Journal of Computational Methods in Molecular Design, 2016, 6 (2):1-12

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Quantitative Structure-Activity relationship (QSAR) models for predicting Toxicity of Dioxin compounds

Sabitu Babatunde Olasupo[1*], Adamu Uzairu[2] and Balarabe Sagagi[1]

[1]Department of chemistry, Kano University of Science and Technology, Wudil Kano, Nigeria
[2]Department of chemistry, Ahmadu Bello University Zaria, Nigeria
Corresponding Email:olasabit@yahoo.com

ABSTRACT

A quantitative structure activity relationship (QSAR) study was developed in order to model the toxicity of dioxin, specifically toxicity of polyhalogenate /polychlorinated dibenzo-p-dioxins (25PHCDs/ PCDDs).The QSAR model was constructed, using Genetic Function Algorithm (GFA).The Quantum chemical descriptors were computed by density functional theory (DFT) at B3LYP/ 6-31G. Model-1 with highest statistical significance was selected and it has squared correlation coefficient (R^2) =0.971, Cross validated correlation coefficient (Q^2) = 0.961 and external prediction ability R^2_{pred} = 0.885. The accuracy of the developed model was evaluated through cross-validation, an external test set, Y-randomization and applicability domain techniques. The result of the present study is expected to be useful to predict and identify other toxic compound or to synthesis non-toxic dioxins.*

Keywords: QSAR, Dioxins, DFT, GFA, PCDDs.

INTRODUCTION

The prediction of toxicity of environmental pollutants or contaminants from their molecular structure is one of the major aspect of modern toxicology research and the potential toxicity of compounds can equally be assessed on the basis of their physicochemical and biological properties [1].

Dioxins are polyhalogenated/polychlorinated aromatic Compounds formed as by-products of various industrial processes (such as incomplete combustion of organic compounds containing chlorine, medical/municipal wastes incineration etc.) and commonly regarded as highly toxic compounds that are environmental pollutants and persistent organic pollutants POPs [2]. The term “dioxins and dioxin-like compound” commonly refers to polychlorinated dibenzodioxins (PCDDs), poly-chlorinated dibenzofurans (PCDFs) and polychlorinated biphenyls (PCBs) [3].

Dioxins (PCDDs) are characterized by hydrophobicity and lipophilicity, and these characteristics make them to concentrate in adipose and hepatic tissues and can persist in an individual for long periods of time [4-5]. These compounds are persistent Organic pollutants (POPs) that can enter water bodies and accumulate in soil, sediments, biota, humans and food webs, posing significant health threats to humans, animals and the environment [6].

Among the health effects in humans to dioxins (PCDDs) exposures includes, immunotoxicity, developmental and neuron developmental effects, and changes in thyroid and steroid hormones and reproductive function [7].Therefore, investigation on toxicity of PCDDs are of great importance as to understand the risk attached to their exposure to human health, animals and to the environment.

The study of the quantitative relationship between activity/toxicity and molecular structure (QSAR/QSTR) is an important aspect of research in computational chemistry being used in the prediction of toxicity and biological activities of organic compounds [8, 9]. Previous studies have also reported several toxic effects exhibited by the dioxin like compounds [10].

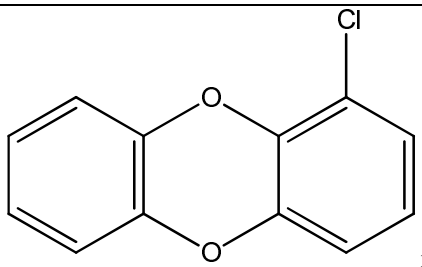
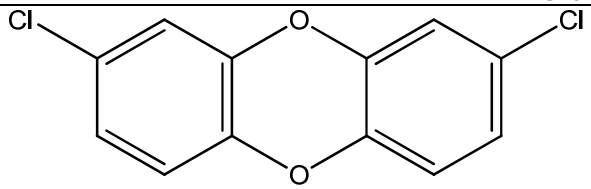
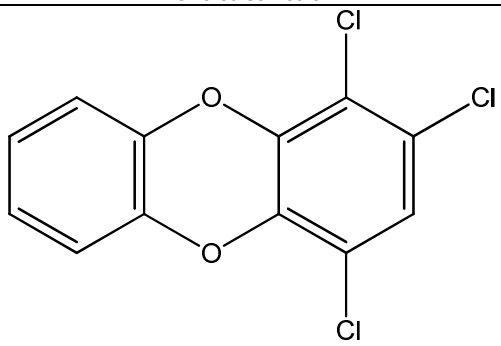
In the present study, the main goal was to build a QSAR model for description and prediction of toxicities of polyhalogenated/ polychlorinated dibenzodioxins (PHCDs/PCCDs) using Genetic Function Approximation (GFA) Algorithm approach. The proposed method was validated using several strategies: Cross validation, Y-randomization, and externally validated using division of the entire data set into training data set and test data sets.

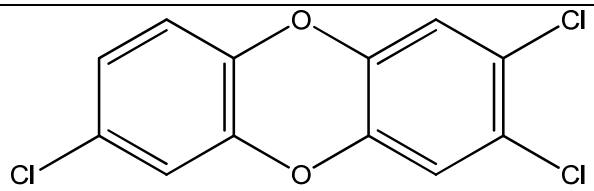
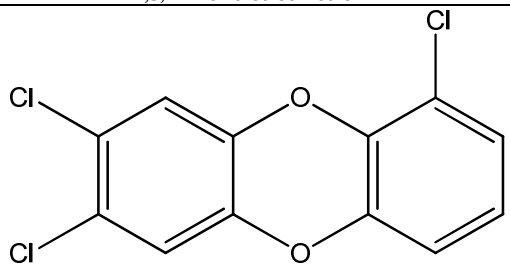
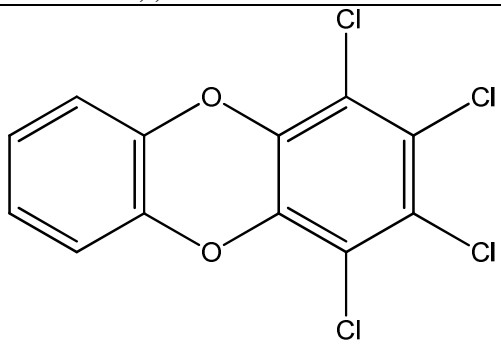
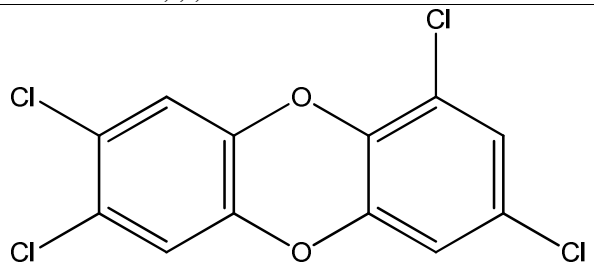
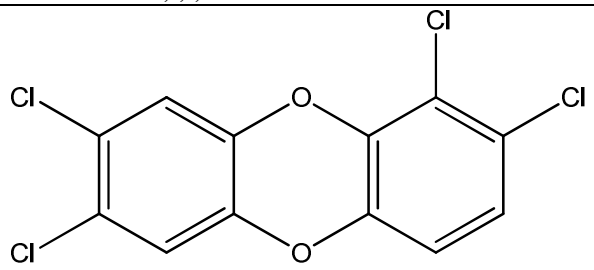
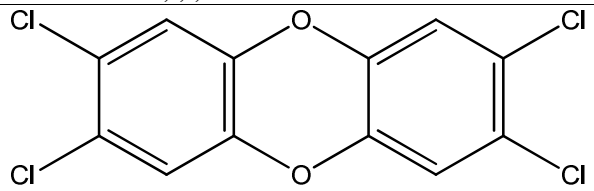
MATERIALS AND METHODS

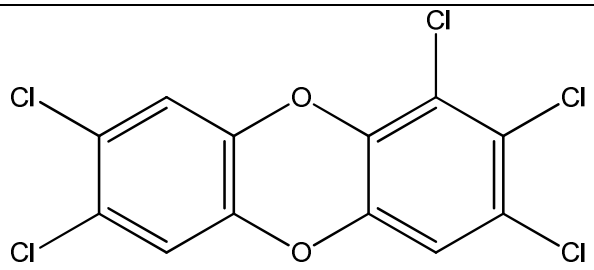
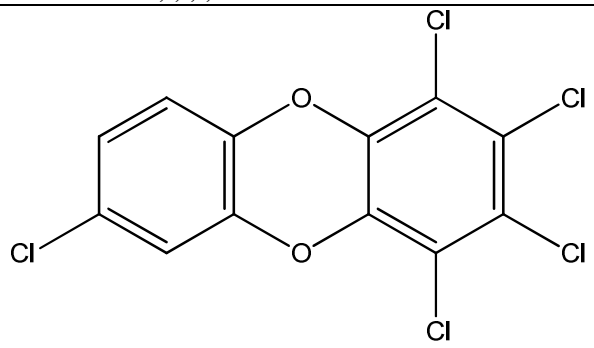
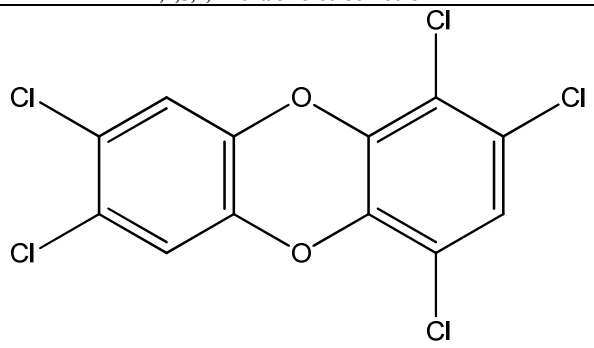
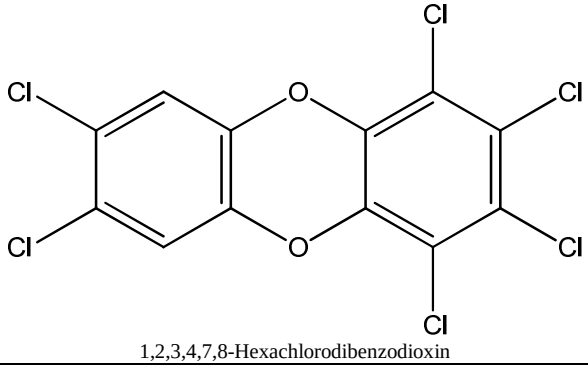
2.1 Data set

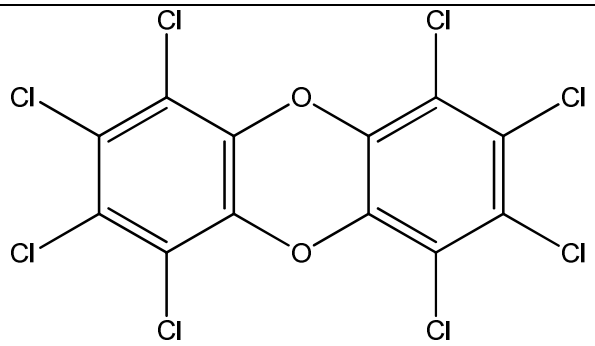
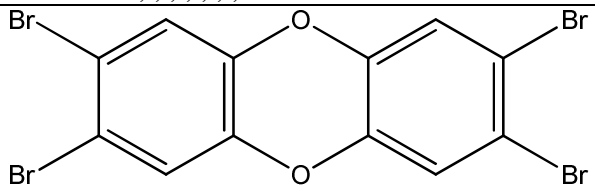
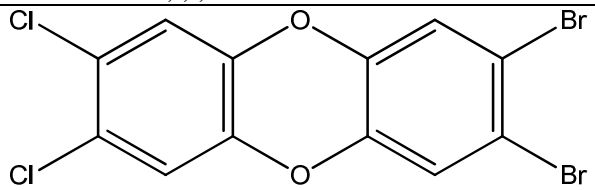
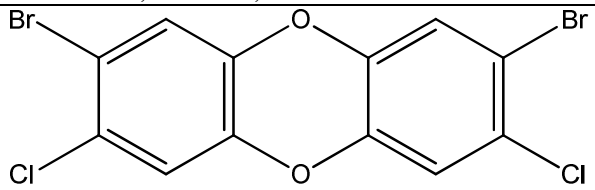
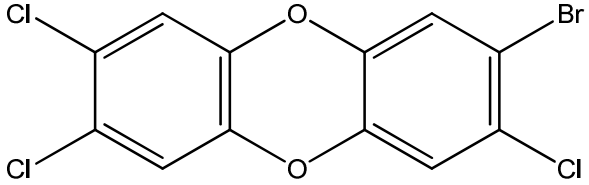
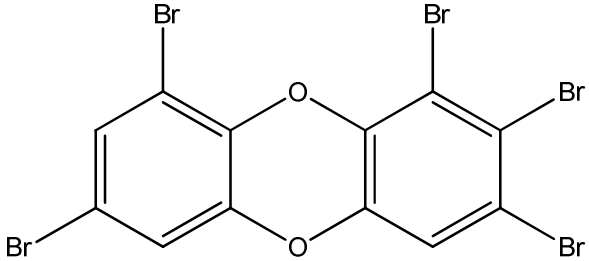
To build a QSAR model for Toxicity of Polychlorinated dibenzodioxins series, a data set of 25 PCDDs compounds were taken from literature [11] with the experimental log ($^1/EC_{50}$) values. The studied compounds with their corresponding experimental log ($^1/EC_{50}$) values is shown in Table1.

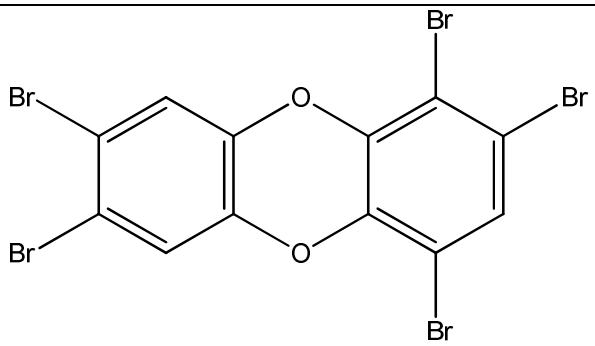
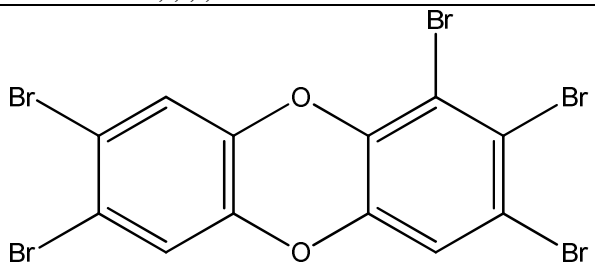
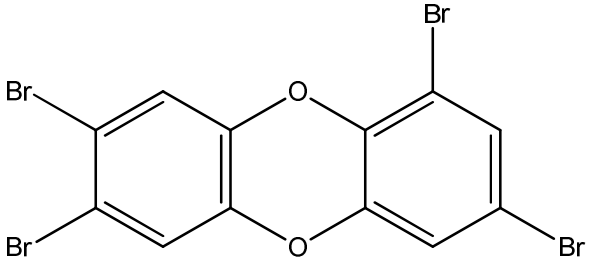
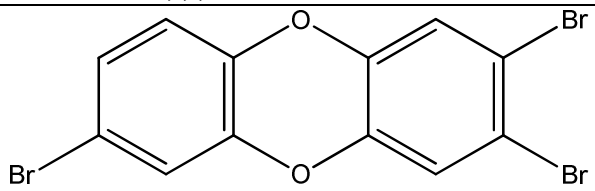
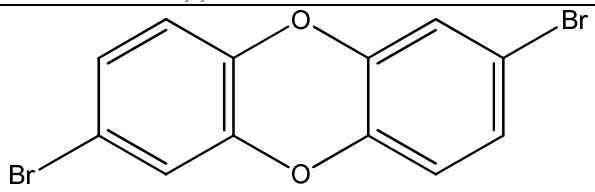
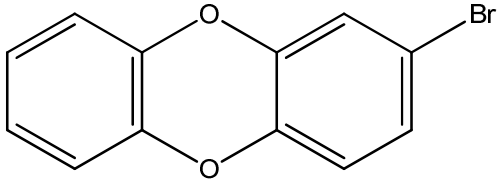
Table1: Chemical structures and experimental log ($^1/EC_{50}$) values for studied compounds

S/N	IUPAC NAME/CHEMICAL STRUCTURE	TOXICITY log(1/EC_{50})
1	 1-ChlorodibenzodioxinS	4.00
2	 2,8-Dichlorodibenzodioxin	5.50
3	 1,2,4-Trichlorodibenzodioxin	4.89

4	 2,3,7-Trichlorodibenzodioxin	7.15
5	 2,3,6-Trichlorodibenzodioxin	6.66
6	 1,2,3,4-Teterachlorodibenzodioxin	5.89
7	 1,3,7,8-Teterachlorodibenzodioxin	6.10
8	 2,3,6,7-Teterachlorodibenzodioxin	6.80
9	 2,3,7,8-Teterachlorodibenzodioxin	8.00

10	 1,2,3,7,8-Pentachlorodibenzodioxin	7.10
11	 1,2,3,4,7-Pentachlorodibenzodioxin	5.19
12	 1,2,4,7,8-Pentachlorodibenzodioxin	5.96
13	 1,2,3,4,7,8-Hexachlorodibenzodioxin	6.55

14	 1,2,3,4,6,7,8,9-Octachlorodibenzodioxin	5.00
15	 2,3,7,8-Tetrabromodibenzodioxin	8.82
16	 2,3-Dibromo-7,8-dichlorodibenzodioxin	8.83
17	 2,8-Dibromo-3,7-dichlorodibenzodioxin	9.35
18	 2-Bromo-3,7,8-Trichlorodibenzodioxin	7.94
19	 1,3,7,8,9-Pentabromodibenzodioxin	7.03

20	 1,2,4,7,8-Pentabromodibenzodioxin	7.77
21	 1,2,3,7,8-Pentabromodibenzodioxin	8.18
22	 1,3,7,8-Tetrabromodibenzodioxin	8.70
23	 2,3,7-Tribromodibenzodioxin	8.93
24	 2,7-Dibromodibenzodioxin	7.81
25	 2-Bromodibenzodioxin	6.53

2.2 Geometry optimization and calculation of molecular descriptors:
Molecular structure of all the compounds were drawn using ChemDraw ultra [12] version 12.02 module of the program and subsequently imported into Wave function program Spartan ''14'' [13] version 1.2.2 for structural minimization. The geometries of all the 25 molecules of PCDDS were optimized with density functional theory (DFT) method at the B3LYP level of theory and 6-31G* as the basis set.

2.21 Model generation:
The structures of 25 Polychlorinated dibenzodioxins (25 PCDDs) were studied by statistical methods based on the Genetic Function Algorithm technique to develop all the models. A peculiar features of genetic function approximation (GFA) algorithm is that it generate a population of equations rather than a single equation as do most other statistical methods.

The range of variation in this population gives added information on the quality of fit and importance of the descriptors [14]. The lack- of- fit (LOF) used here i.e. the fitness function was the leave one-out cross validated correlation coefficient (Q^2_{LOO}) to estimate the quality of the model. This is calculated by

$$LOF = \frac{LSE}{\{[1-[c+d*P]\}2}{M} \quad (1)$$

Where c= number of basic function
d= smoothing parameter
m= number of samples in the training set
LSE = least square error
P = total number of features contained in all basics functions [15]

2.3 Validation of QSAR models.
The predictive ability of the developed QSAR model were evaluated using both internal and external statistical validation parameters for a generally acceptable QSAR model proposed by Ravinchandran et al [16], as presented in Table 2.

Table 2: Validation parameters for an acceptable QSAR model Ravinchandran et al [16]

S/N	Symbol	Name	Range
1	R^2	Coefficient of determination	≥ 0.6
2	Q^2	Gross validation coefficient	> 0.5
3	R^2 pred.	Coefficient of determination for external test set	≥ 0.6
4	R^2 adj	Adjusted square correlation coefficient	> 0.5
5	p (95%)	Confidence interval at 95%	≤ 0.05
6	Next test set	Minimum number of extend test set	≥ 5
7	$R^2 - Q^2$	Difference between R^2 and Q^2	≤ 0.3

Table 3: Training set with their Experimental, Predicted and Residual Toxicity Values in log ($^1/EC_{50}$)

S/N	Chemical Names	Experimental Values	Predicted Values	Residual Values
1	1-ChlorodibenzodioxinS	5.500	5.650	-0.150
2	2,8Dichlorodibenzodioxin	4.890	4.883	0.006
3	1,2,4-Trichlorodibenzodioxin	7.150	7.670	-0.520
4	2,3,7-Trichlorodibenzodioxin	6.660	6.431	0.228
5	2,3,6-Trichlorodibenzodioxin	5.890	5.638	0.251
6	1,2,3,4-Teterachlorodibenzodioxin	6.100	6.042	0.057
7	1,3,7,8-Teterachlorodibenzodioxin	6.800	6.844	-0.044
8	1,2,3,7,8-Pentachlorodibenzodioxin	7.100	6.611	0.489
9	1,2,3,4,7-Pentachlorodibenzodioxin	5.190	5.461	-0.271
10	1,2,3,4,6,7,8,9-Octachlorodibenzodioxin	5.000	5.165	-0.165
11	2,8-Dibromo-3,7-dichlorodibenzodioxin	9.350	9.217	0.133
12	1,3,7,8,9-Pentabromodibenzodioxin	7.030	6.826	0.203
13	1,2,4,7,8-Pentabromodibenzodioxin	7.770	7.832	-0.062
14	1,2,3,7,8-Pentabromodibenzodioxin	8.180	8.404	-0.224
15	1,3,7,8-Tetrabromodibenzodioxin	8.700	8.607	0.093
16	2,3,7-Tribromodibenzodioxin	8.930	8.888	0.042
17	2-Bromodibenzodioxin	6.530	6.596	-0.066

3.0 QSAR Results

A QSAR study was carried out for a set of 25 PCDDs compounds in order to determine a quantitative relationship between structures and toxicity. Dataset Division GUI v 1.2 software was employed to divide the data set of 25 studied compounds into a training set of 17 PCDDs (70%) which was used to build the model and a prediction set (test set) of 8 PCDDs (30%), which was applied to test the built model. The below Table 3 and Table 4 represent the training set and Test set of the studied compounds with their corresponding experimental, predicted and residual values in log ($^1/EC_{50}$) respectively.

From the Table 3, it is obvious that compound numbered-2 ***(2, 8-Dichlorodibenzodioxin)*** is best predicted evidenced by its predicted value to be the closest to the experimental value compared to all other compounds in the set and also with the lowest positive residual value.

Table 4. Test set with their Experimental, Predicted and Residual Toxicity Values in log ($^1/EC_{50}$)

S/N	Chemical Names	Toxicity	Predicted	Residual
1	1-ChlorodibenzodioxinS	4.000	4.528	-0.528
2	2,3,7,8-Teterachlorodibenzodioxin	8.000	8.083	-0.083
3	1,2,4,7,8-Pentachlorodibenzodioxin	5.960	5.529	0.431
4	1,2,3,4,7,8-Hexachlorodibenzodioxin	6.550	5.444	1.106
5	2,3,7,8-Tetrabromodibenzodioxin	8.820	9.662	-0.842
6	2,3-Dibromo-7,8-dichlorodibenzodioxin	8.830	9.154	-0.324
7	2-Bromo-3,7,8-Trichlorodibenzodioxin	7.940	8.715	-0.775
8	2,7-Dibromodibenzodioxin	7.810	8.784	-0.974

As can be seen in the Table 3 and Table 4 above, the predicted values are in good agreement with the corresponding experimental values because of their low residual values.

The GFA analysis generates five (5) models out of which the most statistically significant model (model-1) was selected as presented in Table 5. A brief description of the descriptors in the model is shown in Table 6. The statistical parameters of the model is presented in Table 7.

Table 5: Most statistically significant model; Model-1

S/N	Equation	Definition
1.	Y = 8.018 * X33 - 4.224* X74 + 8.336 * X182 + 8.803 * X186 - 3.808	X33 : ATS7v X74 : AATS3m X182 : ATSC3s X186 : ATSC7s

Table 6. Brief description of the selected descriptors of the best model-1

Descriptor	Regression coefficient	Description	Descriptor Class
ATS7v	8.018	Broto-Moreau autocorrelation - lag 7 / weighted by van der Waals volumes	Autocorrelation Descriptor
AATS3m	- 4.224	Average Broto-Moreau autocorrelation - lag 3 / weighted by mass	Autocorrelation Descriptor
ATSC3s	8.336	Centered Broto-Moreau autocorrelation - lag 3 / weighted by I-state	Autocorrelation Descriptor
ATSC7s	8.803	Centered Broto-Moreau autocorrelation - lag 7 / weighted by I-state	Autocorrelation Descriptor

Table 7. Statistical parameters of the model

Model	R^2	R^2_{adj}	R^2_{pred}	R^2_{cv}	*Friedman LOF*
1	*0.971*	*0.961*	*0.884*	*0.922*	*0.344*

The results in Table 7 are in good agreement with minimum acceptable parameters of a QSAR model as reported in Table 2.

The scattered plot between the experimental and predicted log ($^1/EC_{50}$) of test set (External set validation) and training set (Internal set validation) were presented in Fig 1 and Fig 2. Also, the scattered plot of predicted versus experimental log ($^1/EC_{50}$) for all of the 25 compounds studied (training and test set) is shown in Fig 3.

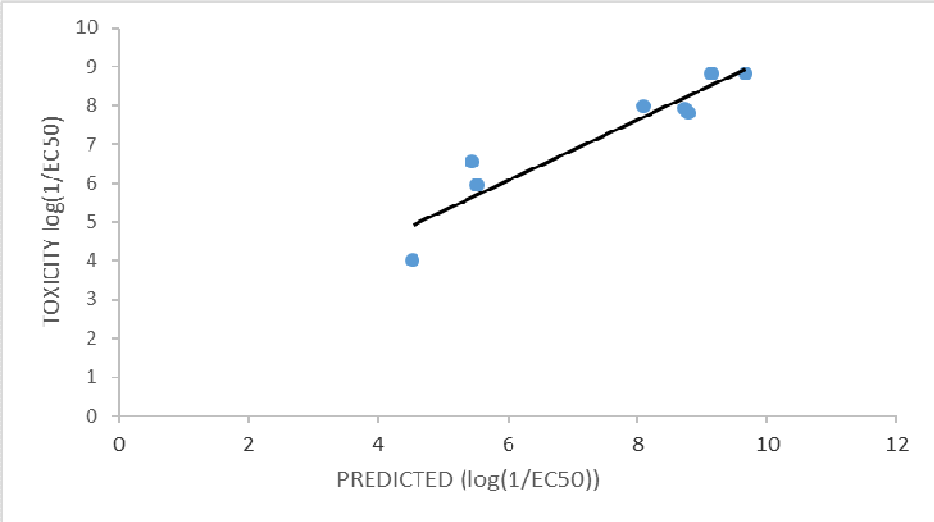


Fig **1: A graphical representation of the model-1 Validation (Test Set)**

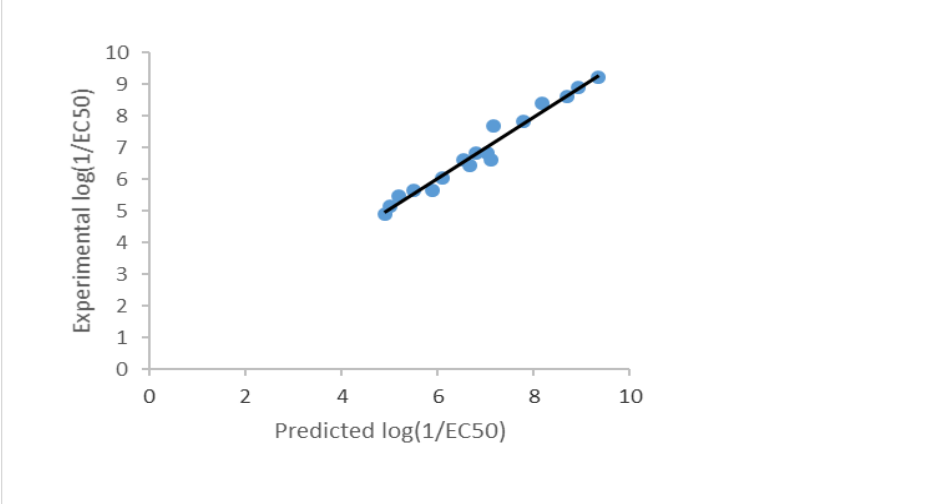


Fig 2: A graphical representation of the model-1 (Training Set)

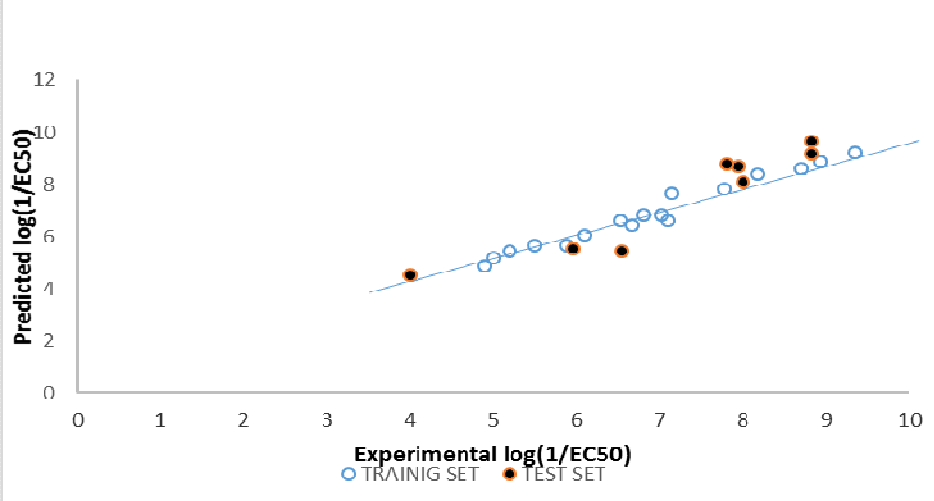


Fig 3: A graphical representation of predicted log ($1/EC_{50}$) (training & test set) versus experimental log (I/EC_{50}) values by GFA modeling.

3.1 Evaluation of the GFA model

3.1.1. **Y-randomization**

The robustness GFA model was tested by applying Y-randomization. The results of Y-randomization test are presented in Table 7. Random models parameters are shown in Table 8.

Table 7: The results of Y-randomization of the Training set

Model	R	R^2	Q^2
Original	0.919	0.846	0.659
Random 1	0.551	0.303	-0.271

Table 8: Random models parameters

Average R :	0.352
Average r^2 :	0.137
Average Q^2 :	-0.556
cRp^2 :	0.782

3.1.1 **Applicability Domain (AD)**

The model was further validated by applying the **Williams plot**, ***the plot of the standardized residuals versus the leverage as shown in Fig 4***. This was exploited to visualize the applicability domain (AD) [17]. Leverage indicates a compound's distance from the centroid of X. The leverage of a compound in the original space is defined as;

$$hi = xi\,(X^{T}X)^{-1}xi$$

Where xi the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values.

The warning leverage (h*) is defined as:

$$h_{*} = 3\frac{(P+1)}{n}$$

Where n = number of training compounds
p = number of predictor variables

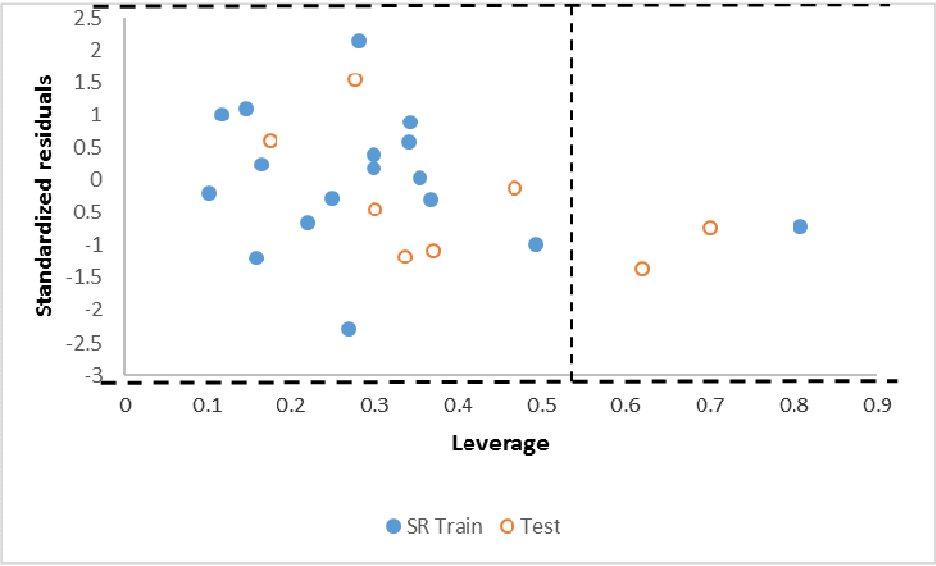


Fig 4: Williams plot of the model

From the Williams plot in (fig 4) above, it is obvious that compound in the test set fall inside the domain of the GFA model (the warning leverage h* =0.60). There are only three compounds (***one in the training set and two in the test set)*** which have the leverage higher than the warning h* value, thus they can be regarded as structural outliers.

DISCUSSION

The model reported in Table – 5 represents most statistically significant model of GFA analysis (R^2=0.971, R^2_{Adj}= 0.961, R^2_{cv} = 0.922, Friedman LOF= 0.344) and are in good agreement with statistical parameters reported in Table 2, this shows the goodness and reliability of the model. A comparison of the predicted toxicities with the experimental log ($^1/_{EC50}$) reported in Table 3 indicated high predictability of the model evidenced by low residual values observed in the Table. The compound No 2 *(2,8 – Dichlorodibenzo dioxin)* is best predicted in the series obvious by its lowest positive residual value. The predicted toxicities of test set in log ($^1/_{EC50}$) in Table – 4 are in good agreement with the experimental values (R^2_{Pred}= 0.884). The robustness of the proposed models and its predictive ability were also guaranteed by the results of Y-randomization test presented in Table 7 and that of random models parameters presented in Table 8. The low R^2 and Q^2 values that were obtained proved the robustness of the model and that the good result in the original model is not due to a chance correlation or structural dependency of the training set. The fact that the cR^2p value of the model is > 0.5 as reported in Table 8 is a good confirmation that the model is robust and promising [18] .

A Comparison of the validation parameters of the best model with the acceptable standard proposed by Ravinchandra et al; reported in Table 2 shows that the parameters are in excellent agreement with the standard as R^2 = 0.971, R^2_{adj} = 0.961, R^2_{pred} = 0.884, Q^2 = 0.961, Friedman LOF = 0.344. This further authenticates the goodness and reliability of the model. The scattered plot of experimental toxicities versus predicted log ($^1/_{EC50}$) is shown in Fig 1 for the test set and Fig 2 for the training set with the $R^2_{Pred=}$ 0.8848 which is in agreement with GFA derived R^2 value, this also confirms the reliability of the model. Also, Fig 3 shows the predicted versus experimental log ($^1/_{EC50}$) for all of the 25 compounds studied, the training set and the test set. As can be seen, the predicted log ($^1/_{EC50}$) values are in good agreement with the experimental log ($^1/_{EC50}$) values.

Furthermore, the Williams plot, the plot of the standardize residuals versus the leverage as shown in Fig 4 was exploited to visualize the applicability domain (AD) [17]. From the plot shown in Fig 4, it is obvious that compound in the test set fall inside the domain of the GFA model (the warning leverage h* =0.60). There are only three compounds (***one in the training set and two in the test set)*** which have the leverage higher than the warning h* value, thus they can be regarded as structural outliers.

3.3.1. ***Descriptors Contribution***

Adrian Beteringhe and Alexandru T.B [19] 2004, reported that Topological charge index of order 3 (GGI_3) and Hydrophilic factor (Hy) are among the descriptors that are responsible for toxicity of Polychlorinated dibenzodioxins. Lakhlifi et al., [20] 2014, reported that Total energy (E_T), energy E_{HOMO}, energy E_{LUMO}, activation energy Ea, the dipole moment N and the factor of oscillation f(so) play an important role for the toxicity of Polychlorinated dibenzodioxins.

Apart from the descriptors reported by (Adrian Beteringhe et al, 2004; Lakhlifi et al; 2014) which are found to be responsible for toxicity of PCDDs, the present study reveals the following four Autocorrelation descriptors contribute more significantly to the toxicities of Polychlorinated dibenzodioxins as indicated by their regression coefficient values. These four descriptors are Broto-Moreau autocorrelation - lag 7 / weighted by van der Waals volumes (ATS7v), Average Broto-Moreau autocorrelation - lag 3 / weighted by mass (AATS3m), Centered Broto-Moreau autocorrelation - lag 3 / weighted by I-state (ATSC3s) and Centered Broto-Moreau autocorrelation - lag 7 / weighted by I-state (ATSC7s) as reported in Table 6.The descriptors ATS7v, ATSC3s and ATSC7s all have a positive regression coefficient; thereby exhibit positive influence on the log ($^1/_{EC50}$) values of the compounds. Thus suggesting that by increasing their values would be favorable to the toxicities of the compounds, this implies that log ($^1/_{EC50}$) is directly related to these descriptors. The descriptor AATS3m, Average Broto-Moreau autocorrelation - lag 3 / weighted by mass displays a negative sign, which indicates that toxicity is inversely related to this descriptor. This means by decreasing the value of this descriptor, it will increase log ($^1/_{EC50}$) value of the compounds. The Model 1 is presented in Table 5 and the descriptors used are shown in table 6. From the above discussion, it is obvious that Atomic van der Waals volumes, atomic mass and I-States in a molecule play an important role in the toxicity of compounds.

CONCLUSION

In this study, the toxicity of 25 PCCDs were successfully modeled by genetic function algorithm using four selected descriptors all belonging to the 2D Autocorrelation descriptors (ATS7v, AATS3m, ATSC3s and ATSC7s) families. These descriptors used relate that Atomic van der Waals volumes, atomic mass and I-States in a molecule proved to be important factors controlling the toxicity of PCDDs. The QSAR model shows good reliability, robustness and high predictability when verified by internal validation (R^2 = 0.971, R^2_{adj} = 0.961, R^2_{pred} = 0.884, Q^2 = 0.961, Friedman LOF = 0.344) and also external validation. The proposed QSAR model can be used in the prediction of the toxicity of congeneric compounds used in this work in order to guide in the identification of not yet known toxic compounds in our laboratory.

Acknowledgements
The authors appreciate the support from Mrs Olasupo Rasheedat Titilayo. We also acknowledge the valuable assistance of Mr David Ebuka Arthur, Samuel Hassan, Philip Ameji and Olaniran Olakunle Daniel.

REFERENCES

[1] SC Basak; GD Grunwald; BD Gute; K Balasubramanions; D Opitz, *J. Chem. Info. Comput. Sci*., **2000**, 40 885.
[2] T Webster; B Commoner, *The dioxin debate*, in: Dioxin and Health, Schecter. A., Ed; Plenum Press, New York, **1994**; 1.
[3] *Safety evaluation of certain food additives and contaminants:* WHO, **2002**, 48.
[4] K Bhupander, *Arch. Appl. Sci. Res*., **2012**, 4, 1906-1914
[5] A Nandan, *IJPSR*, **2013**, 4, 2691-2695.
[6] N Klanova, M Matykiewiczova; P Zdenek; K Prosek; K Laska, *Environmental Pollution*, **2007**, 152, 416–423.
[7] *Code of proactive for the prevention and reduction of dioxin and dioxin-like PCB contamination in food and feeds*. Rome, FAO/WHO (**2006**).
[8] A Rkatrizky; D Rpentrukhin; E Basaka; MK Umaru, *J Chem. Info. Comput. Sci*; **2001**, 40,118.
[9] A Rkatrizky; V Umaran; Y Mkrelson, *Inf. Comput*., **2000**, 41, 345-390.
[10] G W Lucier; GJ Portier; MA Gallo, *Environ. Health Perspect*. **1993,** 101, 36-44.
[11] OG Mekenyan; DJ Veith; GTA Ankley, *Environ Health Perpect.* **1996**, 104, 1302-1310.
[12] CS Chem3D Ultra Cambridge Soft Corporation, Cambridge USA.
[13] Wavefunction Inc., *Spartan 14 v. 1.1.2, Irvine*, California, USA (**2013**).
[14] VK Srivastav; M Tiwari, *Arab. J. Chem.*, **2013**, 7, 1-4.
[15] R Kurial; P Roy; I Mitra, *Molecule*, **2009**, 14, 1660-1701.
[16] V Ravinchandran; H Rajak; A Jain; S Sivadasan; C P Varghese; R Kishore-Agrawal, *Int J. Des. Discov*., **2011**, 2, 511-519.
[17] TI Netzeva; AP Worth; T Aldenberg; R Benigni; MTD Cronin; P Gramatica; JS Jaworska, S.Kahn; G Klopman; CA Marchant; G Myatt; N Nikolova-Jeliazkova; GY Patlewicz; R Perkins; DW Roberts; TW Schultz; DT Stanton; JJM van de Sandt; W Tong; G Veith and CH Yang. ATLA - Alternatives to Laboratory Animals, **2005**, 33:155-173.
[18] P Ambure; RB Aher; A Gajewicz; T Puzyn; K Roy, Modeling. *Chemom. Intell. Lab. Syst.*, **2015**, 147, 1–13.
[19] A Beteringhe, T.B Alexandru, *ARKIVOC*, **2004** (1) 163-182.
[20] L Lakhlifi, *Inter. J. Advanced Res. Comp. Sc. Soft. Eng.* **2014**, 4, 202-213.