13rd International Electronic Conference on Synthetic Organic Chemistry (ECSOC-13), 1-30 November 2009.
http://www.mdpi.org/ecsoc-13/ & http://www.usc.es/congresos/ecsoc/13/

[g003]

Atom-Based Quadratic Indices to Predict Aquatic Toxicity of Benzene Derivatives to *Tetrahymena pyriformis*

Juan A. Castillo-Garit,[a,b,c,*] Jeanette Escobar,[a] Yovani Marrero-Ponce,[b,c,] and Francisco Torrens,[c]

[a]*Applied Chemistry Research Center, Faculty of Chemistry-Pharmacy and Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.* ***e-mail:*** *jacgarit@yahoo.es, juancg.22@gmail.com or juancg@uclv.edu.cu*

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

[c]*Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, 46071 Valencia, Spain*

Abstract

The non-stochastic and stochastic atom-based quadratic indices are applied to develop quantitative structure-activity relationship (QSAR) models for the prediction of aquatic toxicity. The used dataset, consisting of 392 benzene derivatives for which toxicity data to the ciliate *Tetrahymena pyriformis* were available, is divided into training and test sets. The obtained multiple linear regression models are statistically significant (R^2 = 0.787 and s = 0.347, R^2 = 0.806 and s = 0.329, for non-stochastic and stochastic quadratic indices, respectively) and show rather good stability in a cross-validation experiment (q^2 = 0.769 and s_{cv} = 0.357, q^2 = 0.791 and s_{cv} = 0.337, correspondingly). In addition, a validation through an external test set is performed, which yields significant values of R^2_{pred} of 0.745 and 0.742. The comparison with other approaches exposes a good behavior of our method of predicting the aquatic toxicity of benzenes. The obtained results suggest that, the non-stochastic and stochastic quadratic indices seem to provide an interesting alternative to costly and time-consuming experiments for determining toxicity.

Keywords: Atom-based non-stochastic and stochastic linear index, Multiple linear regression, QSAR, *Tetrahymena pyriformis*, Program ***TOMOCOMD-CARDD***.

1. Introduction

Benzene is a parent compound for a wide variety of derivatives, many of which are among the most prevalent industrial organic chemicals in the world, as defined by the High Production Volume Chemicals list [1]. It´s chemical caracteristics (bond angles of 120°, sp^2-hybrid orbitals, as well as π-bonds derived from p-atomic orbitals and equally extending around the ring) impart the aromatic nature of the substance. With this delocalization, benzene does not exhibit the high reactivity typical of polyene compounds. However, this fact changes dramatically when benzene is substituted with unsaturated (e.g., π-bond-containing) funcionalities, especially in conjunction with leaving groups [2, 3]. Therefore, toxicity data on benzene derivatives are important for their use in risk assessment processes [4].

While experimental testing provides the most reliable data about the effects of chemicals, it is not suitable to screen a large number of potential toxicants [5], because the generation of toxicological data is often a lengthy and costly process and, thus, predictive models in the form of quantitative structure-activity relationships (QSARs) are a necessary tool to fill data gaps in environmental risk assessment and regulatory concerns [6]. This kind of studies offers the advantages of higher speed and lower cost, especially when compared to experimental testing [5].

The QSARs are powerful tools in predictive toxicology and are employed, as scientifically credible tools, to predict the acute toxicity of chemicals when few empirical data are available. The Office of Toxic Substances of the U.S. Enviromental Protection Agency has developed QSARs based on as little as one datum and assumptions about the nature of the relationship between a chemical class and its toxicity [7]. Consistent with the development and application of QSARs to the design of more efficacious pharmaceuticals and pesticides, it has been the increasing acceptance of structure-activity relationships for predicting the adverse effects of xenobiotics in risk assessment [8].

In the development of an ecotoxicity-based QSAR, the connection of subjects (biology, chemistry, and statistics) has permitted the development of structure–activity relationships as an accepted sub-discipline in toxicology [9]. There are three elements in this subdiscipline: the toxicological data, the descriptor data, and the statistical method of linking the two data sets [10]. In additon, some issues have been recognized as topics of particular interest [11]: they are quality, transparency, domain identification, and validation. A quality QSAR only can be constructed and validated with quality data, but quality in QSARs is morethan a high

coefficient of determination. Transparency means that the data that are used in the development and validation of the models are available for examination and can also mean the amount of process information obtainable from the statistical method; it goes from the black boxes of genetic algorithms to interpretable multiple linear regression [12]. Since the use of a particular QSAR is only valid within its domain, the identification of that domain is critical to QSAR acceptability [11].

In particular, the database of inhibition of growth database of ciliated protozoan *Tetrahymena pyriformis* [13] is considered to be a high-quality data set [14]. It has been developed in a single laboratory over more than two decades. Moreover, these data have been compiled for the main purpose of QSAR development and validation. In recent years, many works have been reported using *T. pyriformis* to develop linear models [5, 15-23]; additionaly, some non-linear methods were also applied [24-26] to predict aquatic toxicity in *T. pyriformis*.

On the other hand, a novel scheme to the rational *–in silico-* molecular design and to QSAR/QSPR has been introduced by our research group: ***TOMOCOMD*** **(**acronym of ***TO***pological ***MO***lecular ***COM***puter ***D***esign). It calculates several new families of 2D (D=dimention), 3D-Chiral (2.5) and 3D (geometric and topographical) non-stochastic and stochastic atom- and bond-based molecular descriptors, based on algebraic theory and discrete mathematics. These descriptors are denoted quadratic, linear and bilinear indices, and have been defined by analogy with the quadratic, linear and bilinear mathematical maps [27-32]. These approaches describe changes in electron distribution with time throughout molecular backbone, and they have been successfully employed in the prediction of several physical, physicochemical, chemical, biological, pharmacokinetic and toxicological properties of organic compounds [33-42], including studies related to proteomics [43, 44] and nucleic acid-drug interactions [45, 46]. Besides, these indices have been extended to consider the three-dimensional features of small/medium-sized molecules based on the trigonometric 3D-chirality correction factor approach [47-51].

The present report is written with the objective of testing the applicability of the atom-based quadratic indices in ecotoxicological research. Therefore, we shall develop QSAR models for the prediction of aquatic toxicity for a large group of substituted benzenes, tested on the impairment assay of the population growth of *T. pyriformis*.

2. Materials and Methods

2.1. TOMOCOMD-CARRD approach.

For the computation of the atom-based quadratic indices we used software ***TOMOCOMD*** [52]. It is an interactive program for molecular design and bioinformatic research, which contains

four routines: ***CARDD***(*C*omputed-*A*ided *R*ational *D*rug *D*esign), ***CAMPS*** (*C*omputed-*A*ided *M*odeling in *P*rotein *S*cience), ***CANAR*** (*C*omputed-*A*ided *N*ucleic *A*cid *R*esearch) and ***CABPD*** (*C*omputed-*A*ided *B*io-*P*olymers *D*ocking); every one of them allows both drawing the structures (drawing mode) and calculating molecular 2D/3D descriptors (calculation mode). In the present report, we outline salient features concerned with only one of these routines, ***CARDD***, and with the calculation of atom-based non-stochastic and stochastic quadratic indices, considering and not considering H-atoms in the molecular pseudograph (G).

The main steps for the application of this method in quantitative structure-activity/toxicity relationships (QSAR/QSTR) and for drug design were the same as the ones that we used in an earlier publication for the non-stochastic and stochastic atom-based linear indices [42].

The descriptors computed in this work were the following:

1) $\boldsymbol{q_k}(x)$ and $\boldsymbol{q_k}^{H}(x)$ are the k^{th} atom-based non-stochastic total quadratic indices, not considering and considering H-atoms, respectively, in the molecule

2) $\boldsymbol{q_k}_{L}(x_E)$ and $\boldsymbol{q_k}_{L}^{H}(x_E)$ are the k^{th} atom-based non-stochastic local (atom-type = heteroatoms: S, N, O) quadratic indices, not considering and considering H-atoms, respectively, in the molecule.

3) $\boldsymbol{q_k}_{L}^{H}(x_{E\text{-}H})$ are the k^{th} atom-based non-stochastic local (atom-type = H-atoms bonding to heteroatoms: S, N, O) quadratic indices, considering H-atoms in the molecular pseudograph (G).

Therefore, the k^{th} atom-based stochastic total [${}^{s}\boldsymbol{q_k}(x)$ and ${}^{s}\boldsymbol{q_k}^{H}(x)$], as well as local [${}^{s}\boldsymbol{q_k}(x_E)$, ${}^{s}\boldsymbol{q_k}^{H}(x_E)$ and ${}^{s}\boldsymbol{q_k}^{H}(x_{E\text{-}H})$] quadratic indices were also computed.

2.2 Chemical database selection.

Biological data is central to the issues of quality, transparency, and domain identification as they relate to toxicological QSAR. High-quality toxicity data, in a structurally diverse set of molecules, are required to formulate and validate high-quality QSARs. Quality toxicity data typically come from standardized assays, measured in a consistent manner, with a clear and unambiguous endpoint, and lower experimental error [12]. Toxicity assessments that are made in a single laboratory by a single protocol tend to be the most precise ones. Taking into consideration these points, we select the database of the inhibition of growth of the ciliated protozoan *T. pyriformis*. This database has been developed in a single laboratory over more than two decades, and it has been recognized as a high-quality data set [14]. While numerous workers, using slight variations in the static protocol and nominal concentrations, have generated the data, the data set still remains an excellent primary source of information: it is also unique in terms of size, molecular diversity, and quality.

The general data set used in this study has been recently published by other researchers [12]. It consists of almost 400 substituted benzenes, representing several mechanisms of toxic action. Some compounds were reported by Schultz and Netzeva as non-toxic at saturation; hence these compounds were not used in the present work. A horizontal validation was performed using a training set, composed of 313 benzene derivatives, for model development and a validation set (79 compounds) to assess the predictive capability of the QSAR models. In order to split the database into training and prediction series, a *k*-means cluster analyses (*k*-MCA) was carried out for the entire data set to design, in a rational representative way, the training (learning) and prediction (test) series [53, 54].

2.3. Chemometric Methods.

2.3.1. Cluster Analysis. The cluster analysis (CA) is the name of a group of methods used to recognize similarities among cases (objects) or among variables and to single out some categories as a set of similar cases (or variables) [55]. This CA comprehends a number of different 'classification algorithms' and allows organizing the data into subsystems. These algorithms are grouped into two categories: hierarchical clustering and partitional (non-hierarchical) clustering. Hierarchical clustering rearranges objects in a tree-structure (joining clustering), in an agglomerative (bottom-up) procedure. On the other hand, partitional clustering assumes that the objects have non-hierarchical characters [53-56].

The most used cluster algorithms are the *k*-means cluster analysis (*k*-MCA) and Jarvis-Patrick algorithm (also known as k-nearest neighbor cluster analysis, *k*-NNCA); in our case, in order to design the training and test series to guarantee structural and toxicity variabilities in both series of the present database, we carried out both kinds of cluster analyses (*k*-MCA and *k*-NNCA) for the entire dataset of compounds [53-56]. The number of members in every cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taken into account to have an acceptable statistical quality of data partition into clusters. The values of the standard deviation (SD) between and within clusters, those of the respective Fisher ratio and their *p*-level of significance were also examined [53-56]. Finally, before carrying out the cluster processes, all the variables were standardized. In the standardization, all values of selected variables (molecular descriptors) were replaced by standardized values, which were computed as follows: Std. score = (raw score - mean)/Std. deviation.

2.3.2. Multiple Linear Regression. In the prediction of aquatic toxicity against *T. pyriformis*, the multiple linear regression (MLR) analysis was used as statistical method. This experiment was performed with software package STATISTICA [56]. The considered tolerance parameter (proportion of variance that is unique to the respective variable) was the default value for

minimum acceptable tolerance, which was 0.01. Forward stepwise procedure was fixed as the strategy for variable selection. The principle of maximal parsimony (Occam's razor) was taken into account as the strategy for model selection. Therefore, we selected the model with the highest statistical signification, but having as few parameters (a_k) as possible. The **log $(IGC_{50})^{-1}$** (decimal logarithm of the inverse 50 percent growth inhibitory concentration) values, concentration reported as mmol/L, were used as the dependent variable.

The quality of the models was determined by examining the regression's statistical parameters and those of the cross-validation procedures [57, 58]. Therefore, the following parameters were verified: the correlation coefficient (R), determination coefficient or square correlation coefficient (R^2), Fisher-ratio's *p*-level [*p*(F)], standard deviation of the regression (*s*) and the leave-one-out (LOO) press statistics (q^2, s_{cv}). The predictive powers of the obtained models were assessed by using an external prediction (test) set.

3. Results and Discussion

3.1. Similarity Analysis and the Design of Training and Test Sets.

As we mentioned above, the quality of any QSAR model depends on the quality of the selected data set, but one of the most critical aspects is to warrant enough molecular diversity for the training set. We performed a hierarchical CA of the entire dataset to demonstrate the structural diversity of this data set, [53, 54]. The dendrogram (binary tree) is given in Figure 1; using the Euclidean distance (X-axis) and the complete linkage (Y-axis), it illustrates the results of the *k*-NNCA developed for the dataset. As it can be observed in the binary tree there is a number of different subsets, which proves the molecular variability of the selected chemicals in these database.

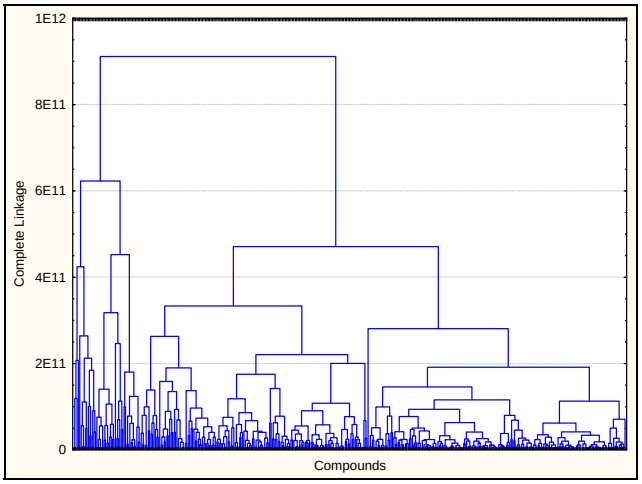

Figure 1. A dendrogram illustrating the results for the hierarchical *k*-NNCA developed for the dataset.

Due to the difficulty in evaluating the output dendrogram, other kind of CAs is usually performed. Therefore, we perform a *k*-MCA with the objective of spliting the whole group into two data sets (training and predicting ones). The main idea of this procedure consists in making a partition of the chemicals into several statistically representative classes of compounds. This procedure ensures that any chemical class (as determined by the clusters derived from *k*-MCA) will be represented in both compounds' series. This "rational" design of the training and predicting series allowed us to devise both sets that are representative of the whole "experimental universe". This procedure splits the dataset of benzene derivatives into 9 clusters.

Finally, we select the training and prediction sets by taking, in a random way, compounds belonging to every cluster. From these 392 benzene derivatives, 313 compounds were chosen as the training set. The remaining subset, composed of 79 compounds, was used as the test set for the external validation of the models. These compounds were never used in the development of the QSAR models. This procedure is illustrated graphically in Figure 2. The CA was performed to select a representative sample of the training and test sets.

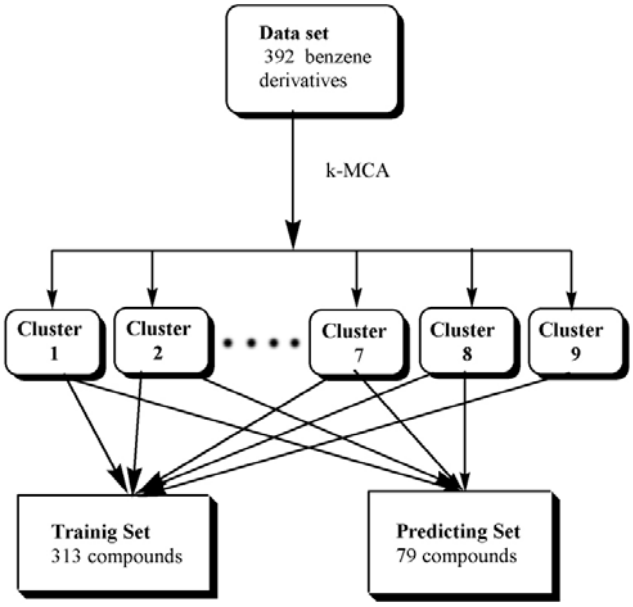

Figure 2. General algorithm used for designing training and test sets throughout *k*-MCA

3.2. Development of the models of prediction of aquatic toxicity.

In order to evaluate the applicability of the non-stochastic and stochastic atom-based quadratic indices for predicting aquatic toxicity, the whole data set was divided into training and test sets, as we described above. The MLR analysis was used to develop QSAR models for the prediction of aquatic toxicity against *T. pyriformis*. The toxicity values to *T. pyirformis* for the benzene derivatives of the training set are presented in Table **1**.

The model obtained by using atom-based non-stochastic linear indices is the following:

$$\textbf{Log (1/IGC}_{\textbf{50}}\textbf{)} = -0.899(\pm 0.106) + 7.06\text{x}10^{-2}(\pm 0.58\text{x}10^{-2})^{P}\boldsymbol{q}_{1L}(x_E) + 2.87\text{x}10^{-2}(\pm 0.20\text{x}10^{-2})^{K}\boldsymbol{q}_{0}(x) - 1.68\text{x}10^{-2}(\pm 0.18\text{x}10^{-2})^{G}\boldsymbol{q}_{2L}{}^{H}(x_E) + 1.29\text{x}10^{-2}(\pm 0.17\text{x}10^{-2})^{G}\boldsymbol{q}_{2L}(x_E) - 1.97\text{x}10^{-6}(\pm 0.29\text{x}10^{-6})^{G}\boldsymbol{q}_{7}{}^{H}(x) - 1.85\text{x}10^{-3}(\pm 0.33\text{x}10^{-3})^{V}\boldsymbol{q}_{2L}{}^{H}(x_{E\text{-}H}) + 2.88\text{x}10^{-10}(\pm 0.55\text{x}10^{-10})^{V}\boldsymbol{q}_{15L}{}^{H}(x_{E\text{-}H}) \quad \textbf{(1)}$$

N = 313 $R^2 = 0.730$ $s = 0.396$ F = 118.04 $p < 0.0001$

$q^2 = 0.697$ $s_{cv} = 0.415$

where N is the size of the data set, R is the correlation coefficient, R^2 is the determination coefficient, s is the standard deviation of the regression, F is the Fischer ratio, q^2 (s_{cv}) is the square correlation coefficient (standard deviation) of the cross-validation performed with the LOO procedure.

As can be seen, the obtained model (Eq. **1**) explains 73% of the experimental variance of the aquatic toxicity with adequate value of 0.396 of standard deviation. However, eight compounds were detected as statistical outliers (006, 020,074, 156, 215, 335, 354 and 360) and showed large values of standard residual. These compounds and their residual values are reported in Table **2**. Once rejected these outlier compounds, a new non-stochastic model (Eq. **2**) was obtained with better statistical parameters:

$$\textbf{Log (1/IGC}_{\textbf{50}}\textbf{)} = -1.302(\pm 0.116) + 7.55\text{x}10^{-2}(\pm 0.53\text{x}10^{-2})^{P}\boldsymbol{q}_{1L}(x_E) + 3.12\text{x}10^{-2}(\pm 0.18\text{x}10^{-2})^{K}\boldsymbol{q}_{0}(x) - 1.77\text{x}10^{-2}(\pm 0.16\text{x}10^{-2})^{G}\boldsymbol{q}_{2L}{}^{H}(x_E) + 1.33\text{x}10^{-2}(\pm 0.16\text{x}10^{-2})^{G}\boldsymbol{q}_{2L}(x_E) - 1.47\text{x}10^{-6}(\pm 0.32\text{x}10^{-6})^{G}\boldsymbol{q}_{7}{}^{H}(x) - 1.48\text{x}10^{-3}(\pm 0.30\text{x}10^{-3})^{V}\boldsymbol{q}_{2L}{}^{H}(x_{E\text{-}H}) + 2.41\text{x}10^{-10}(\pm 0.50\text{x}10^{-10})^{V}\boldsymbol{q}_{15L}{}^{H}(x_{E\text{-}H}) \quad \textbf{(2)}$$

N = 305 $R^2 = 0.787$ $s = 0.347$ F = 156.61 $p < 0.0001$

$q^2 = 0.769$ $s_{cv} = 0.357$ $R^2_{pred} = 0.745$

where R^2_{pred} is the square correlation coefficient for the external prediction set.

This new model explains almost the 79 % of the experimental variance, and a small value of standard deviation of 0.347; the other statistical parameters were also improved. In order to assess the predictability and stability of the obtained models using non-stochastic linear indices (Eqs. **1** and **2**) for data variation, we performed here a LOO cross-validation (LOO-CV). The second model obtained with non-stochastic quadratic indices (Eq. **2**) showed a good value of square correlation coefficient q^2=0.769; this value of q^2 ($q^2 > 0.5$) can be considered as a proof of the high-predictive ability of the model [57, 58]. This was corroborated with the predicition of an external set of compounds that were not included in the trainig set used to develop the model. The second QSAR model achieved a 13.97% decrease in s_{cv} with regard to the initial model, which contains some outliers.

Table 1. Experimental and predicted values [Log ($1/IGC_{50}$)] for the training set.

Compounds	CAS	Log ($1/IGC_{50}$) Obs.[a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
benzene	71-43-2	-0.12	-0.083	-0.248	-0.195	-0.377
p-xylene	106-42-3	0.25	0.199	0.110	0.109	0.014
1-phenyl-2-butanol	120055-09-6	-0.16	0.605	0.563	0.105	0.159
toluene	108-88-3	0.25	0.058	-0.068	-0.042	-0.180
n-amylbenzene[b]	538-68-1	1.79	0.784	*-np-*	0.885	0.878
benzylamine	100-46-9	-0.24	-0.725	-0.723	-0.791	-0.819
5-phenyl-1-pentanol	10521-91-2	0.42	0.374	0.435	0.440	0.496
α,α-dimethylbenzenepropanol	103-05-9	-0.07	0.578	0.642	0.309	0.399
4-phenyl-1-butanol	3360-41-6	0.12	0.193	0.227	0.205	0.229
benzyl alcohol	100-51-6	-0.83	-0.240	-0.308	-0.246	-0.345
sec-phenethyl alcohol	98-85-1	-0.66	0.037	0.000	0.001	-0.054
4-ethylbenzyl alcohol	768-59-2	0.07	0.084	0.083	0.120	0.100
3-phenyl-1-butanol	2722-36-3	0.01	0.186	0.223	0.168	0.199
(R)-1-phenyl-1-butanol	22144-60-1	-0.01	0.412	0.429	0.518	0.522
4-biphenylmethanol	3597-91-9	0.92	0.391	0.612	0.219	0.432
4-ethylbiphenyl[b,c]	5707-44-8	1.97	0.832	*-np-*	0.621	*-np-*
biphenyl	92-52-4	1.05	0.519	0.647	0.265	0.394
(±)-1,2-diphenyl-2-propanol	5342-87-0	0.8	1.084	1.335	0.768	1.054
3,4-dimethylaniline	95-64-7	-0.16	0.002	-0.019	-0.217	-0.224
4-pentyloxyaniline	39905-50-5	0.97	0.645	0.702	0.922	0.990
4-hexyloxyaniline	39905-57-2	1.38	0.830	0.915	1.155	1.256
4-isopropylaniline	99-88-7	0.22	0.180	0.185	-0.005	0.025
3-ethylaniline	587-02-0	-0.03	0.007	-0.018	-0.167	-0.179
4-ethylaniline	589-16-2	0.03	-0.003	-0.027	-0.168	-0.181
(2-bromoethyl)benzene	103-63-9	0.42	0.603	0.721	0.794	0.780
2-methylaniline	95-53-4	-0.16	-0.110	-0.148	-0.342	-0.406
2,6-diisopropylaniline	24544-04-5	0.76	0.869	0.971	0.518	0.668
aniline	62-53-3	-0.23	-0.391	-0.463	-0.556	-0.652
2,6-diethylaniline	579-66-8	0.31	0.448	0.480	0.161	0.232
thioanisole	100-68-5	0.18	0.550	0.478	0.412	0.376
3,4,5-trimethylphenol	527-54-8	0.93	0.410	0.400	0.371	0.342
benzyl chloride	100-44-7	0.06	0.277	0.201	0.502	0.391
2,4,6-trimethylphenol	527-60-6	0.42	0.422	0.410	0.289	0.272
4-*tert*-butylphenol	98-54-4	0.91	0.590	0.606	0.555	0.565
4-*tert*-pentylphenol	80-46-6	1.23	0.769	0.814	0.878	0.909
2,3,6-trimethylphenol	2416-94-6	0.28	0.441	0.423	0.280	0.265
anisole	100-66-3	-0.1	0.011	-0.103	0.077	-0.065
2,4-dimethylphenol	105-67-9	0.14	0.267	0.215	0.155	0.092
2-phenyl-3-butyn-2-ol	127-66-2	-0.18	0.443	0.503	0.262	0.305
p-cresol	106-44-5	-0.16	0.070	-0.006	0.058	-0.059
4-ethylphenol	123-07-9	0.21	0.250	0.203	0.279	0.197
4-propylphenol	645-56-7	0.64	0.436	0.416	0.528	0.477
nonylphenol	104-40-5	2.47	1.586	1.719	1.913	2.062
m-cresol	108-39-4	-0.08	0.061	-0.014	0.041	-0.073
o-cresol	95-48-7	-0.29	0.070	-0.006	0.008	-0.102
2-ethylphenol	90-00-6	0.16	0.254	0.206	0.215	0.142
phenol	108-95-2	-0.35	-0.104	-0.213	-0.128	-0.285
2-allylphenol	1745-81-9	0.33	0.393	0.386	0.301	0.278
iodobenzene	591-50-4	0.36	0.633	0.535	0.279	0.257

Table 1. *Cont…*

Compounds	CAS	Log (1/IGC$_{50}$) Obs. [a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
2-tolunitrile	529-19-1	-0.24	0.180	0.167	-0.002	-0.045
4-hydroxyphenethyl alcohol	501-94-0	-0.83	-0.102	-0.085	-0.004	-0.036
2-chloro-4-methylaniline	615-65-6	0.18	0.297	0.282	0.275	0.248
2-chloroaniline	95-51-2	-0.17	0.108	0.062	0.102	0.033
5-pentylresorcinol	500-66-3	1.31	0.921	0.983	1.247	1.269
3-methoxyphenol	150-19-6	-0.33	0.059	-0.016	0.160	0.045
4-hexylresorcinol[b]	136-77-6	1.8	0.734	*-np-*	0.997	0.989
4-chloro-3,5-dimethylphenol	88-04-0	1.2	0.731	0.717	0.735	0.697
4-bromotoluene	106-38-7	0.47	0.541	0.468	0.475	0.398
1-bromo-4-ethylbenzene	1585-07-5	0.67	0.713	0.670	0.682	0.642
4-chloro-3-methylphenol	59-50-7	0.8	0.558	0.510	0.605	0.520
bromobenzene	108-86-1	0.08	0.402	0.290	0.310	0.191
4-chlorophenol	106-48-9	0.54	0.386	0.308	0.486	0.352
4-iodophenol	540-38-5	0.85	0.699	0.636	0.607	0.603
2-(4-chlorophenyl)ethylamine	156-41-2	0.14	-0.260	-0.162	-0.077	-0.041
2.4-dichloroaniline	554-00-7	0.56	0.597	0.583	0.779	0.732
chlorobenzene	108-90-7	-0.13	0.329	0.214	0.359	0.200
3-chloroaniline	108-42-9	0.22	0.107	0.058	0.125	0.054
1,2-dimethyl-4-nitrobenzene	99-51-4	0.59	0.730	0.711	0.704	0.649
4-(pentyloxy)benzaldehyde	5736-91-4	1.18	1.055	1.090	1.467	1.503
4-nitrotoluene	99-99-0	0.65	0.597	0.537	0.525	0.430
4-isopropylbenzaldehyde	122-03-2	0.67	0.543	0.534	0.601	0.597
1,2-dimethyl-3-nitrobenzene	83-41-0	0.56	0.724	0.706	0.654	0.606
3-chlorophenol	108-43-0	0.87	0.386	0.308	0.470	0.338
3-nitrotoluene	99-08-1	0.42	0.597	0.537	0.516	0.422
1,4-dibromobenzene	106-37-6	0.68	0.974	0.892	0.760	0.712
benzaldehyde	100-52-7	-0.2	0.066	-0.045	0.049	-0.073
4-hydroxypropiophenone	70-70-2	0.12	0.645	0.639	0.453	0.449
2,4-dichlorophenol	120-83-2	1.04	0.872	0.822	1.017	0.918
valerophenone	1009-14-9	0.56	0.951	0.963	0.928	0.949
propiophenone	93-55-0	-0.07	0.583	0.540	0.319	0.294
butyrophenone	495-40-9	0.21	0.766	0.751	0.576	0.581
2-hydroxybenzaldehyde	90-02-8	0.42	0.141	0.065	0.120	0.029
heptanophenone	1671-75-6	1.56	1.323	1.388	1.434	1.515
acetophenone	98-86-2	-0.46	0.406	0.334	0.107	0.046
nitrobenzene	98-95-3	0.14	0.459	0.360	0.332	0.199
octanophenone	1674-37-9	1.89	1.508	1.601	1.658	1.773
2,5-dichloroaniline	95-82-9	0.58	0.598	0.578	0.783	0.741
3,4-dichlorotoluene	95-75-0	1.07	0.997	0.939	0.985	0.903
3-nitroaniline	99-09-2	0.03	0.270	0.233	0.096	0.048
3,5-dichloroaniline	626-43-7	0.71	0.596	0.577	0.795	0.750
3-nitroanisole	555-03-3	0.72	0.638	0.568	0.601	0.508
benzophenone	119-61-9	0.87	0.885	1.064	0.633	0.810
3-chloro-5-methoxyphenol	65262-96-6	0.76	0.951	0.948	0.823	0.810
4-nitrobenzyl chloride	100-14-1	1.18	0.903	0.871	1.037	0.974
2,4-dibromophenol	615-58-7	1.4	1.026	0.983	1.143	1.117
2-amino-5-chlorobenzonitrile	5922-60-1	0.44	0.729	0.705	0.209	0.235
2-hydroxy-4-methoxyacetophenone	552-41-0	0.55	0.662	0.657	0.504	0.511

Table 1. *Cont…*

Compounds	CAS	Log (1/IGC_{50}) Obs. [a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
3,5-dichlorophenol	591-35-5	1.56	0.887	0.833	1.022	0.924
4-chlorobenzophenone	134-85-0	1.5	1.355	1.567	1.146	1.351
1,3,5-trichlorobenzene	108-70-3	0.87	1.327	1.264	1.312	1.219
2,4,5-trichloroaniline	636-30-6	1.3	1.086	1.098	1.430	1.414
4-bromobenzophenone	90-90-4	1.26	1.428	1.644	1.101	1.346
2,4,6-trichlorophenol	88-06-2	1.41	1.350	1.333	1.559	1.494
4-ethoxy-2-nitroaniline	616-86-4	0.76	0.798	0.815	0.756	0.785
5-bromovanillin	2973-76-4	0.62	0.836	0.834	0.981	1.003
4-nitrophenetole	100-29-8	0.83	1.152	1.168	0.917	0.932
1-bromo-3-nitrobenzene	585-79-5	1.03	1.030	0.961	0.781	0.719
4-bromo-2,6-dichlorophenol	3217-15-0	1.78	1.443	1.424	1.601	1.576
2-chloro-6-nitrotoluene	83-42-1	0.68	1.088	1.056	0.944	0.890
2,3,5,6-tetrachloroaniline	3481-20-7	1.76	1.577	1.619	2.048	2.067
2,4,5-trichlorophenol	95-95-4	2.1	1.366	1.344	1.567	1.502
1,2,4,5-tetrachlorobenzene	95-94-3	2	1.796	1.768	1.786	1.727
4-methyl-2-nitroaniline	89-62-3	0.37	0.528	0.514	0.258	0.255
1-chloro-3-nitrobenzene	121-73-3	0.73	0.957	0.885	0.831	0.728
2,3,4,5-tetrachloroaniline	634-83-3	1.96	1.575	1.618	2.049	2.065
2,4,6-tribromophenol	118-79-6	1.91	1.598	1.585	1.756	1.802
2-bromo-5-nitrotoluene	7149-70-4	1.16	1.138	1.116	0.968	0.944
1-fluoro-3-iodo-5-nitrobenzene	3819-88-3	1.09	1.579	1.552	1.072	1.099
2-nitrophenol	88-75-5	0.67	0.549	0.482	0.391	0.293
2-chloro-4-nitroaniline	121-87-9	0.75	0.745	0.742	0.732	0.711
5-hydroxy-2-nitrobenzaldehyde	42454-06-8	0.33	0.711	0.694	0.730	0.679
3,4,5,6-tetrabromo-o-cresol	576-55-6	2.57	2.326	2.383	2.448	2.609
Pentafluoroaniline[b]	771-60-8	0.26	1.224	*-np-*	1.047	0.965
1-bromo-2-nitrobenzene	577-19-5	0.75	0.999	0.939	0.759	0.699
3,5-dichloro-nitrobenzene	618-62-2	1.13	1.426	1.388	1.334	1.261
2,3,4,5-tetrachlorophenol	4901-51-3	2.72	1.843	1.855	2.093	2.064
thiobenzamide	2227-79-4	0.09	0.325	0.357	-0.179	-0.125
α,α,α, 4-tetrafluoro-*m*-toluidine	2357-47-3	0.77	1.009	1.037	0.724	0.690
1-chloro-2-nitrobenzene	88-73-3	0.68	0.926	0.862	0.802	0.703
4-chloro-6-nitro-*m*-cresol	7147-89-9	1.63	1.198	1.195	1.129	1.101
pentachlorophenol	87-86-5	2.07	2.321	2.366	2.604	2.613
1,3-dinitrobenzene	99-65-0	0.76	1.054	1.007	0.870	0.784
2,4-dinitrotoluene	121-14-2	0.87	1.185	1.179	1.036	0.992
4,5-dichloro-2-nitroaniline	6641-64-1	1.66	1.317	1.335	1.378	1.382
pentafluorophenol	771-61-9	1.63	1.530	1.546	1.190	1.060
pentabromophenol	608-71-9	2.66	2.708	2.768	2.975	3.158
3-chloro-4-fluoronitrobenzene	350-30-1	0.8	1.246	1.209	1.183	1.068
1,4-dinitrobenzene	100-25-4	1.3	1.025	0.985	0.909	0.817
3,4-dichloronitrobenzene	99-54-7	1.16	1.396	1.366	1.361	1.284
2,4-dichloro-6-nitroaniline	2683-43-4	1.26	1.295	1.313	1.346	1.359
3,4-dinitrobenzyl alcohol	79544-31-3	1.09	0.996	1.089	1.115	1.139
2,3-dichloronitrobenzene	3209-22-1	1.07	1.395	1.365	1.294	1.226
1,2-dinitrobenzene	528-29-0	1.25	1.013	0.976	0.826	0.746
phenyl isothiocyanate[c]	103-72-0	1.41	0.969	0.919	0.202	*-np-*
3-trifluoromethyl-4-nitrophenol	88-30-2	1.65	0.800	0.792	0.937	0.854
2,6-iodo-4-nitrophenol	305-85-1	1.81	1.169	1.157	1.030	0.925

Table 1. *Cont…*

Compounds	CAS	Log (1/IGC$_{50}$) Obs. [a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
2,4-chloro-6-nitrophenol	609-89-2	1.75	1.511	1.507	1.502	1.465
1,3,5-trichloro-2-nitrobenzene	18708-70-8	1.43	1.861	1.867	1.763	1.730
1,2,4-trichloro-5-nitrobenzene	89-69-0	1.53	1.893	1.890	1.781	1.745
1,2,3-trichloro-4-nitrobenzene	17700-09-3	1.51	1.893	1.890	1.759	1.726
2-chloro-5-nitrobenzaldehyde	6361-21-3	0.53	1.070	1.061	1.116	1.067
pentafluorobenzaldehyde	653-37-2	0.82	1.651	1.680	1.312	1.198
2,4-dinitro-1-iodobenzene	709-49-9	2.12	1.826	1.831	1.271	1.355
2,3,5,6-tetrachloronitrobenzene	117-18-0	1.82	2.359	2.392	2.198	2.204
2,5-dinitrophenol	329-71-5	1.04	1.153	1.140	1.019	0.960
2,4-dinitroaniline	97-02-9	0.72	0.973	0.977	0.729	0.725
2,3,4,5-tetrachloronitrobenzene	879-39-0	1.78	2.361	2.393	2.235	2.236
1,2-dichloro-4,5-dinitrobenzene	6306-39-4	2.21	1.979	2.004	1.833	1.812
2,6-dinitroaniline	606-22-4	0.84	1.005	1.004	0.683	0.690
4,6-dinitro-2-methylphenol	534-52-1	1.73	1.324	1.342	1.163	1.150
4-*tert*-butyl-2.6-dinitrophenol	4097-49-8	1.8	1.856	1.965	1.847	1.936
1,5-dichloro-2.3-dinitrobenzene	28689-08-9	2.42	1.977	2.002	1.785	1.771
6-chloro-2.4-dinitroaniline	3531-19-9	1.12	1.441	1.476	1.340	1.368
2-bromo-4.6-dinitroaniline	1817-73-8	1.24	1.525	1.563	1.487	1.548
2,3,4,6-tetrafluoronitrobenzene	314-41-0	1.87	1.755	1.760	1.287	1.169
Pentafluoronitrobenzene[c]	880-78-4	2.43	2.071	2.104	1.449	*-np-*
1,4-dinitrotetrachlorobenzene	20098-38-8	2.82	2.921	3.014	2.649	2.715
1,5-difluoro-2,4-dinitrobenzene	327-92-4	2.08	1.652	1.670	1.419	1.330
1,3-dinitro-2,4,5-trichlorobenzene	2678-21-9	2.6	2.453	2.511	2.234	2.257
1,3,5-trichloro-2,4-dinitrobenzene hemihydrate	6284-83-9	2.19	2.452	2.511	2.193	2.222
4-chloro-3,5-dinitrobenzaldehyde[b,c]	1930-72-9	2.66	1.660	*-np-*	1.592	*-np-*
1-phenyl-2-propanol	14898-87-4	-0.62	0.115	0.114	0.057	0.053
4-methylbenzyl alcohol	589-18-4	-0.49	-0.116	-0.149	-0.087	-0.143
(±)1-phenyl-2-pentanol	705-73-7	0.16	0.490	0.543	0.656	0.701
2-(*p*-tolyl)ethylamine	3261-62-9	-0.04	-0.596	-0.504	-0.568	-0.504
4-methyl benzylamine	104-84-7	-0.01	-0.601	-0.566	-0.632	-0.617
3-methylbenzyl alcohol	587-03-1	-0.24	-0.116	-0.148	-0.090	-0.145
3-phenyl-2-propen-1-ol	104-54-1	-0.08	-0.051	-0.030	-0.066	-0.065
4-*tert*-buthylbenzyl alcohol	877-65-6	0.48	0.387	0.450	0.371	0.447
4-methylphenetyl alcohol	699-02-5	-0.26	-0.014	-0.006	0.011	0.003
1-phenylethylamine	618-36-0	-0.18	-0.485	-0.458	-0.489	-0.479
2-methyl-1-phenyl-2-propanol	100-86-7	-0.41	0.387	0.417	0.109	0.166
(±)-1-phenyl-1-propanol	93-54-9	-0.43	0.191	0.179	0.262	0.236
phenetyl alcohol	60-12-8	-0.59	-0.157	-0.187	-0.145	-0.196
2-phenyl-1-butanol	89104-46-1	-0.11	0.196	0.226	0.265	0.287
benzhydrol	91-01-0	0.5	0.684	0.859	0.482	0.674
benzaldoxime	622-32-2	-0.11	-0.224	-0.298	-0.232	-0.281
3,5-dimethylaniline	108-69-0	-0.36	-0.039	-0.066	-0.228	-0.233
4-*tert*-buthylaniline	769-92-6	0.36	0.305	0.338	0.097	0.178
4-phenylbutyronitrile	2046-18-6	0.15	0.419	0.399	0.201	0.260
2,4,6-trimethylaniline	88-05-1	-0.05	0.195	0.192	-0.104	-0.059
3-phenylpropionitrile	645-59-0	-0.16	0.232	0.186	0.044	0.060
4-*sec*-butylaniline	30273-11-1	0.61	0.370	0.390	0.270	0.330
benzyl cyanide	140-29-4	-0.36	0.039	-0.032	-0.062	-0.097

Table 1. *Cont…*

Compounds	CAS	Log (1/IGC$_{50}$) Obs. [a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
2,5-dimethylaniline	95-78-3	-0.33	-0.017	-0.047	-0.239	-0.241
α-methylbenzyl cyanide	1823-91-2	0.01	0.173	0.142	0.101	0.108
2-isopropylaniline	643-28-7	0.12	0.179	0.173	-0.030	0.005
2,6-dimethylaniline	87-62-7	-0.43	0.005	-0.029	-0.259	-0.257
N-ethylaniline	103-69-5	0.07	-0.207	-0.250	-0.200	-0.221
2-propylaniline	1821-39-2	0.08	0.178	0.170	0.069	0.091
N-methylaniline	100-61-8	0.06	-0.452	-0.519	-0.607	-0.659
2-amino-4-*tert*-buthylaniline	1199-46-8	0.37	0.383	0.441	0.262	0.368
2-methoxyaniline	90-04-0	-0.69	-0.332	-0.393	-0.297	-0.336
3-phenylpyridine	1008-88-4	0.47	0.245	0.324	0.195	0.309
2-aminobenzyl alcohol	5344-90-1	-1.07	-0.524	-0.528	-0.561	-0.568
2-benzylpyridine	101-82-6	0.38	0.554	0.649	0.373	0.529
3,5-di-*tert*-buthylphenol	1138-52-9	1.64	1.269	1.400	1.161	1.350
phenyl propargyl sulfide	5651-88-7	0.54	0.483	0.465	0.472	0.530
4-ethoxyphenol	622-62-8	0.01	0.227	0.159	0.434	0.365
4-benzylpyridine	2116-65-6	0.63	0.472	0.567	0.383	0.536
3.4-dimethylphenol	95-65-8	0.12	0.218	0.166	0.193	0.124
3-*tert*-buthylphenol	585-34-2	0.74	0.582	0.586	0.527	0.543
3,5-dimethylphenol	108-68-9	0.11	0.224	0.170	0.179	0.112
6-*tert*-buthyl-2,4-dimethylphenol	1879-09-0	1.16	0.935	1.005	0.737	0.855
3-isopropylphenol	618-45-1	0.61	0.415	0.386	0.418	0.383
2,5-dimethylphenol	95-87-4	0.14	0.233	0.178	0.153	0.090
4-hydroxy-3-methoxybenzyl alcohol	498-00-0	-0.7	-0.179	-0.183	0.114	0.091
3-amino-2-cresol	53222-92-7	-0.55	-0.176	-0.203	-0.293	-0.309
4-chloro-2-methylaniline	95-69-2	0.35	0.172	0.137	0.277	0.252
2,4,6- tris(dimethylaminomethyl)phenol[c]	90-72-2	-0.52	-0.055	0.024	0.647	*-np-*
2-fluoroaniline	348-54-9	-0.37	-0.138	-0.198	-0.179	-0.276
4-aminobenzyl cyanide	3544-25-0	-0.76	-0.212	-0.220	-0.396	-0.341
3-iodoaniline	626-01-7	0.65	0.128	0.071	0.495	0.547
3-cinnamonitrile	4360-47-8	0.16	0.098	0.086	0.000	0.018
3-fluorobenzyl alcohol	456-47-3	-0.39	0.029	-0.015	0.141	0.028
3-cyanoaniline	2237-30-1	-0.47	-0.298	-0.345	-0.462	-0.458
4-fluorophenol	371-41-5	0.02	0.189	0.091	0.241	0.076
2-iodoaniline	615-43-0	0.35	0.131	0.074	0.528	0.582
3-fluoroaniline	372-19-0	-0.1	-0.137	-0.198	-0.134	-0.238
4-chloro-2-methylphenol	1570-64-5	0.7	0.444	0.380	0.593	0.511
2-chloro-4,5-dimethylphenol	1124-04-5	0.69	0.595	0.568	0.762	0.722
3,5-dimethoxyphenol	500-99-2	-0.09	0.047	-0.019	0.431	0.363
4-hydroxybenzyl cyanide	14191-95-8	-0.38	0.061	0.024	0.002	-0.004
4-bromo-2,6-dimethylphenol	2374-05-2	1.16	0.654	0.624	0.755	0.759
2-bromobenzyl alcohol	18982-54-2	0.1	0.141	0.099	0.352	0.321
2-chloro-5-methylphenol	615-74-7	0.54	0.436	0.373	0.590	0.510
2-fluorophenol	367-12-4	0.19	0.188	0.090	0.180	0.024
4-(dimethylamino)benzaldehyde	100-10-7	0.23	-0.114	-0.183	0.169	0.170
4-bromophenol	106-41-2	0.68	0.306	0.209	0.510	0.419
3-chloro-2-methylaniline	95-79-4	0.5	0.173	0.138	0.277	0.252
3-chloro-4-methylaniline	95-74-9	0.39	0.144	0.113	0.260	0.237
4-chlorophenethyl alcohol	1875-88-3	0.32	0.212	0.210	0.454	0.423
4-chlorobenzyl alcohol	873-76-7	0.25	0.107	0.064	0.347	0.271

Table 1. *Cont….*

Compounds	CAS	Log (1/IGC$_{50}$) Obs. [a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
2-bromo-4-methylphenol	6627-55-0	0.6	0.470	0.408	0.657	0.612
1,3,5-trimethyl-2-nitrobenzene	603-71-4	0.86	0.583	0.576	0.754	0.757
2-bromophenol	95-56-7	0.33	0.307	0.210	0.493	0.407
4-hydroxy-3-methoxybenzonitrile	4421-08-3	-0.03	-0.048	-0.079	0.096	0.095
3-nitrobenzyl alcohol	619-25-0	-0.22	0.012	-0.022	0.358	0.301
4-methoxybenzonitrile	874-90-8	0.1	-0.119	-0.178	0.064	0.025
2-hydroxy-4,5-dimethylacetophenone	36436-65-4	0.71	0.660	0.680	0.532	0.584
2-anisaldehyde	135-02-4	0.15	0.008	-0.079	0.264	0.190
methyl-4-methylaminobenzoate	18358-63-9	0.31	-0.164	-0.169	-0.175	-0.105
4-phenoxybenzaldehyde	67-36-7	1.26	0.795	0.945	0.852	1.031
3-hydroxy-4-methoxybenzaldehyde	621-59-0	-0.14	0.047	-0.008	0.378	0.332
4-benzoylaniline	1137-41-3	0.68	0.562	0.786	0.332	0.596
3-anisaldehyde	5991-31-1	0.23	0.012	-0.076	0.306	0.226
n-propyl cinnamate	7778-83-8	1.23	0.873	0.911	0.987	1.091
(*trans*)ethyl cinnamate	103-36-6	0.99	0.686	0.697	0.624	0.713
hexanophenone	942-92-7	1.19	1.012	1.043	1.225	1.271
n-butyl cinnamate	538-65-8	1.53	1.058	1.123	1.355	1.473
4-chlorobenzyl cyanide	140-53-4	0.66	0.407	0.363	0.455	0.448
(*trans*)methyl cinnamate	103-26-4	0.58	0.430	0.415	0.267	0.324
ethyl-4-methoxybenzoate	94-30-4	0.77	0.479	0.441	0.741	0.766
phenylacetic acid hydrazide	937-39-3	-0.48	-1.093	-0.971	-1.337	-1.160
2,6-dichlorophenol	87-65-0	0.73	0.647	0.576	0.968	0.879
benzyl methacrylate	2495-37-6	0.65	0.720	0.723	0.572	0.668
isoamyl-4-hydroxybenzoate	6521-30-8	1.48	0.984	1.033	1.513	1.589
benzyl-4-hydroxyphenyl ketone[b,c]	2491-32-9	1.07	2.509	*-np-*	2.100	*-np-*
benzyl benzoate	120-51-4	1.45	0.947	1.115	0.948	1.176
2-methyl-5-nitrophenol	5428-54-6	0.66	0.369	0.311	0.602	0.539
3-acetoamidophenol	621-42-1	-0.16	0.047	0.039	-0.216	-0.149
2-nitrobiphenyl	86-00-0	1.3	0.766	0.936	0.831	1.003
5-chloro-2-hydroxybenzamide	7120-43-6	0.59	0.024	0.062	0.110	0.172
3-nitrophenol	554-84-7	0.51	0.195	0.104	0.436	0.330
phenyl-1.3-dialdehyde	626-19-7	0.18	-0.021	-0.093	0.309	0.244
ethyl-4-bromobenzoate	5798-75-4	1.33	0.815	0.775	0.913	0.969
2,4-dihydroxyacetophenone	89-84-9	0.25	0.371	0.353	0.340	0.333
phenyl-4-hydroxybenzoate	17696-62-7	1.37	0.965	1.142	0.974	1.181
2-hydroxy-4-methoxybenzophenone	131-57-7	1.42	0.966	1.182	1.009	1.255
benzylidene malononitrile	2700-22-3	0.64	-0.201	-0.086	0.149	0.247
4-nitrophenyl phenyl ether	620-88-2	1.58	1.029	1.170	1.112	1.282
resorcinol monobenzoate	136-36-7	1.11	0.947	1.127	0.957	1.167
4-bromophenyl-3-pyridyl ketone[b,c]	14548-45-9	0.82	2.348	*-np-*	2.444	*-np-*
3-nitroacetophenone	121-89-1	0.32	0.543	0.501	0.683	0.664
3-nitrobenzaldehyde	99-61-6	0.11	0.212	0.131	0.590	0.514
ethyl phenylcyanoacetate	4553-07-5	-0.02	0.558	0.578	0.657	0.782
2-nitroanisole	91-23-6	-0.07	0.240	0.145	0.508	0.428
3-methyl-2-nitrophenol	4920-77-8	0.61	0.373	0.315	0.520	0.470
2,5-diphenyl-1,4-benzoquinone[b]	844-51-9	1.48	0.575	*-np-*	1.396	1.821
2-nitrobenzamide	610-15-1	-0.72	-0.107	-0.091	-0.092	-0.036
methyl-2,5-dichlorobenzoate	2905-69-3	0.81	0.800	0.784	1.234	1.230
4-methyl-2-nitrophenol	119-33-5	0.57	0.375	0.316	0.570	0.513

Table 1. *Cont…*

Compounds	CAS	Log (1/IGC$_{50}$) Obs.[a]	Non-stochastic		Stochastic	
			Eq. 1	Eq. 2	Eq. 3	Eq. 4
2,2',4,4'-tetrahydroxybenzophenone	131-55-5	0.96	1.141	1.402	1.065	1.353
4-nitrobenzaldehyde	555-16-8	0.2	0.214	0.132	0.585	0.510
3,5-dichlorosalicylaldehyde	90-60-8	1.55	0.740	0.706	1.262	1.227
2-(benzylthio)-3-nitropyridine	69212-31-3	1.72	1.088	1.218	1.341	1.631
ethyl-4-nitrobenzoate	99-77-4	0.71	0.679	0.648	1.026	1.055
2,4-dichlorobenzaldehyde	874-42-0	1.04	0.683	0.620	1.054	0.991
2',3',4'-trichloroacetophenone	13608-87-2	1.34	1.349	1.360	1.652	1.676
2,2'-dihydroxybenzophenone	835-11-0	1.16	1.020	1.225	0.828	1.060
2-chloromethyl-4-nitrophenol	2973-19-5	0.75	0.604	0.570	1.135	1.105
α,α,α-trifluoro-*p*-cresol	402-45-9	0.62	0.852	0.809	0.790	0.691
dimethylnitroterephthalate	5292-45-5	0.43	0.712	0.839	0.764	0.960
thioacetanilide	637-53-6	-0.01	0.262	0.233	0.166	0.242
2-nitroresorcinol	601-89-8	0.66	0.257	0.195	0.479	0.419
3,5-dibromo-4-hydroxybenzonitrile	1689-84-5	1.16	0.701	0.693	1.073	1.157
methyl-4-chloro-2-nitrobenzoate	42087-80-9	0.82	0.711	0.702	1.298	1.308
1-fluoro-2-nitrobenzene	1493-27-2	0.23	0.473	0.372	0.574	0.444
α,α,α-tetrafluoro-*o*-toluidine	393-39-5	-0.02	0.897	0.911	0.838	0.788
benzoyl cyanide	613-90-1	0.31	-0.098	-0.105	0.212	0.173
2,5-difluoronitrobenzene	364-74-9	0.33	0.766	0.690	0.899	0.761
4-hydroxy-3-nitrobenzaldehyde	3011-34-5	0.61	0.279	0.226	0.679	0.637
benzoyl isothiocyanate	532-55-8	0.10	0.429	0.367	0.415	0.484

[a]Experimental values (cocentration in mmol/L) taken from [12], [b]statistical outliers for Eq. **1**, [c]statistical outliers for Eq. **3**, np: Not performed

On the other hand, the stochastic quadratic indices were also employed to develop a QSAR model to predict the aquatic toxicity of benzene derivatives. The first obtained model, using these atom-based quadratic indices as molecular descriptors, together with its statistical parameters, is given below:

$$\textbf{Log (1/IGC}_{50}\textbf{)} = -0.870(\pm0.105) + 8.04\text{x}10^{-2}(\pm0.64\text{x}10^{-2})\,{}^{Ks}\boldsymbol{q}_{0L}(x_E) + 3.65\text{x}10^{-2}(\pm0.50\text{x}10^{-2})\,{}^{Ps}\boldsymbol{q}_{1}(x) - 6.65\text{x}10^{-2}(\pm0.79\text{x}10^{-2})\,{}^{Ks}\boldsymbol{q}_{4L}{}^{H}(x_E) - 0.187(\pm0.024)\,{}^{Ks}\boldsymbol{q}_{3L}{}^{H}(x_{E\text{-}H}) + 0.101(\pm0.017)\,{}^{As}\boldsymbol{q}_{14L}{}^{H}(x_{E\text{-}H}) + 0.167(\pm0.017)\,{}^{Ks}\boldsymbol{q}_{15}{}^{H}(x) - 0.201(\pm0.021)\,{}^{Gs}\boldsymbol{q}_{9}{}^{H}(x) \quad \textbf{(3)}$$

$N = 313$ $R^2 = 0.745$ $s = 0.385$ $F = 127.43$ $p < 0.0001$

$q^2 = 0.712$ $s_{cv} = 0.405$

This model showed a square correlation coefficient of 0.745, which is slightly better than the one obtained with the first non-stochastic model (R^2=0.730); the same behaviour can be observed with the value of the standard deviation. In the development of the first stochastic model (Eq. **3**), seven compounds (020, 182, 210, 215, 279, 335 and 354) were detected as statistical outliers. The residual values of these compounds, together with their chemical names, are also shown in Table **2**. The removal of the above-noted compounds and subsequent

reanalysis lead to Eq. **4**, which exhibits better statistics. This new model obtained with stochastic atom-based quadratic indices, together with its statistical parameters, is given below:

$$\textbf{Log (1/IGC}_{\textbf{50}}\textbf{)} = -1.337(\pm 0.108) + 7.09\text{x}10^{-2}(\pm 0.58\text{x}10^{-2})\,{}^{Ks}\boldsymbol{q}_{0L}(x_E) + 5.07\text{x}10^{-2}(\pm 0.49\text{x}10^{-2})\,{}^{Ps}\boldsymbol{q}_{1}(x) - 5.69\text{x}10^{-2}(\pm 0.71\text{x}10^{-2})\,{}^{Ks}\boldsymbol{q}_{4L}{}^{H}(x_E) - 0.175(\pm 0.021)\,{}^{Ks}\boldsymbol{q}_{3L}{}^{H}(x_{E\text{-}H}) + 9.69\text{x}10^{-2}(\pm 1.51\text{x}10^{-2})\,{}^{As}\boldsymbol{q}_{14L}{}^{H}(x_{E\text{-}H}) + 0.144(\pm 0.015)\,{}^{Ks}\boldsymbol{q}_{15}{}^{H}(x) - 0.178(\pm 0.019)\,{}^{Gs}\boldsymbol{q}_{9}{}^{H}(x) \quad \textbf{(4)}$$

N = 306 $R^2 = 0.806$ $s = 0.329$ F = 176.99 $p < 0.0001$

$q^2 = 0.791$ $s_{cv} = 0.337$ $R^2_{pred} = 0.742$

This improved model explains more than the 80% of the experimental values of aquatic toxicicty, with a standard deviation 14.5% lower than the one of the former model obtained with the entire dataset. The predictability and stability of the new obtained models, using stochastic linear indices (Eqs. **3** and **4**) for data variation, were also carried out here by means of LOO cross-validation. The second stochastic model (Eq. **4**) showed a good value of square correlation coefficient q^2=0.791, which is 11.09% greater than the value of q^2 of the first stochastic model (0.712). Moreover, the standard deviation of the LOO-CV was improved in 16.79% with regard to the one of the previously obtained model (Eq. **3**). The value of q^2 (0.791) can be considered as a proof of the high-predictive ability of the model. However, the external validation is the only way to establish the real predictivity of the models [59]; this topic will be disscussed in the next subsection.

Table 2. Statistical outliers and residual values from Eqs. **1** and **3**

Compound	Residual value
Non-stochastic model (Eq. 1)	
n-amylbenzene	1.006
4-ethylbiphenyl	1.139
4-hexylresorcinol	1.066
pentafluoroaniline	-0.964
4-chloro-3,5-dinitrobenzaldehyde	0.999
benzyl-4-hydroxyphenyl ketone	-1.439
4-bromophenyl-3-pyridyl ketone	-1.528
2,5-diphenyl-1,4-benzoquinone	0.905
Stochastic model (Eq. 3)	
4-ethylbiphenyl	1.349
phenyl isothiocyanate	1.209
pentafluoronitrobenzene	0.981
4-chloro-3,5-dinitrobenzaldehyde	1.068
2,4,6-tris(dimethylaminomethyl)phenol	-1.167
benzyl-4-hydroxyphenyl ketone	-1.030
4-bromophenyl-3-pyridyl ketone	-1.624

3.3. Validation of the toxicity-based QSAR models

All toxicity-related QSARs require validation to ensure they are capable of making accurate predictions of toxicity for compounds not included in the training set. The best means of validation is by using of an external data set. This is the most demanding method because it requires additional testing and attention to the selection of compounds for validation [12]. Efforts should be made to ensure chemical diversity within the training set, and the chemicals in the validation set should be similar to the ones in the training set [59]. The training chemicals should represent the depth and breadth of all existing chemicals within the domain. The chemicals selected for the test set should also represent the distribution of existing chemicals within the training domain. In this work, CA was used to assess both diversity for training and representation for validation.

The principal importance of the horizontal validation is to prove the predictability and the robustness of the model. An external set of 79 benzene derivatives was used as a test set to judge the predictability of the best model obtained with the non-stochastic quadratic indices (Eq. **2**). Therefore, the determination coefficient for the test set (R^2_{pred}) with model **2** was of 0.745; the good prediction for the tested compounds confirms the significance of the selected molecular descriptors and the model based on them. Two compounds (pentafluorobenzyl alcohol, Res=3.049 and 6-phenyl-1-hexanol, Res=1.022) were detected as ouliers. The predicted values for the compounds of the prediction set, using the non-stochastic linear indices (Eq. **2**) are shown in Table **3**.

Likewise, the real predictive power of the best stochastic quadratic indices' model (Eq. **4**) was validated by the same external test set of 79 compounds, achieving a value of R^2_{pred} of 0.742, with two compounds as outliers (pentafluorobenzyl alcohol, Res=0.832 and 2,4,5-trimethoxybenzaldehyde, Res=0.771). The obtained values for the test sets, using stochastic linear indices (Eq. **4**), are also shown in Table **3**.

Now we shall give a little discussion about the presence of outliers in the developed QSAR models. Outliers are useful in QSAR development as they assist in establishing the chemical domain of the model. Outliers from a QSAR are compounds that do not fit the model, or that are poorly predicted by it [60]. By the use of several methods, it is possible for us to highlight outliers including, at the most basic level, the identification of those compounds with significantly high standard residuals from regression-based techniques by use of several methods. In this work, outliers' detection was performed by using the following standard statistical tests: residual, standardized residual, Mahalanobis distance, deleted residual and

Cooks' distance [56, 58]. After their identification, outliers were removed from the data set, and the QSAR recalculated (as we described in the previous section) to develop new models.

Table 3. Experimental and predicted values [Log ($1/IGC_{50}$)] for the test set.

Compounds	CAS	Log ($1/IGC_{50}$) Obs. [a]	Non-stochastic	Stochastic
			Eq. 2	Eq. 4
n-butylbenzene	104-51-8	1.25	0.557	0.606
isopropylbenzene	98-82-8	0.69	0.332	0.245
6-phenyl-1-hexanol	2430-16-2	0.87	-outlier-	0.729
3-phenyl-1-propanol	122-97-4	-0.21	0.025	0.010
(±)-2-phenyl-2-butanol	1565-75-9	0.06	0.518	0.459
1,1-diphenyl-2-propanol	29338-49-6	0.75	1.282	0.849
3-aminobenzyl alcohol	1877-77-6	-1.13	-0.468	-0.548
4-butoxyaniline	4344-55-2	0.61	0.490	0.717
4-methylaniline	106-49-0	-0.05	-0.244	-0.433
3-methylaniline	108-44-1	0.28	-0.238	-0.436
4-butylaniline	104-13-2	1.07	0.402	0.401
2-ethylaniline	578-54-1	-0.22	0.008	-0.203
4-methoxyphenol	150-76-5	-0.14	-0.006	0.067
4-methylanisole	104-93-8	0.25	0.074	0.142
2,3,5-trimethylphenol	697-82-5	0.36	0.400	0.300
phenetole	103-73-1	-0.14	0.206	0.223
3-ethylphenol	620-17-7	0.29	0.196	0.183
4-chloroaniline	106-47-8	0.05	0.061	0.061
4-chloroanisole	623-12-1	0.6	0.400	0.503
1,3-dihydroxybenzene	108-46-3	-0.65	-0.116	-0.160
4-chlorobenzylamine	104-86-9	0.16	-0.220	-0.142
2-nitrotoluene	88-72-2	0.26	0.532	0.384
3-ethoxy-4-hydroxybenzaldehyde	121-32-4	0.02	0.561	0.691
3-methoxy-4-hydroxybenzaldehyde	121-33-5	-0.03	0.249	0.332
4-bromo-6-chloro-*o*-cresol	7530-27-0	1.28	1.120	1.178
1,2-dichlorobenzene	95-50-1	0.53	0.760	0.687
4-chlorobenzaldehyde	104-88-1	0.4	0.459	0.478
1,2,4-trichlorobenzene	120-82-1	1.08	1.264	1.218
4-chloro-2-nitrotoluene	89-59-8	0.82	1.034	0.941
3-nitrobenzonitrile	619-24-9	0.45	0.664	0.278
2-nitroaniline	88-74-4	0.08	0.292	0.024
2,3,4,6-tetrachlorophenol	58-90-2	2.18	1.855	2.052
1-fluoro-4-nitrobenzene	350-46-9	0.1	0.706	0.554
3,5-dibromo-salicylaldehyde	90-59-5	1.65	1.248	1.402
4-chloro-3-nitrophenol	610-78-6	1.27	0.979	0.901
1-chloro-4-nitrobenzene	100-00-5	0.43	0.863	0.757
2,5-dichloronitrobenzene	89-61-2	1.13	1.387	1.225
2,4-dichloronitrobenzene	611-06-3	0.99	1.365	1.246
1,2,3-trifluoro-4-nitrobenzene	771-69-7	1.89	1.417	0.980
1-bromo-2.4-dinitrobenzene	584-48-5	2.31	1.563	1.309
2,4-dinitrophenol	51-28-5	1.06	1.134	0.952
2,6-dinitrophenol	573-56-8	0.83	1.145	0.886
1-chloro-2,4-dinitrobenzene	97-00-7	2.16	1.486	1.315
2,4-dinitro-1-fluorobenzene	70-34-8	1.71	1.328	1.081
4-isopropylbenzyl alcohol	536-60-7	0.18	0.388	0.445
2-methylbenzyl alcohol	89-95-2	-0.43	-0.147	-0.147

Table 3. *Cont ...*

Compounds	CAS	Log (1/IGC$_{50}$) Obs. [a]	Non-stochastic	Stochastic
			Eq. 2	Eq. 4
N-methylphenethylamine	589-08-2	-0.41	-0.519	-0.659
β-methylphenethylamine	582-22-9	-0.28	-0.478	-0.443
(±)-1-phenyl-1-butanol	22135-49-5	-0.09	0.393	0.522
2-phenyl-1-propanol	1123-85-9	-0.4	0.016	0.029
2-phenyl-2-propanol	617-94-7	-0.57	0.266	0.129
2.4-dimethylaniline	95-68-1	-0.29	-0.054	-0.242
2.3-dimethylaniline	87-59-2	-0.43	-0.048	-0.238
4-buthoxyphenol	122-94-1	0.7	0.586	1.050
2-phenylpyridine	1008-89-5	0.27	0.406	0.310
4-isopropylphenol	99-89-8	0.47	0.381	0.384
2.3-dimethylphenol	526-75-0	0.12	0.176	0.095
2-isopropylphenol	88-69-7	0.61	0.398	0.336
2-methoxy-4-propenylphenol	97-54-1	0.75	0.441	0.499
4-chloro-3-ethylphenol	14143-32-9	1.08	0.582	0.772
3-chloro-2-methylaniline	87-60-5	0.38	0.138	0.224
3-chlorobenzyl alcohol	873-63-2	0.15	0.064	0.270
4-bromophenyl acetonitrile	16532-79-9	0.6	0.396	0.441
4-chlororesorcinol	95-88-5	0.13	0.209	0.507
4-biphenylcarboxaldehyde	3218-36-8	1.12	0.724	0.688
2.4.5-trimethoxybenzaldehyde	4460-86-0	-0.1	0.106	-outlier-
3-hydroxybenzaldehyde	100-83-4	0.08	-0.116	0.047
4-benzoylphenol	1137-42-4	1.02	1.007	0.914
4-cyanobenzamide	3034-34-2	-0.38	-0.409	-0.597
3-chlorobenzophenone	1016-78-0	1.55	1.325	1.356
phenyl benzoate	93-99-2	1.35	1.085	1.053
2-nitrobenzaldehyde	552-89-6	0.17	0.124	0.498
5-methyl-2-nitrophenol	700-38-9	0.59	0.319	0.515
methyl-4-nitrobenzoate	619-50-1	0.39	0.366	0.706
pentafluorobenzyl alcohol	440-60-8	-0.2	-outlier-	-outlier-
3-hydroxy-4-nitrobenzaldehyde	704-13-2	0.27	0.233	0.633
2.5-dibromonitrobenzene	3460-18-2	1.37	0.911	1.213
4.5-difluoro-2-nitroaniline	78056-39-0	0.75	0.532	0.775
2.4-dibromo-6-nitroaniline	827-23-6	1.62	0.800	1.725

[a]Experimental values (cocentration in mmol/L) taken from [12]

There are several potential reasons for a chemical to be an outlier from a QSAR. Usually, such compounds have been recognized as acting by a different mechanism of action from the other chemicals, which are well modeled by the QSAR. Examples of outliers from toxicological QSARs abound for all endpoints and have actually been extremely useful in their development. In the 1980s and more recently, the analysis of outliers proved to be the spur for the further analysis and identification of mechanisms of action [61].

A closer analysis of the outlier compounds showed that two compounds, 335 and 354 (benzyl-4-hydroxyphenyl ketone and 4-bromophenyl-3-pyridyl ketone, respectively), were detected as outliers for both models (Eqs. **1** and **3**); adittionally, compound 306 (2,5-diphenyl-1,4-

benzoquinone) was detected as outlier by the Eq. **1**; all three compounds belong to cluster number nine. That is a logical result, because this cluster is composed of only these three compounds, so the structures of these three compounds are markedly different from the rest of the structures in the whole data set. Taking this into account, we can expect an outlier behavior for these compounds, as was shown in the development of the models. For that reason these compounds were included only in the training set. On the other hand, compounds 020, 074, 182 and 215 (4-ethylbiphenyl, 4-hexylresorcinol, phenyl isothiocyanate and 4-chloro-3,5-dinitrobenzaldehyde) were also detected as outliers in previous reports [2, 12, 42]. Other outliers without any apparent structural pattern were detected (see Table **1**).

3.4. Comparison with other approaches

In this subsection, we proceed to develop a comparison between the ability of non-stochastic and stochastic atom-based quadratic indices for the prediction of aquatic toxicity of benzene derivatives against *T. pyriformis*. In a recent publication [42], we developed several QSAR models using five kinds of bidimentional (2D) descriptors, implemented in the Dragon software [62]; these descriptors were: Topological, BCUT, Gálvez´s topological charge indices, 2D Autocorrelations and Molecular Walk Counts. The corresponding models were developed with the same data set as was used in the development of the former models, obtained with non-stochastic and stochastic atom-based quadratic indices (Eqs. **1** and **3,** respectively). Aditionally, we compare the models obtained here with those previously obtained with atom-based linear indices [42]. The statistical parameters of the previously obtained models are shown in Table **4**.

The comparison was based mainly on the quality of the statistical parameters of the regression. Specifically, the results of the present approach (atom-based non stochastic quadratic indices) showed the highest square correlation coefficient value of 0.745 with stochastic quadratic indices, while the model obtained with non-stochastic quadratic indices achieved a value of R^2 of 0.730. These results are similar-to-better than those achieved with stochastic (R^2=0.733) and non-stochastic (R^2=0.721) linear indices to predict aquatic toxity of benzene derivatives. The achieved values of R^2, for the QSAR models developed with Dragon´s 2D molecular descriptors, were between 0.516 and 0.716; the model obtained with molecular walk count descriptors was not considered in the comparison because of the poor shown behavior. Similar behaviour was achieved in the values of standard deviation, *s*=0.385 and *s*=0.396, for stochastic and non-stochastic quadratic indices´ models, correspondingly. The values of standard deviation, for the reported models with the 2D Dragons´ MDs, were between 0.406 and 0.530.

On the oder hand, the models were also compared according to their result in the LOO cross-validation procedure. In particular, the atom-based quadratic models achieved the best values of press statistics, q^2 and s_{cv}. As it can be seen, our models have statistical parameters better than the models obtained with Dragon's molecular descriptors. The model obtained with stochastic quadratic indices showed the highest value of q^2=0.712 and the lowest value of s_{cv}=0.405; the model obtained with non-stochastic quadratic indices had a similar behavior: q^2=0.697 and s_{cv}=0.415. These results are quite similar to the ones achieved with stochastic (q^2=0.704 and s_{cv}=0.411) and non-stochastic (q^2=0.687 and s_{cv}=0.425) linear indices. The values of these statistical parameters for the other models are for q^2 between 0.682 and 0.478, and for s_{cv} between 0.423 and 0.545. All these results are summarized in Table **3**, where a detailed comparison can be more easily performed. Finally, we can say that, for the entire data set the model developed with stochastic indices achieved results slightly better than the model developed with non-stochastic indices, as well as that the models obtained with quadratic indices were also rather better than the one previously obtained with linear indices, correspondingly. In addition, the models obtained with atom-based quadratic and linear indices were superior to those developed with 2D Dragon´s MDs to describe the aquatic toxicity.

Table 4. Comparison between the QSAR models obtained using atom-based quadratic indices with other approaches previously reported [42] to predict aquatic toxicity.

index	*N*	R^2	*s*	F	q^2	s_{cv}
Non-Stochastic Quadratic Indices	313	0.730	0.396	118.04	0.697	0.415
Stochastic Quadratic Indices	313	0.745	0.385	127.43	0.712	0.405
Non-Stochastic Linear Indices[a]	313	0.721	0.403	131.79	0.687	0.421
Stochastic Linear Indices[a]	313	0.733	0.394	139.94	0.704	0.411
2D autocorrelations[a]	313	0.609	0.476	79.54	0.585	0.486
BCUT[a]	313	0.690	0.424	113.56	0.675	0.431
Gálvez topological charge indices[a]	313	0.516	0.530	54.30	0.478	0.545
Topological descriptors[a]	313	0.716	0.406	128.70	0.682	0.423

[a]QSAR Model reported in a previous work [42].

4. Conclusions

In recent publications, it has been recognized the growing necessity of developing more reliable QSAR/QSTR models to assess drug discovery and chemical environmental risk [17, 63, 64]. Therefore, it is necessary the continuos development of predictive regression/classification-based models, in order to predict aquatic toxicity by means of QSAR. Consequently, we have developed fairly good MLR models that could permit us to predict, by fast *"in silico"* screening, the aquatic toxicity of benzenes against *T. pyriformis*.

In the current study, the use of non-hierarchical cluster analysis permits us to split carefully the data into training and validation sets, guaranteeing enough molecular diversity in

each subset. The obtained models, with non-stochastic and stochastic atom-based quadratic indices, were statistically significant and robust in terms of the R^2, s, q^2 and s_{cv} values. The best model was developed with stochastic quadratic indices; it showed good values of $R^2 = 0.806$ and $q^2 = 0.791$. In the impairment of the population growth of *T. pyriformis* with our two models, the capability of predicting the aquatic toxicity of benzene derivatives was assessed by the good values of predictive R^2_{pred} (0.745 and 0.742 for non-stochastic and stochastic model, respectively), achieved for the test set. The results achieved with the stochastic model showed results slightly better than the ones with the non-stochastic model, but both models can be efficiently used to predict the aquatic toxicity of benzene derivatives. The comparison with other approaches, previously reported [42], assesses a good behavior of our method.

Finally, those models obtained in the current work are not ideal because the data set used here, although of good quality and reliable, is limited. Therefore, based on increasing data the learning/modeling will need to be an ongoing, iterative process in which the models will be continuously refined. However, the method proposed here (atom-based quadratic indices) could be a substitute for costly and time-consuming experiments to determine toxicity.

Acknowledgements

Castillo-Garit, J.A. and M-P, Y.; thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University in 2008.We sincerely thank Dr. T. W. Schultz for providing some manuscript reprints from his works, which significantly contribute to the development of this report.

References

1. Green, S., A. Goldberg, and J. Zurlo, *TestSmart-High Production Volume Chemicals: An Approach to Implementing Alternatives into Regulatory Toxicology.* Toxicol. Sci., 2001. **63**(1): p. 6-14.
2. Schultz, T.W., *Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis.* Chem Res Toxicol, 1999. **12**(12): p. 1262-7.
3. Cronin, M.T., B.W. Gregory, and T.W. Schultz, *Quantitative structure-activity analyses of nitrobenzene toxicity to Tetrahymena pyriformis.* Chem. Res. Toxicol., 1998. **11**(8): p. 902-8.
4. Gagliardi, S.R. and T.W. Schultz, *Regression comparisons of aquatic toxicity of benzene derivatives: Tetrahymena pyriformis and Rana japonica.* Bull Environ Contam Toxicol, 2005. **74**(2): p. 256-62.
5. Netzeva, T.I. and T.W. Schultz, *QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data.* Chemosphere, 2005. **61**(11): p. 1632-1643.
6. DeWeese, A.D. and T.W. Schultz, *Structure-activity relationships for aquatic toxicity to Tetrahymena: halogen-substituted aliphatic esters.* Environ Toxicol, 2001. **16**(1): p. 54-60.

7. Auer, C.M., J.V. Nabholz, and K.P. Baetcke, *Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5.* Environ. Health Perspect., 1990. **87**: p. 183-197.
8. Bradbury, S.P., *Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research.* Toxicol. Lett., 1995. **79**(1-3): p. 229-237.
9. McKinney, J.D., et al., *The practice of structure activity relationships (SAR) in toxicology.* Toxicol. Sci., 2000. **56**(1): p. 8-17.
10. Schultz, T.W., et al., *Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective.* J. Mol. Struct. (THEOCHEM), 2003. **622**(1): p. 1-22.
11. Schultz, T.W. and M.T. Cronin, *Essential and desirable characterisctics of ecotoxicity quantitative structure-activity relationships.* Environ Toxicol Chem, 2003. **22**(3): p. 599-607.
12. 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.
13. Schultz, T.W., *TERATOX: Tetrahymena pyriformis population grow impairment endpoint- A surrogate for fish lethality.* Toxicol. Methods, 1997. **7**: p. 289-309.
14. 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.
15. Chen, D., et al., *Holographic QSAR of selected esters.* Chemosphere, 2004. **57**(11): p. 1739-45.
16. Cronin, M.T., et al., *Assessment and modeling of the toxicity of organic chemicals to Chlorella vulgaris: development of a novel database.* Chem. Res. Toxicol., 2004. **17**(4): p. 545-54.
17. Gonzalez, M.P., et al., *A novel approach to predict a toxicological property of aromatic compounds in the Tetrahymena pyriformis.* Bioorg. Med. Chem., 2004. **12**(4): p. 735-44.
18. Schultz, T.W., et al., *Population growth impairment of aliphatic alcohols to Tetrahymena.* Environ Toxicol, 2004. **19**(1): p. 1-10.
19. Aptula, A.O., et al., *Chemistry-toxicity relationships for the effects of di- and trihydroxybenzenes to Tetrahymena pyriformis.* Chem Res Toxicol, 2005. **18**(5): p. 844-54.
20. Cheng, Y.Y. and H. Yuan, *Quantitative study of electrostatic and steric effects on physicochemical property and biological activity.* J Mol Graphics Model, 2005.
21. Spycher, S., E. Pellegrini, and J. Gasteiger, *Use of structure descriptors to discriminate between modes of toxic action of phenols.* J Chem Inf Model, 2005. **45**(1): p. 200-8.
22. Costescu, A. and M. Diudea, V,, *QSTR Study on Aquatic Toxicity Against Poecilia reticulata and Tetrahymena pyriformis Using Topological Indices.* Internet Electron. J. Mol. Des., 2006. **5**: p. 116-134.
23. Zvinavashe, E., et al., *Quantum chemistry based quantitative structure-activity relationships for modeling the (sub)acute toxicity of substituted mononitrobenzenes in aquatic systems.* Environ Toxicol Chem, 2006. **25**(9): p. 2313-21.
24. Ivanciuc, O., *Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against Pimephales promelas and Tetrahymena pyriformis.* Internet Electron. J. Mol. Des., 2004. **3**: p. 802-821.

25. Ivanciuc, O., *Applications of Support Vector Machines in Chemistry*, in *Rev. Comput. Chem*, K.B. Lipkowitz and T.R. Cundari, Editors. 2007, Whiley-VCH, : Weinheim.
26. Melagraki, G., et al., *Prediction of toxicity using a novel RBF neural network training methodology.* Journal of Molecular Modeling, 2006. **12**(3): p. 297-305.
27. Marrero-Ponce, Y., *Total and Local Quadratic Indices of the Molecular Pseudograph´s Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds.* Molecules, 2003. **8**: p. 687-726.
28. Marrero-Ponce, Y., *Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors.* J. Chem. Inf. Comput. Sci., 2004. **44**(6): p. 2010-2026.
29. Marrero-Ponce, Y., et al., *Novel 2D TOMOCOMD-CARDD Descriptors: Atom-based Stochastic and non-Stochastic Bilinear Indices and their QSPR Applications.* J. Math. Chem., 2008: p. DOI 10.1007/s10910-008-9389-0.
30. Marrero-Ponce, Y., et al., *Bond-based 2D TOMOCOMD-CARDD approach for drug discovery: aiding decision-making in 'in silico' selection of new lead tyrosinase inhibitors.* J. Comput.-Aided Mol. Design, 2007. **21**(4): p. 167-188.
31. Casanola-Martin, G.M., et al., *TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices.* Bioorg Med Chem, 2007. **15**(3): p. 1483-503.
32. Marrero-Ponce, Y., et al., *Bond-based global and local (bond, group and bond-type) quadratic indices and their applications to computer-aided molecular design. 1. QSPR studies of diverse sets of organic chemicals.* J Comput-Aided Mol Design, 2006. **20**(10-11): p. 685-701.
33. Casañola-Martin, G.M., et al., *New tyrosinase inhibitors selected by atomic linear indices-based classification models.* Bioorg Med Chem Lett, 2006. **16**(2): p. 324-30.
34. Marrero Ponce, Y., et al., *Predicting antitrichomonal acitivity: A computational screening using atom-based bilinear indices and experimental proofs.* Bioorg. Med. Chem., 2006. **14**: p. 6502-6524.
35. Marrero-Ponce, Y., *Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications.* Bioorg. Med. Chem. , 2004. **12**: p. 6351-6369.
36. Marrero-Ponce, Y., et al., *Prediction of Intestinal Epithelial Transport of Drug in (Caco-2) Cell Culture from Molecular Structure using in silico Approaches During Early Drug Discovery.* Internet Electron. J. Mol. Des., 2005. **4** p. 124-150.
37. Marrero-Ponce, Y., A. Huesca-Guillen, and F. Ibarra-Velarde, *Quadratic indices of the ¨molecular pseudograph´s atom adjacency matrix¨ and their stochastic forms: a novel approach for virtual screening and in silico discovery of new lead paramphistomicide drugs-like compounds.* J. Mol. Struct. (Theochem), 2005. **717**: p. 67-79.
38. Marrero-Ponce, Y., et al., *A computer-based approach to the rational discovery of new trichomonacidal drugs by atom-type linear indices.* Curr. Drug Discov. Technol., 2005. **2**(4): p. 245-65.
39. Marrero-Ponce, Y., et al., *Atom, atom-type, and total non-stochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity.* Bioorg. Med. Chem., 2005. **13**(8): p. 2881-2899.
40. Marrero-Ponce, Y., et al., *Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic.* Bioorg. Med. Chem., 2005. **13**(4): p. 1005-1020.

41. Castillo-Garit, J.A., et al., *Estimation of ADME Properties in Drug Discovery: Predicting Caco-2 Cell Permeability Using Atom-Based Stochastic and Non-Stochastic Linear Indices.* J. Pharm. Sci., 2008. **97**: p. 1946-1976.
42. Castillo-Garit, J.A., et al., *A novel approach to predict aquatic toxicity from molecular structure.* Chemosphere, 2008: p. doi:10.1016/j.chemosphere.2008.05.024
43. Marrero-Ponce, Y., et al., *Protein Quadratic Indices of the ¨Macromolecular Pseudograph´s α-Carbon Atom Adjacency Matrix¨. 1. Prediction of Arc Repressor Alanine-mutant´s Stability.* Molecules 2004. **9** p. 1124-1147.
44. Marrero-Ponce, Y., et al., *Protein linear indices of the 'macromolecular pseudograph alpha-carbon atom adjacency matrix' in bioinformatics. Part 1: prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor.* Bioorg. Med. Chem., 2005. **13**(8): p. 3003-3015.
45. Marrero-Ponce, Y., et al., *Nucleic Acid Quadratic Indices of the ¨Macromolecular Graph´s Nucleotides Adjacency Matrix¨. Modeling of Footprints after the Interaction of Paromomycin with the HIV-1 Ψ-RNA Packaging Region.* Int. J. Mol. Sci. , 2004. **5**: p. 276-293.
46. Marrero Ponce, Y., J.A. Castillo Garit, and D. Nodarse, *Linear indices of the 'macromolecular graph's nucleotides adjacency matrix' as a promising approach for bioinformatics studies. Part 1: prediction of paromomycin's affinity constant with HIV-1 psi-RNA packaging region.* Bioorg. Med. Chem., 2005. **13**(10): p. 3397-3404.
47. Marrero-Ponce, Y., et al., *3D-Chiral quadratic indices of the ¨molecular pseudograph´s atom adjacency matrix¨ and their application to central chirality codification: classification of ACE inhibitors and prediction of r-receptor antagonist activities.* Bioorg. Med. Chem. , 2004. **12**: p. 5331-5342.
48. Marrero-Ponce, Y. and J.A. Castillo-Garit, *3D-chiral Atom, Atom-type, and Total Non-stochastic and Stochastic Molecular Linear Indices and their Applications to Central Chirality Codification.* J. Comput.-Aided Mol. Design, 2005. **19**(6): p. 369-83.
49. Castillo-Garit, J.A., Y. Marrero-Ponce, and F. Torrens, *Atom-based 3D-chiral quadratic indices. Part 2: prediction of the corticosteroid-binding globulinbinding affinity of the 31 benchmark steroids data set.* Bioorg. Med. Chem., 2006. **14**(7): p. 2398-2408.
50. Castillo-Garit, J.A., et al., *Atom-based Stochastic and non-Stochastic 3D-Chiral Bilinear Indices and their Applications to Central Chirality Codification.* J. Mol. Graphics Model., 2007. **26**: p. 32-47.
51. Castillo-Garit, J.A., et al., *Bond-Based 3D-Chiral Linear Indices: Theory and QSAR Applications to Central Chirality Codification.* J. Comput. Chem., 2008: p. DOI:10.1002/jcc.20964
52. Marrero-Ponce, Y. and V. Romero, *TOMOCOMD software. TOMOCOMD (TOpological MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version will be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es*. 2002: Central University of Las Villas.
53. Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*. 1988, Englewood Cliffs, NJ: Prentice-Hall.
54. Mc Farland, J.W. and D.J. Gans, *Cluster Significance Analysis*, in *Chemometric Methods in Molecular Design*, H. Waterbeemd, Editor. 1995, VCH Publishers: Winheim, Ger. p. 295–307.
55. Xu, J. and A. Hagler, *Chemoinformatics and Drug Discovery.* Molecules, 2002. **7**: p. 566-700.
56. STATISTICA version 6.0, *StatSoft*. 2001: Tulsa.

57. Wold, S. and L. Erikson, *Chemometric Methods in Molecular Design*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, Editor. 1995, VCH Publishers: Weinheim. p. 309-318.
58. Belsey, D.A., E. Kuh, and R.E. Welsch, *Regression Diagnostics*. 1980, New York: Wiley.
59. Golbraikh, A. and A. Tropsha, *Beware of q2!* J. Mol. Graphics Model., 2002. **20**(4): p. 269-76.
60. Egan, W.J. and S.L. Morgan, *Outlier detection in multivariate analytical chemical data.* Anal. Chem., 1998. **70**: p. 2372-2379.
61. Cronin, M.T.D. and T.W. Schultz, *Pitfalls in QSAR.* J. Mol. Struct. (THEOCHEM), 2003. **622**(1): p. 39-51.
62. Todeschini, R., V. Consonni, and M. Pavan. Dragon Software version 2.1, 2002.
63. Kulkarni, A.S. and A.J. Hopfinger, *Membrane-interaction QSAR analysis: application to the estimation of eye irritation by organic compounds.* Pharm. Res., 1999. **16**(8): p. 1245-1253.
64. Devillers, J., *New trends in (Q)SAR modeling with topological indices.* Curr. Opin. Drug Discov. Dev., 2000. **3**: p. 275-279.