Genetic Functional Algorithm Prediction of Toxicity of some Polychlorinated Dioxins using DFT and Semi-empirical Calculated Molecular Descriptors

Hassan Samuel*[1,2], Adamu Uzairu[2], Paul Andrew Mamza[2] and Okunola Oluwale Joshua[3]

[1]Nigerian Institute of Leather and Science Technology, Samaru-Zaria, Nigeria
[2]Department of Chemistry, Faculty of Science, Ahmadu Bello University, Zaria, Nigeria
[3]Department of Applied Chemistry, Faculty of Science, Federal University Dutsin-ma, Katsina State, Nigeria
The correspondence should be addressed to Hassan Samuel; hassansamuel84@gmail.com

Abstract

A set of twenty five compounds of polyhalogenated dioxins with toxicity data in EC50 was subjected to quantitative structure activity relationship studies using Material Studio software 7.0. Large number of molecular descriptors was calculated from the level of theory DFT (BLYP/6-31G*) and semi-empirical (AM1) using the softwares Spartan 14v1.1.2 and PaDel descriptor. The correlation between the toxicities and the DFT and semi-empirical calculated descriptors was examined. Genetic Function Approximation (GFA) technique was used to generate ten QSAR models for each of the two level of theory, out of these models the one with the highest statistical significance was selected as the best for the two methods. DFT (R^2 = 0.9516, R^2_{adj} = 0.9389, R^2_{cv} = 0.9091, LOF = 0.5882, significance of regression F-value = 74.8019) and Semi-empirical (R^2 = 0.96803, R^2_{adj} = 0.9596, R^2_{cv} = 0.9518, LOF = 0.3877, significance of regression F-value = 115.0703). These descriptors were found to be responsible for the toxicities of polyhalogenated dioxins. DFT (BCUTc-1h, VP-3, SssssGe, ETA_dAlpha_B and ETA_BetaP) and semi-empirical (EHOMO, SP-7, ETA_Shape_P, ETA_EtaP_L and GRAV-4). From the comparison of the models generated using DFT and semi-empirical and based on their statistical parameters, semi-empirical (AM1) has slightly better predictive power than DFT (BLYP/6-31G*).

Introduction

Several studies have shown that since the middle of the 20th century there has been an increasing concern about the exposure of humans and wildlife to certain xenobiotic that were released into the environment due to diverse anthropogenic activities.[1] One group of environmental toxicants that is of particular interest relative to potential environmental health effects is dioxin-like chemicals (DLCs).These ubiquitous compounds are hydrophobic, lipophilic and resistant to biological and chemical degradation, properties that impart persistency and a propensity to bio-accumulate and biomagnify to concentrations that can cause harmful effects. DLCs include polychlorinated dibenzo-p-dioxins and dibenzo furans (PCDD/Fs), dioxin-like polychlorinated biphenyls (DL-PCBs), polycyclic aromatic hydrocarbons (PAHs), as well as a multitude of other partially known and unknown compounds [2-7]. Dioxins and dioxin-like compounds are persistent organic pollutants (POPs) that can enter water bodies and eventually sink into the sediment through various transportation routes [8]Wildlife and people are constantly exposed to dioxins through ingestion of dioxins that are present at low levels as environmental contaminants in food. Although they are at low levels in food, some dioxins are very slowly removed from the body and therefore they accumulate in fat tissue. In laboratory animals, dioxins are highly toxic, cause cancer, and alter reproductive, developmental and immune function. The in vivo behavior of these compounds depends on their uptake, distribution and metabolism as well as modifying factors such as species, age and reproductive status [9-10]Food is the major source for human exposure to and dioxins and dioxin-like compounds, especially fatty foods: dairy products (butter, cheese, fatty milk), meat, egg, and fish. Food of animal origin accounts for 95 % of total exposure. The current average body burden of dioxins is about 5–50 ng/kg (as WHO TEq in fat; pg/g = ng/kg) or 100–1000 ng (WHO-TEq) per person which is close to the lowest concentrations possibly causing health effects. Some subgroups within the society (e.g., nursing babies and people consuming plenty of fish) may be exposed to higher than average amounts of these compounds and are thus at greater risk. Dioxin concentrations have been screened in five WHO international studies, and in Central Europe the concentrations have decreased in breast milk from about 40 ng/kg (as TEq in milk fat) in 1987 to below 10 ng/kg in 2006. PCBs have decreased at about the same rate. The decrease in environmental concentrations is due to cessation of PCB use and improved incineration technology [11].

The way in which dioxin affects cells is similar in some way to the way in which hormones such as estrogen work. Dioxin enters a cell and binds to a protein present in cells known as the Ah receptor. The receptor when bound to dioxin can then bind to DNA and alter the expression of some genes. This can lead to alterations in the level of specific proteins and enzymes in the cell. While it is not known exactly how changes in the levels of these different proteins cause the toxicity of dioxin, it is believed by most scientists that the initial binding of dioxin the Ah receptor is the first step. [12-14]. Binding of these dioxin-like compounds can cause a great diversity of biological effects including hepatotoxicity, endocrine effects, immunotoxicity, body weight loss, teratogenicity, carcinogenicity and the induction of diverse enzymes such a aryl hydrocarbon hydroxylase (AHH) and 7-ethoxyresorufinOdeethylase (EROD) in various organisms [15-16]. Due to the problems of assessing the fate and toxicity of large number of chemicals, alternative method has been sought to classical in vivo animal texting. In the area of computer – aided toxicity prediction, quantitative structure activity relationship (QSAR) have been seen as an attractive method for toxicity and fate assessment [17].The study of the quantitative relationship between toxicity/activity and molecular structure (QSTR/QSAR) is an important area of research in computational chemistry and has been widely used in the prediction of toxicity and other biological activities of organic compounds [18-19].

In this study, genetic function approximation (GFA) which is a statistical modeling algorithm that builds functional models of experimental data. Since its inception, several applications of this algorithm in the area of quantitative structure–activity relationship modeling have been reported [20-19]. The genetic function approximation (G FA) algorithm is a genetic algorithm (GA) [21] derived from the previously reported G/SPLINES algorithm [22] and has been recently applied to the generation of QSAR models [23]. The main purpose of this work is to find out how accurate QSAR analysis (using Material studio 7.0 software and the statistical tool Genetic functional algorithm) predicted the toxicity of polychlorinated aromatic compounds, and also to find out the descriptors responsible for producing such toxicity other than the once reported by [24-25].

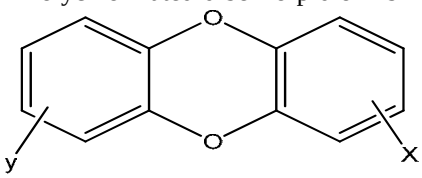
Density functional theory (DFT) is a computational quantum mechanical modelling method used in physics, chemistry and materials science to investigate the electronic structure (principally the ground state) of many body systems, in particular atoms, molecules, and the condensed phases. Using this theory, the properties of a many-electron system can be determined by using functions, i.e. functions of another function, which in this case is the spatially dependent electron_density. Hence the name density functional theory comes from the use of functions of the electron density. DFT is among the most popular and versatile methods available in condensed-matter physics, computational physics, and computational chemistry [26]. In contrast, the semi-empirical Austin Method 1 (AM1) deals only with the valence electrons, thus significantly reducing the complexity and hence time of one of the most computationally expensive steps [27]. The aim of this research is to find how accurate QSAR analysis (using Material studio 7.0 software and the statistical tool Genetic Function Approximation) can predict the toxicities of polyhalogenated dioxins and to compare the predictive power of the models generated using DFT calculated molecular descriptors and the once generated using semi-empirical calculated descriptors.

QSAR METHODOLOGY

Data Set Biological Toxicity

A data set of 25 molecules of polychlorinated dibenzo-p-dioxin and polybrominated dibenzo-p-dioxin was taken from the literature. These molecules were randomly divided into training set of 17 molecules and test set of 8 molecules. The general structure of all the compounds used for QSAR analysis and their experimental biological toxicity (EC50-Effective Concentration at 50%) are given in Table-1. The toxicities of the compounds were converted to negative logarithmic scale (pEC_{50}) to achieve normal distribution.

Table 1: Structures and biological activities of training set compounds.

S/N	Polychlorinated-dibenzo-p-dioxins  X=Cl or Br and y=Cl or Br	Toxicity(pEC_{50})
1	2,3,7,8-tetrachlorodibenzo-p-dioxin	8.00
2	1,2,3,7,8-pentachlorodibenzo-p-dioxin	7.10
3	2,3,6,7-tetrachlorodibenzo-p-dioxin	6.80
4	2,3,6-trichlorodibenzo-p-dioxin	6.66
5	1,2,3,4,7,8-hexachlorodibenzop-dioxin	6.55
6	1,3,7,8-tetrachlorodibenzo-p-dioxin	6.10
7	1,2,4,7,8-pentachlorodibenzo--dioxin	5.96
8	1,2,3,4-tetrachlorodibenzo-p-dioxin	5.89
9	2,3,7-tetrachlorodibenzo-p-dioxin	7.15
10	1,2,3,4,7-pentachlorodibenzo-p-dioxin	5.19
11	1,2,4-Trichlorodibenzo-p-dioxin	4.89
12	2,8-dichlorodibenzo-p-dioxine	5.49
13	1,2,3,4,6,7,8,9-Octachlorodibenzo-o-dioxin	5.00
14	1-chlorodibenzo-p-dioxin	4.00
15	2,3,7,8-tetrabromodibenzo-p-dioxin	8.82
16	2,3-dibromo-7,8-chlorodibenzo-p-dioxin	8.83
17	2,8- dibromo-3,7-dichlorodibenzo-p-dioxin	9.35
18	2-Bromo-3,7,8-trichlorodibenzo-p-dioxin	7.94
19	1,3,7,9-tetrabromodibenzo-p-dioxin	7.03
20	1,3,7,8-tetrabromodibenzo-p-dioxin	8.70
21	1,2,4,7,8-pentabromodibenzo-p-dioxin	7.77
22	1,2,3,7,8-pentabromodibenzo-p-dioxin	8.18
23	2,3,7-tribromodibenzo-p-dioxin	8.93
24	2,7-dibromodibenzo-p-dioxin	7.81
25	2-Bromodibenzo-p-dioxin	6.53

Geometry optimization and Computation of molecular descriptors

Molecular descriptor can be defined as the essential information of a molecule in terms of its physicochemical properties such as constitutional, electronic, geometrical, hydrophobic, lipophilicity, solubility, steric, quantum chemical, and topological descriptors [28]. The structures of all the 25 molecules were optimized by Density Functional Theory (BLYP/6-31G*) and Semi-empirical (AM1) level of theory using Spartan 14v1.1.2 software. Using the optimized structures, Spartan 14v1.1.2 and PaDel descriptor software were used to compute about 1700 different types of molecular descriptors.

Elimination and Selection of molecular descriptors

Prior to any statistical modeling, a preprocessing step was employed to discard descriptors having the same values for all the molecules. After preprocessing, 332 molecular descriptors remained. Since this number of the remaining descriptors was still very much larger than the training compounds [29], a feature selection procedure

was employed to select those descriptors that have high level of correlation with the pEC_{50} toxicity. This correlation analysis was performed using Multiple Linear Regression.

Generation of QSAR Models

The selected descriptors were used to construct 10 models from DFT and Semi-empirical techniques using the well-known statistical algorithm, Genetic Function Algorithm (GFA) which was recently used in QSAR study [30]. A distinctive feature of GFA is that it produces a population of several models instead of generating a single model, as do most other statistical methods. Genetic algorithm makes superior models to those developed using stepwise regression techniques because it selects the basis function genetically. The main use of genetic algorithms in QSAR targets variable selection and model identification [31-33]. The genetic algorithm handles the selection, while the model paradigm generates the evaluation function

Model Validation and statistical parameters

The best models from optimization by DFT and by SE were selected on the basis of various statistical parameters such as squared correlation coefficient (R^2), cross-validation squared coefficient (R^2cv), adjusted squared correlation coefficient (R^2_{adj}), F-test value Lack-Of-Fit (LOF) and Standard error of estimate (SE) and Variance Inflation Factor (VIF. The predictive capacity of the model, was tested using internal validation techniques for both DFT and SE. Univariate analysis was performed to check the symmetry distribution of toxicity data. Skewness, kurtosis and other parameters were calculated as shown in Table-

Coefficient of multiple determination (R^2)

To assess the goodness-of-fit, the coefficient of multiple determination is used. R^2 estimates the proportion of the variation in the response that is explained by the predictor.

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})} \quad (1)$$

Where yi is the observed dependent variable, $\bar{y}$ the mean value of the dependent variable and $\hat{y}$ the calculated dependent variable. If there is no linear relationship between the dependent variable and the descriptors then R^2 = 0.00; if there is a perfect fit then R^2 = 1.00. R^2 values higher than 0.5 indicates that the explained variance by the model is higher the unexplained one.

Internal validation- R^2cv

Cross-validation square correlation coefficient R^2 (LOO-Q^2) is calculated according to the formula:

$$Q^2 = 1 - \frac{\sum(Y_{pred} - Y)^2}{\sum(Y - \bar{Y})^2} \ldots \quad (2)$$

Adjusted R^2 (R^2_{adj})

The value of R2 can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. It follows R^2 should be used be caution. This can be avoided by using another statistical parameter the so-called adjusted R^2 (R^2_{adj}).

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{I-1}{I-K}\right) \quad (3)$$

R^2_{adj} is interpreted similarly to the R^2 value, except that it takes into consideration the number of degrees of freedom. The value of R^2_{adj} decreases if an added variable to the equation does not reduce the unexplained variable.

Standard error of estimate (SEE)

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{(I-(K+1))}} \quad (5)$$

The smaller the value of SEE is, the higher the reliability of the prediction. However, it is not recommended to have the standard error of estimate smaller than the experimental error of the biological data, because it is an indication of over fitted model.

F-value

The F-value is determined using equation-6

$$F = \frac{\sum_{i=0}^{I}(y_i - \bar{y})^2/(K-1)}{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2/(I-K)} \quad (6)$$

The higher the F-value, the greater the probability that the equation is significant [34].

Lack of fit (LOF)

A “fitness function” or lack of fit (LOF) was used to estimate the quality of the model, so that best model receives the best fitness score. The error measurement term is determine by equation-1

$$LOF = \frac{LSE}{(1-\frac{c+d*p}{M})^2} \quad (7)$$

where 'c' is the number of basic functions (other than constant term); 'd' is smoothing parameter (adjustable by the user); 'M' is the number of samples in the training set; LSE is least squares error and 'p' is the total numbers of the features contained in all basis functions [35].

Variance Inflation Factor (VIF)

The multi-collinearity between the above five descriptors was detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$VIF = \frac{1}{1-R^2} \quad (8)$$

Where R^2 is the correlation coefficient of the multiple regression between the variables within the models. If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [36-37].

Y-Randomization Test.

The statistical significance of the relationship between the toxicity of polyhalogeneted compounds and chemical structure descriptors was further demonstrated by randomization procedure. Y-randomization is the most popular and probably the most powerful technique for validation of a given QSAR model [38-39]. In this approach, dependent variable vector (the toxicities in this study) is randomly shuffled and a new QSAR model is built using the independent variables. The procedure is repeated number of times. If the new QSAR model has lower R^2 and R^2_{cv} values for several trials (100 times in this study), then the given QSAR model is thought to be robust. Therefore, Y-randomization is useful to avoid any chance-comer correlation between dependent variable vector and independent variables. This Y-randomization was tested for model and low values of R^2 and R^2_{cv} were observed

Table-2: criteria for selection of good model

S/N	CRITERIA FOR SELECTION OF MODEL
1	N = number of molecules (> 20 molecules)
2	K= number of descriptors in a model (statistically N/5 descriptor in a model)
3	df = degree of freedom (N-K-1) (higher is better).
4	R^2 =coefficient of determination (> 0.7)
5	R^2cv = cross-validation square correlation (> 0.5)
6	R^2_{adj} = adjusted squared correlation coefficient (> 0.5)
7	R^2_{pred} = predicted coefficient of determination (> 0.5)
8	SEE = standard error of estimate (smaller is better)
9	F-test = F-test for statistical significance of the model (higher is better, for some set of descriptors and compounds)

RESULTS AND DISCUSSION

Using DFT and Semi-empirical calculated molecular descriptors, genetic function approximation was used to performed QSAR regression on 25 molecules of polyhaloginated dioxins compounds to generate 10 models from both DFT and Semi-empirical techniques. The toxicity (pEC_{50}) values were used as dependent variables and calculated molecular descriptors as independent variables as described by the equations in Table. The best models from both the DFT and SE techniques were selected on the basis of their statistical parameters as shown in Table-4.

Table-3: Validation parameters for DFT and SE

Density Functional Theory (DFT)		Semi-Empirical (SE)	
Parameter	Equation1	Parameter	Equation1
Friedman LOF	0.58817700	Friedman LOF	0.38776300
R-squared	0.95165500	R-squared	0.96803200
Adjusted R-squared	0.93893300	Adjusted R-squared	0.95962000
Cross validated R-squared	0.90908100	Cross validated R-squared	0.95180300
Significant Regression	Yes	Significant Regression	Yes
Significance-of-regression F-value	74.80187300	Significance-of-regression F-value	115.07027100
Critical SOR F-value (95%)	2.76172000	Critical SOR F-value (95%)	2.76172000
Replicate points	0	Replicate points	0
Computed experimental error	0.00000000	Computed experimental error	0.00000000
Lack-of-fit points	19	Lack-of-fit points	19
Min expt. error for non-significant LOF (95%)	0.28327500	Min expt. error for non-significant LOF (95%)	0.23000500

Table-4: Generated models from DFT and SE techniques

DFT	Y=42.7676*(BCUTc_1h)+7.7158*(VP-3)+5.9318*(SssssGe) -2.2284*(ETA_dApha_B)+32.7353*(ETA_BetaP)+1.0527
SE	Y=0.1841*(EHOMO)+5.0800*(SP-7)+10.6744*(ETA_Shape_P) +284.0681*(ETA_EtaP_L)-0.0023*(GRAV_4)+101.5783

Table-5: Correlation for DFT calculated descriptors

	pEC50	*BCUTc-1h*	*VP-3*	*SssssGe*	*ETA_dAlpha_B*	*ETA_BetaP*
pEC50	1					
BCUTc-1h	-0.3672	1				
VP-3	0.1166	0.4877	1			
SssssGe	-0.4491	0.6806	0.6530	1		
ETA_dAlpha_B	-0.3274	0.6973	0.8278	0.9602	1	
ETA_BetaP	0.5501	0.0551	0.7339	0.0072	0.2750	1

Table-6: Correlation for SE calculated descriptors

	pEC50	*EHOMO*	*SP-7*	*ETA_Shape_P*	*ETA_EtaP_L*	*GRAV-4*
pEC50	1					
EHOMO	-0.4682	1				
SP-7	0.1118	-0.4024	1			
ETA_Shape_P	0.5341	-0.8841	0.7369	1		
ETA_EtaP_L	0.6588	-0.9450	0.2396	0.8140	1	
GRAV-4	0.4395	-0.9255	0.6399	0.9507	0.8773	1

Table-7: Variance Inflation factor (VIF) for DFT and SE approaches

Density Functional Theory (DFT)		Semi-empirical (SE)	
Descriptors	VIF	Descriptors	VIF
BCUTc-1h	1.158	EHOMO	1.281
VP-3	1.014	SP-7	1.0127
SssssGe	1.2526	ETA_Shape_P	1.399
ETA_dAlpha_B	1.120	ETA_EtaP_L	1.7668
ETA_BetaP	1.439	GRAV-4	1.2394

Table-8: Definition of the descriptors used in the models

Descriptors	Definition
BCUTc-1h	nlow highest partial charge weighted BCUTS
VP-3	Chipath descriptor: valence path, order 3
SssssGe	Sum of atom type E-state:>Ge<
ETA_dAlpha_B	Extended Topochemical Atom descriptor: measure of count of hydrogen bond acceptor atoms and/ or polar surface area
ETA_BetaP	Extended Topochemical Atom descriptor: a measure of electronic features of the molecules relative to molecular size.
EHOMO	Energy of Highest Occupied Molecular Orbital.
SP-7	Chipath descriptor: Simple path, order 7.
ETA_Shape_P	Extended Topochemical Atom descriptor: Shape index P
ETA_EtaP_L	Extended Topochemical Atom descriptor:
GRAV-4	Gravitational index descriptor: gravitational index of all pairs of atoms (not just bonded pairs).

Table-9: Actual and predicted toxicity values for DFT and SE.

S/N	Density Functional Theory (DFT)			Semi-Empirical (SE)		
	Actual values pEC50	Predicted values	Residual values	Actual values pEC50	Predicted values	Residual values
1	8.0000	7.8504	0.1496	8.0000	7.6360	0.3640
2	7.1000	7.0987	0.0013	7.1000	7.1386	-0.0386
3	6.8000	6.4319	0.3680	6.8000	6.8130	-0.0130
4	6.6600	6.5102	0.1498	6.6600	6.2991	0.3609
5	6.5500	6.0009	0.5491	6.6500	6.4638	0.1862
6	6.1000	6.4722	-0.3722	6.1000	6.4554	-0.3554
7	5.9600	5.8006	0.1594	5.9600	6.0871	-0.1271
8	5.8900	6.3043	-0.4143	5.8900	5.6780	0.2118
9	7.1500	6.6762	0.4738	7.1500	6.8399	0.3100
10	5.1700	5.6075	-0.4375	5.1900	5.6462	-0.4562
11	4.8900	5.0809	-0.1909	4.8900	4.7512	0.1388
12	5.4900	5.2047	0.2853	5.4900	5.7988	-0.3088
13	5.0000	5.1862	-0.1862	5.0000	4.9205	0.0795
14	4.0000	4.2921	-0.2921	4.0000	4.1177	-0.1177
15	8.8200	8.9274	-0.1074	8.8200	9.1818	-0.3618
16	8.8300	9.1015	-0.2715	8.8300	8.7976	0.0324

17	9.3500	9.1085	0.2415	9.3500	9.1098	0.2402
18	7.9400	8.6603	-0.7203	7.9400	8.5207	-0.5807
19	7.0300	7.1834	-0.1534	7.0300	6.9865	0.0435
20	8.7000	8.6930	0.0070	8.7000	8.6295	0.0705
21	7.7700	7.8034	-0.0334	7.7700	7.5725	0.1975
22	8.1800	8.1236	0.0564	8.1800	8.3289	-0.1490
23	8.9300	8.5154	0.4146	8.9300	8.6621	0.2679
24	7.8100	7.7868	0.0232	7.8100	7.8609	-0.0510
25	6.5300	6.2299	0.3001	6.5300	6.4739	0.0560

Table-10. Univariate analysis for the toxicity data

Density Functional Theory (DFT)		Semi-Empirical (SE)	
Parameter	Value	Parameter	Value
Number of sample points	25	Number of sample points	25
Range	5.3500	Range	5.3500
Maximum	9.3500	Maximum	9.3500
Minimum	4	Minimum	4
Mean	6.9860	Mean	6.9908
Median	7.0300	Median	7.0300
Variance	2.0054	Variance	1.9994
Standard deviation	1.4453	Standard deviation	1.4432
Mean absolute deviation	1.1834	Mean absolute deviation	1.1784
Skewness	-0.1804	Skewness	-0.1876
Kurtosis	-1.0268	Kurtosis	-1.0170

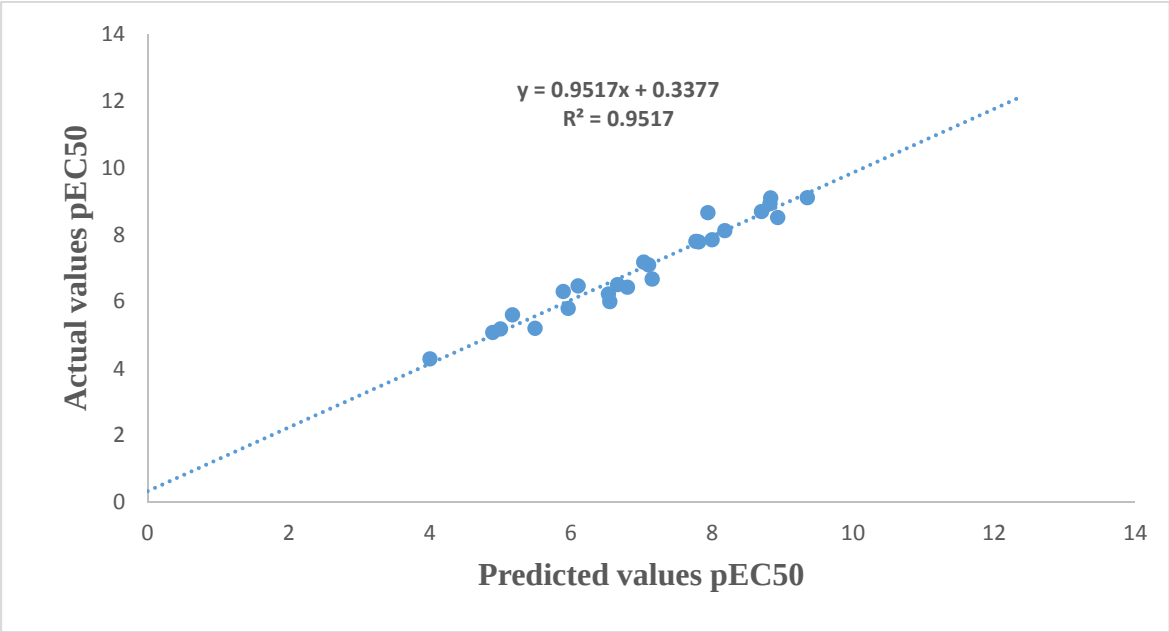


Fig. 1: Linear relationship between actual toxicity (pEC50) and the predicted for DFT Approach.

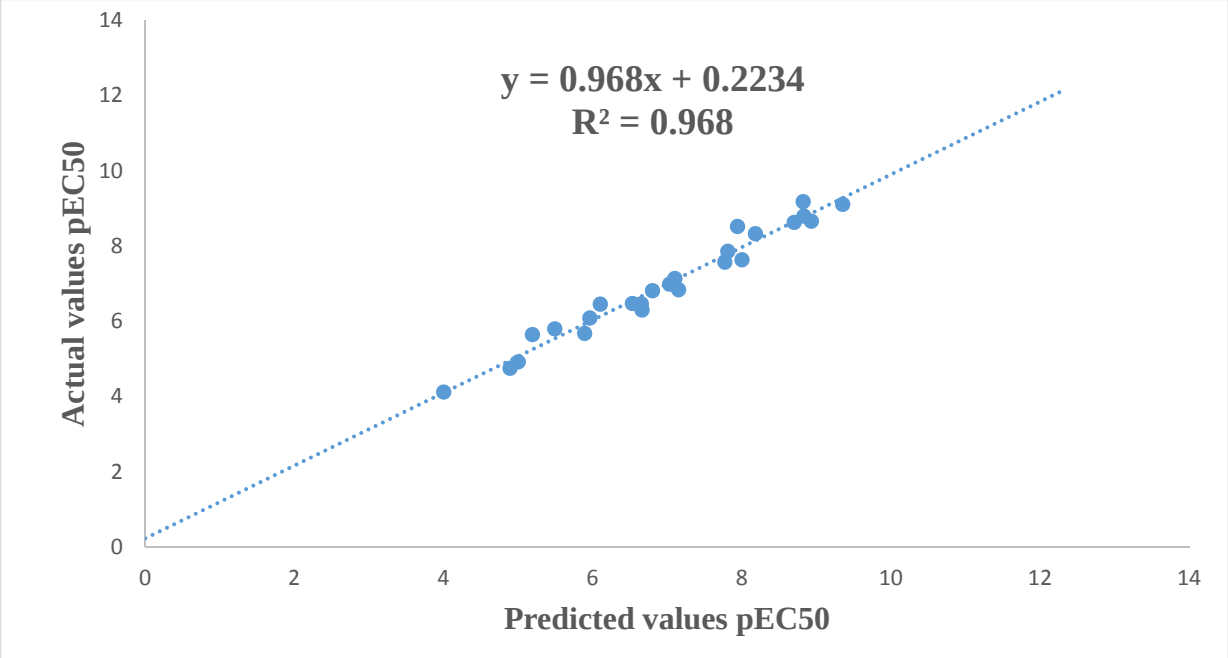


Fig. 2: Linear relationship between Actual toxicity (pEC50) and the predicted for SE approach.

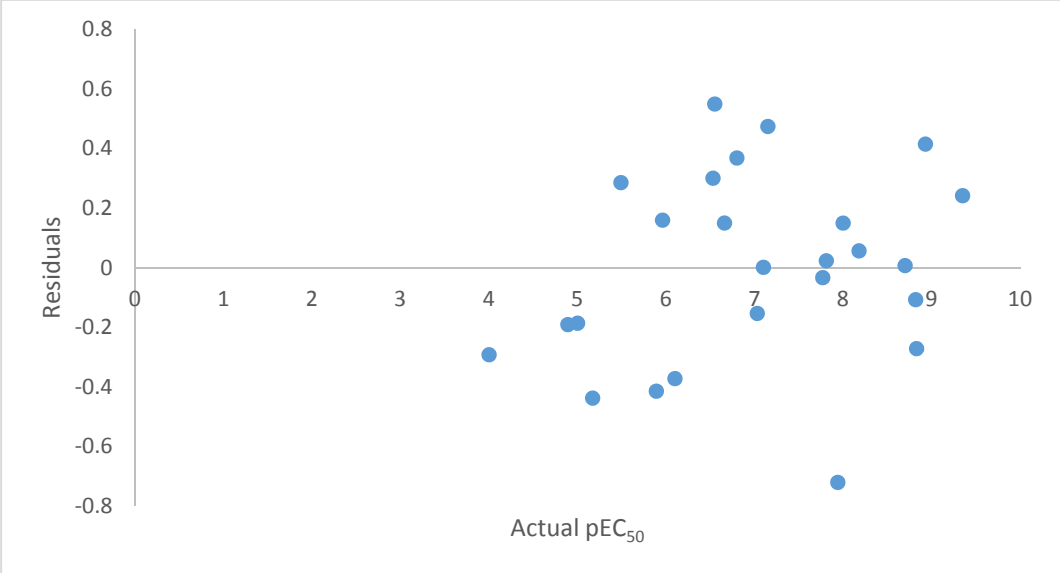


Fig. 3. Plot of residual versus actual toxicity values DFT clculated descriptors

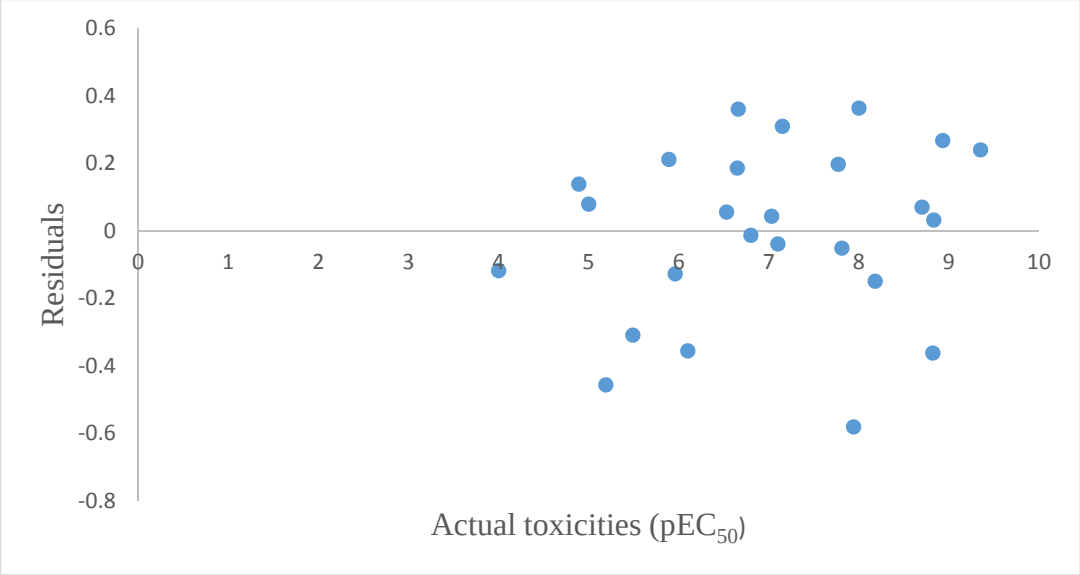


Fig.4. Plot of residuals versus actual toxicity values for SE calculated descriptors**.**

DISCUSSION

Among the ten QSAR models generated from both DFT and SE calculated molecular descriptors approaches, one model as presented in Table-4 was selected from both DFT and SE on the basis of various statistical parameters such as correlation coefficient squared (R^2), cross-validation squared correlation coefficient (R^2_{cv}), adjusted (R^2), lack of fit (LOF), standard error of estimate (SEE), and F-value. These parameters are presented in Table-3. The statistical parameters obtained using semi-empirical calculated molecular descriptors are a little bit better than those obtained using DFT approach. The predictive power of the model is determined based of these statistical parameters which explained in detail in the methodology. (DFT: Friedman LOF = 0.5882, R^2 = 0.9517, $R^2_{adj.}$ = 0.9389, R^2_{cv} = 0.9091, significance of regression F-value = 74.8019) and (Semi-empirical: Friedman LOF = 0.3878, R^2 = 0.9650, $R^2_{adj.}$ = 0.9596, R^2_{cv} = 0.95180, significance of regression F-value = 115.0703). All of these parameters are in very good agreement with the standard reported in Table-2. Table-4 shows the best model selected from both the methods, DFT and SE. The toxicity (Y) was used as independent variables and the descriptors (Xi) as dependent variables in the equations. Each model contains five descriptors as this agrees with the second criteria reported in Table-2. The correlation matrix in Table-5 and Table-6 show that the toxicities of these polyhalogenated dioxins are correlated with their descriptors for DFT and SE approaches respectively. The Variance Inflation factor (VIF) of all five descriptors were calculated using equation-8 and the corresponding VIF values of the five descriptors are presented in Table-7. As can be seen from this table, all the variables have VIF values of less than five, indicating that the obtained model from both the DFT and SE approaches has statistical significance, and the descriptors were found to be reasonably orthogonal. Table-9 shows the predicted toxicities in pEC_{50} of all the 25 molecules which in very good agreement with the experimental toxicities.Table-10 shows the statistical parameters of univariate analysis which describe the toxicity data. The most important data here are skewness and kurtosis. Skewness is the third moment of the distribution, which indicate symmetry of distribution. As skewness is positive, the distribution of the value within the column is skewed toward positive values. For a symmetry distribution, the skewness is close to zero. Kurtosis is the fourth moment of the distribution which indicates the profile of the column of data relative to normal distribution [41].

Contribution of descriptors

This study reveals that the following descriptors are found to be responsible for producing toxicities of polychlorinated dioxins, (DFT: BCUTc-1h, VP-3, SssssGe, ETA_dAlpha_Shape and ETA_BetaP, then for Semi-empirical, we have EHOMO, SP-7, ETA_Shape_P, ETA_Etap_P and GRAV-4). As can be seen in Table-4, Extended Topochemical Atom descriptor and Chipath descriptor appear to be found in both of the models generated using DFT and Semi-empirical calculated molecular descriptors. For DFT calculated descriptors, BCUTc-1h, VP-3, SssssGe and ETA_BetaP contribute positively in producing the toxicities of polyhologenated dioxins because of their positive coefficients in the model. This mean that increasing the values of these descriptors will produce high toxicities of these compounds and vice-versa. In the other hand, the descriptor, ETA_dAlpha_Shape which has negative coefficient in the model contributes negatively in producing the toxicities of these compounds. This indicates that decreasing the values of this descriptor will produce high toxicities of polyhalogenated dioxins. For the Semi-empirical calculated descriptors, the descriptors EHOMO,

SP-7, ETA_Shape_P and ETA_Etap_P which have positive coefficients in the models contribute positively in producing the toxicities of polyhalogenated dioxins. Increasing the values of these descriptors in polyhalogenated dioxins will produce higher toxicities [42]. The descriptor GRAV-4 has negative coefficient in the model and hence contributes negatively in producing the toxicities of these compounds.

Fig.1 and Fig.2 show plot that describe the linear relationship between the calculated toxicity (pEC_{50}) and the experimental toxicities of the 25 molecules of polyhalogenated dioxins for both the DFT and Semi-empirical calculated descriptors. Most of the compounds in the two figures are along the linear line of the plot. This indicates that, the calculated toxicities (pEC_{50}) are in very good agreement with the experimental values. Fig.3 and Fig.4 for DFT and Semi-empirical calculated descriptors respectively show the plot of the residual values versus experimental values of all the 25 molecules. The propagation of the residual values on both sides of zero indicates that no systematic error exists in the development of the models.

CONCLUSION

Genetic Function Approximation (GFA) technique was used to establish a correlation between (DFT and Semi-empirical calculated molecular descriptors) and experimental toxicities of polyhalogenated dioxins. Each of DFT (BLYP/31G*) and Semi-empirical (AM1) calculated descriptors were used to generate ten QSAR models. The model with the highest statistical significant for both DFT (BLYP/31G*) and Semi-empirical (AM1) calculated descriptors was selected. These models were selected based on their statistical parameters and were used to predict the toxicities of polyhalogenated dioxins. The prediction of the toxicity efficiencies for both DFT (BLYP/31G*) and Semi-empirical (AM1) calculated descriptors matched with the experimental measurements. This work reveals that semi-empirical (AM1) calculated descriptors gives a little bit better predictions than DFT (BLYP/31G*) calculated molecular descriptors.

For DFT-calculated molecular descriptors, the descriptors BCUTc-1h, VP-3, SssssGe, ETA_dAlpha_B, and ETA_BetaP were found to be the once responsible for producing toxicities of polyhalogenated dioxins. For Semi-empirical calculated descriptors, EHOMO, SP-7, ETA_Shape_P, ETA_EtaP_L and GRAV-4 were found to be the once responsible for toxicities of polyhalogeneted dioxins. All of these calculated molecular descriptors were aimed to encode some important information about the structural features of polyhalogenated dioxins which could influence the receptor binding affinity.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCES

[1] H Sadegh, M Yari, R Shahryari-ghoshkandi, S Ebrahimiasl, B Mazinejad, M Jalili and M Chahardon. Dioxin: a review of its environmental risk. Pyrex Journal of research inEnviromental studies. 2014, 1, 1-7.

[2] Eichbaum, K. Brinkmann, M.,Buchinger, S. Reifferscheid, G., Hecker, M., Giesy, J. P.,& Hollert, H. (2014). In vitro bioassays for detecting dioxin-like activity—Application potentials and limits of detection, a review. Science of the Total Environment, 487, 37-48.

[3] Giesy, J. P., Ludwig, J. P., &Tillitt, D. E. (1994). Deformities in birds of the Great Lakes region. Environmental science & technology, 28(3), 128A-135A.

[4] Larsson, M., Hagberg, J., Rotander, A., van Bavel, B., &Engwall, M. (2013). Chemical and bioanalyticalcharacterisation of PAHs in risk assessment of remediated PAH-contaminated soils. Environmental Science and Pollution Research, 20(12), 8511-8520.

[5] Poland, A., & Knutson, J. C. (1982). 2, 3, 7, 8-Tetrachlorodibenzo-thorn-dioxin and related halogenated aromatic hydrocarbons: examination of the mechanism of toxicity. Annual review of pharmacology and toxicology, 22(1), 517-554.

[6] Song, M., Jiang, Q., Xu, Y., Liu, H., Lam, P. K., O'Toole, D. K. & Jiang, G. (2006). AhR-active compounds in sediments of the Haihe and Dagu Rivers, China. Chemosphere, 63(7), 1222-1230.

[7] Vries, M. D., Kwakkel, R. P., &Kijlstra, A. (2006). Dioxins in organic eggs: a review. NJAS-Wageningen Journal of Life Sciences, 54(2), 207-221.

[8] S. Khan, Q. Cao, A.J. Lin, and Y.G. Zhu, Concentrations and bioaccessibility of polycyclic aromatic hydrocarbons in wastewater-irrigated soil using in vitro gastrointestinal test, Environ. Sci. Pollut. Res. 2008; 15: 344–353

[9] Safe, S. H. (1986). Comparative toxicology and mechanism of action of polychlorinated dibenzo-p-dioxins and dibenzofurans.Annual review of pharmacology and toxicology, 26(1), 371-399.

[10] Whyte J.J, Jung R E, Schmitt CJ, Tillitt D.E. Ethoxyresorufin-O-deethylase (EROD) activity in fish as a biomarker of chemical exposure, 30. London, ROYAUME-UNI: Informa Healthcare; 2000

[11] Jouko Tuomisto, Terttu Vartianen and Jouni T. Tuomisto. Synopsis on Dioxins and PCBs. Nation Institution for health and welfare Mannerheimintie, 166, FIN-00300 Helsinki, Finland, 2011.

[12] G.W Lucier, C.J Portier, M.A Gallo., Receptor and dose-response models for the effects of dioxins. Environ. Health perspect. 101 1993 36-44.

[13] D.W Nebert, A Puga, V Vasiliou. Role of the Ah receptor and the dioxin-inducible [Ah] gene battery in toxicity, cancer, and signal-transduction. Ann. N.Y. Acad. Sci. 685, 1993: 624-640.

[14] US National Institute of Health, National Institute of Environmental Health Sciences (NIEHS). Dioxin Research at the National Institute of Environmental Health Sciences (NIEH). 2/28/2006).

[15] Chovancova, J., Kocan, A., Jursa, S. PCDDs, PCDFs and dioxin-like PCBs in food of animal origin (Slovakia). Chemosphere, 2005; 61:1305–1311.

[16] Domingo, J.L., Bocio, A. Levels of PCDD/PCDFs and PCBs in edible marine species and human intake: a literature review. Environ. Int. 2007; 33: 397–405.

[17] Knight, D.J and Behenu, D. Alternative to animal testing in the safety evaluation of products, alternative to laboratory animals, 2002; 30: 7.

[18] Katritzky, A.R., Maran, U., V.S. Lobanov and krelson M., Vhem Y. Info. Computer

[19] Katritzky, A.R., Pentrukhin, R Tatham, D., Basak, S Benfenatim, E. Karelaon, M., and Maran U. Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. Journal of Chemical Information and Computer Sciences 2001; 40: 1–18.
[20] D. Roggers, Some theory and examples of genetic function approximation with comparison to evolutionary techniques, in: J. Devillers (Ed.), Techniques,Genetic Algorithms in Molecular Modeling, Academic Press, London, 1996, pp.87–107
[21] J.H. Holland, Adaption in Natural and Artificial Systems, University of MichiganPress, Ann Arbor, MI, 1975
[22] D. Rogers, Data analysis using G/SPLINES, in: Advances in Neural Processing Systems 4, Morgan Kaufmann, San Mateo, CA, 1992
[23] Tahar, L, Azeddine, A.; Rachid, H.; Majdouline, L.; Mohammed, B.; and. Binding Affinities (AhR) of Polychlorinated Biphenyls (PCBs), Dibenzo-p-dioxins (PCDDs) and Dibenzofurans (PCDFs) Study Combining DFT and QSAR Results. IJARCSSE 2014; 4: 304-305.
[24] Huifeng Wu, Fei Li, Jianmen Zhao, Xiali Liu, Linbao Zhang. Docking and –QSAR studies on the Ah receptor binding affinities of polychlorinated biphenyls (PCDDs), dibenzo-p-dioxins (PCDDs) and dibenzofuran (PCDFs). Environmental Toxicology and pharmacology 2011; 32: 478-485.
[25] K Nandan, K ranjan, Md B Ahmad and B Sah. QSAR studies on polychlorinated aromatic compounds using topological descriptors. IJPSR, 2013, 4(7): 2691-2695.
[26] Van Mourik, Tanja; Gdanitz, Robert J. (2002). "A critical note on density functional theory studies on rare-gas dimers". Journal of Chemical Physics 116 (22): 9620–9623. Bibcode, 2002JChPh.116.9620V. doi:10.1063/1.1476010
[27] M Dewar. A semiempirical life. Columbus, OH: American Chemical Society. (1992) ISBN 0-8412-1771-8.
[28] Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN (2008). Applications of 2D descriptors in drug design: a DRAGON tale. Curr Top Med Chem 8: 1628-55.
[29] Minghu Song and Mathew Clark. Development and Evaluation of an in Silico Model for hERG Binding, J. Chem. Inf. Model., 2006, 46 (1), 392-400.
[30] Arodola Olayide, Adebimpe, Radha Charan Dash, Mahmoud E. S. Soliman. QSAR study on Diketo Acid and Carboxiamide Derivatives as Potent HIV-1 Integrase Inhibitor, letter in Drug & Discovery, 2014, 11, 000-000.
[31] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, second ed., Wiley-VCH, Weinheim, 1999.
[32] H. Kubinyi, Quant. Struct.–Act. Relat. 13 (1994) 393
[33] R. Leardi, in: J. Devillers (Ed.), Genetic Algorithms in Molecular Modeling, Academic Press, London, 1996, p. 67.
[34] Sofie Van Damme. Quantum Chemistry in QSAR, Quantum Chemical Descriptors, use, benefits and draw back. Thesis, department of inorganic and physical chemistry (2009).
[35] R Kunal, P P Roy, S Paul and I Mitra. On two Novel parameters for validation of predictive QSAR models. Molecules, 2009, 14: 1660-1701.
[36] S Shapiro, S., Guggenhein, B., Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. Quant. Struct.-Act. Relat. 1998, 17, 327-337.
[37] Jaiswal, M., Khadikar, P.V., Scozzafava, A., Supuran, C.T. carbonic anhydrase inhibitors: the first QSAR on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioog. Med. Chem. Lett. 2004, 14, 3283-3290.
[38] Masand V. H, Jawarkar R. D, Patil K. N, Mahajad D. T, Hadda T. B, Kurhade G. H. COMFA analysis and toxicity risk assessment of coumarin analogues as Mao-A inhibitors: attempting better insight in drug design. Der Pharm. Lett., (2010a), vol. 2 (6): 350-357
[39] Masand V. H, Jawarkar R. D, Patil K. N, Nazerruddin G. M, Bajaj S. O. Correlation potential of Wiener index vis-a- vis molecular refractivity, Antimalarial activity of xanthone derivatives. Org Chem Indian J. 2010b, vol. 6 (1): 30-38.
[40] A K Pathak; A B Mundada; A Shrivastava, Pharmacia, 2011, 1, P, 57.
[41] K F Khaled, Corrosion sci. 2011, 53, 3457-3465
[42] R Navin; K J Sanmati, ChemTech, 2012; 4, 1350-1360