portion
$AC^{1}(N1)^{L}{}_{CB}$	Autocorrelation{1}(Manhattan norm)- A LogP- carbon atoms
$AC^{T}(N3)^{IN2}{}_{HA}$	Autocorrelation{Total}(Minkowsky norm{3})- Ivanciuc's Vertex Degree[45]- hydrogen bond acceptors
$PN^{IN15}{}_{CB}$	Penrose- Ivanciuc's Vertex Degree{12}- carbon atoms
$Q3^{Y}{}_{IS}$	Percentile 75- Eccentric Connectivity- unsaturation bonds

Table 5. Comparison between MD-AWV and others approaches from the literature

Index	**N**	**n**	**R^2**	**Q^2_{loo}**	**Q^2_{boot}**	**Q^2_{ext}**	**SDEC**	**F**	**S**	**Statistical Method**	**Eq/Ref.**
MD-AWV	313	6	0.837	0.830	0.827	0.789	0.305	262.63	0.309	MLR-GA	5
MD-AWV	313	5	0.820	0.813	0.811	0.742	0.321	279.49	0.324	MLR-GA	6
CoMSIA	313	7	0.844	0.514	-	-	0.295	-	-	PLS	[3]
CoMFA	313	7	0.816	0.611	-	-	0.319	-	-	PLS	[3]
VolSurf	313	7	0.806	0.758	-	-	0.322	-	-	PLS	[3]
ERM-VolSurf	313	7	0.797	0.783	-	-	0.330	-	-	MLR	[3]
Quantum chemical descriptors	309	16	0.820		-	0.81	-	-	0.29	Stepwise MLR-PLS	[12]
Stochastic linear indices	308	6	0.799	0.786	-	0.797	0.350	198.88	0.343	MLR	[5]
Non-Stochastic linear indices	307	6	0.791	0.781	-	0.762	0.348	189.43	0.344	MLR	[5]
Stochastic linear indices	313	6	0.733	0.704	-	-	0.411	139.94	0.394	MLR	[5]
Non-Stochastic linear indices	313	6	0.721	0.687	-	-	0.421	131.79	0.403	MLR	[5]
2D autocorrelations	313	6	0.609	0.585	-	-	0.486	79.54	0.476	MLR	[5]
BCUT	313	6	0.690	0.675	-	-	0.431	113.56	0.424	MLR	[5]
Gálvez topological charge indices	313	6	0.516	0.478	-	-	0.545	54.30	0.530	MLR	[5]
Topological descriptors	313	6	0.716	0.682	-	-	0.423	128.70	0.406	MLR	[5]
Molecular walk count	313	6	0.346	0.303	-	-	0.630	40.80	0.614	MLR	[5]

N: number of compounds; **n**: number of variables used to develop the model