Journal of Computational Methods in Molecular Design, 2015, 5 (4):46-60

Scholars Research Library
(http://scholarsresearchlibrary.com/archive.html)

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Insilico prediction of octanol-air partition coefficient of some persistent organic pollutants through QSPR modelling

John Philip Ameji[1*], Adamu Uzairu[1], Hassan Samuel[1], Adedirin Oluwaseye[2], Adawara Ndaghiya Samuel[1] and Onoyima Christian Chinweuba[3]

[1]Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria
[2]Chemistry Advance Laboratory, Sheda Science and Technology Complex, FCT, Abuja Nigeria
[3]Nigeria Police Academy Wudil, Kano State Nigeria

ABSTRACT

Quantitative Structure Property Relationship (QSPR) analysis was applied to 36 Persistent Organic Pollutants (POPs) using a combination of 0D, 1D, 2D and 3D molecular descriptors obtained by Semi empirical (pm3) method. The computed descriptors were correlated with the log of their experimental octanol-air partition coefficient (pK_{OA}).Genetic function approximation was used to derive the most statistically significant QSPR model as a calibration model for predicting the pK_{OA} of this class of molecules. Among the obtained QSPR models, the most statistically significant one was a five parameter linear equation with the squared correlation coefficient R^2 value of 0.9889, adjusted squared correlation coefficient R^{2}_{adj} value of 0.9860 and Leave one out (LOO) cross validation coefficient (Q^2) value of 0.9827. An external set was used for confirming the predictive power of the model ($R^{2}_{pred.}$ = 0.7471). It is envisaged that the QSPR results identified in this study will offer an efficient and cost effective method of assessing the fate of POPs in the environment.

Keywords: POPs, GFA, QSAR, Descriptors, octanol-air partition coefficient.

INTRODUCTION

Persistent Organic Pollutants (POPs) are chemical substances that persist in the environment, bio-accumulate through the food web, and pose a risk of causing adverse effects to human health and the environment. With the evidence of long-range transport of these substances to regions where they have never been used or produced and the consequent threats they pose to the environment of the whole globe, the international community has now, at several occasions, called for urgent global actions to reduce and eliminate releases of these chemicals, because they are in a nutshell: Highly toxic to humans and the environment, Persistent in the environment, resisting bio-degradation, Taken up and bio-accumulated in terrestrial and aquatic ecosystems Capable of long-range, trans boundary atmospheric transport and deposition. In nature these substances affect plant and animal development and growth. They can cause reduced reproductive success, birth defects, behavioral changes and death. They are suspected human carcinogens and disrupt the immune and endocrine systems [1].

The fate and behavior of persistent organic pollutants (POPs) in the environment has attracted substantial scientific and political interest, arising from concern over human exposure to these chemicals and their discovery in primeval environments far from source regions. The ability of certain POPs to undergo long range atmospheric transport has resulted in the negotiation of protocols (e.g. UN/ECE, UNEP) for their reduction or elimination, to reduce the risks to regional and global environments. These chemicals are released into the environment through a range of processes which include; release during the production process, release during use or accidental release during combustion processes [2].

The fate of these chemicals in the environment; where they are released, the physical processes governing their transport, where they accumulate need to be given important consideration. This requires absorption of enormous spectrum of information and ultirnately lead to models of chemical fate [3, 4]. One of these important information is the octanol-air partition coefficient, K_{OA}.

The transport of chernical through the globe is often compared to chromatography [5] where the air is the mobile phase and terrestrial lipids represent the stationary phase. The sinks for lipophilic chemicals thus include waxy cuticle on vegetation, the organic matter in soil and the oily filmwhich coats atmospheric particulate matter. Because octanol is a replacement for organic or lipid phases, the octanol-air partition coefficient (K_{OA})is recognized as a good descriptor for atmosphere-terrestrial lipid exchange [6, 7].

However, experimental determination of K_{OW} is costly and time consuming, and sometimes restricted by lack of sufficiently pure chemicals [8] hampering effective and transparent risk assessment process to the regulated and the regulator. To achieve the sustainable use of chemicals, they is a needfor validated process of risk assessment (in this case, bio-accumulations of POPs) through which we can evaluate the impact of both existing chemicals and those which will be produced in the future [2]. This has necessitated the development of a predictive Quantitative-structure property relationship model for K_{OA} of POPs.

The aim of this work is to build a rational and predictive Quantitative-structure property relationship model for octanol-air partition coefficient K_{OA} at room temperature of POPs.

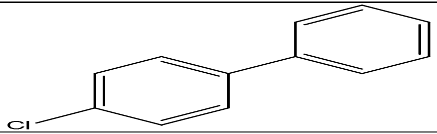
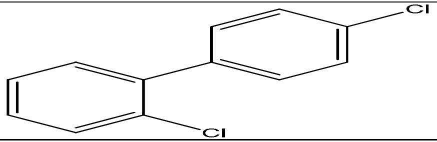
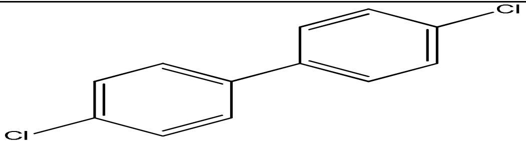
MATERIALS AND METHODS

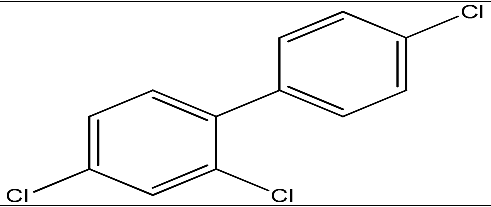
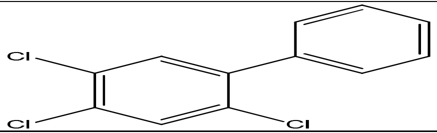
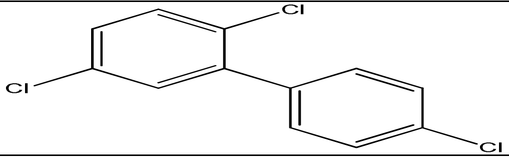
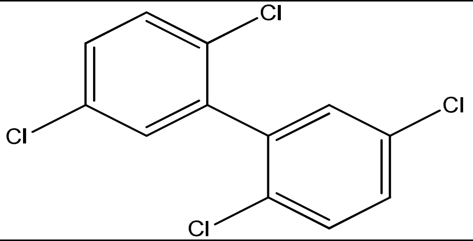
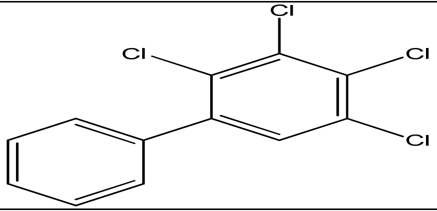
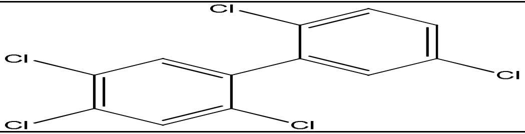
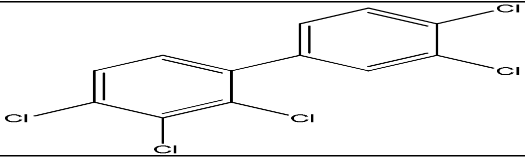
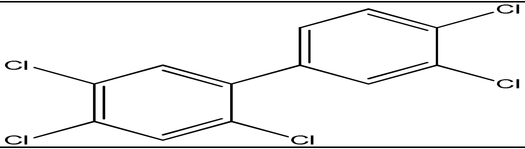
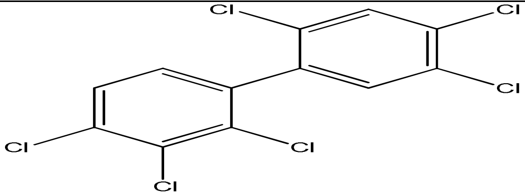
Hansch's approach [9] was used in the QSPR studies. In this approach, structural properties of compounds are calculated in terms of different physicochemical parameters and these parameters were correlated with biological activity through equation using regression analysis.

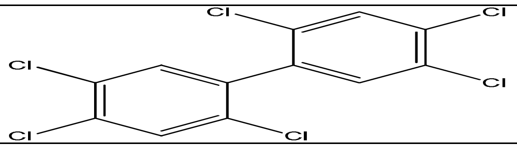
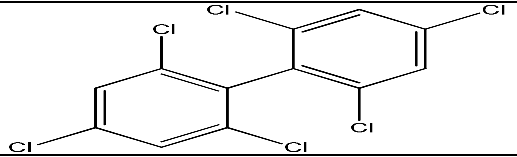
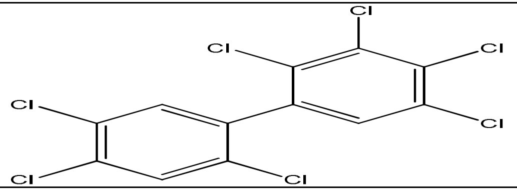
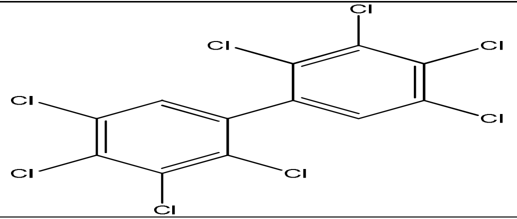
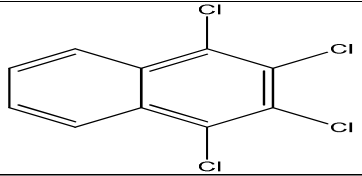
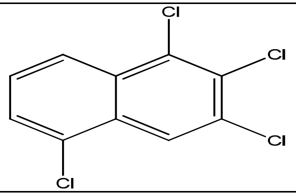
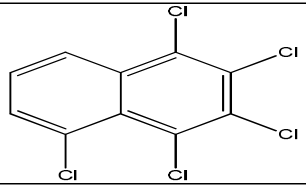
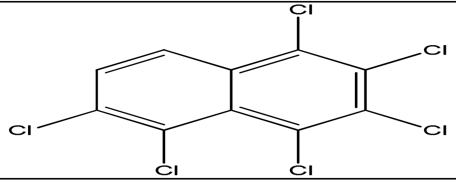
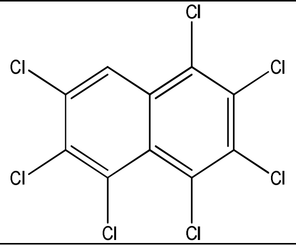
Data collection

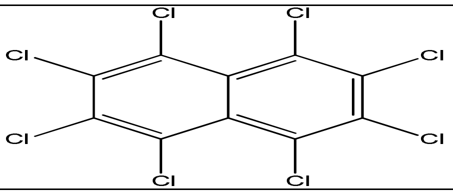
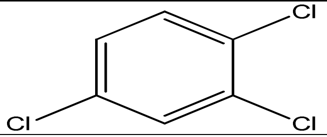
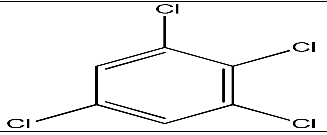
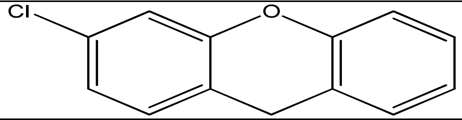
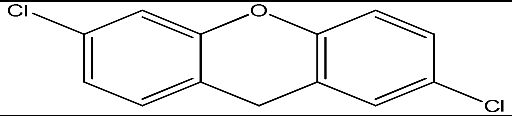
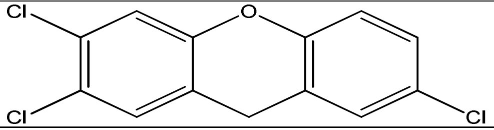
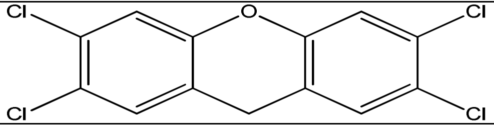
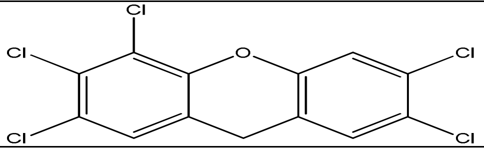
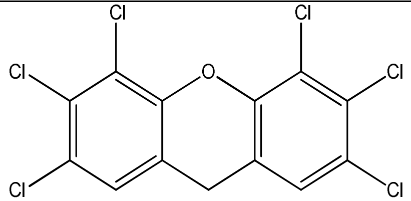
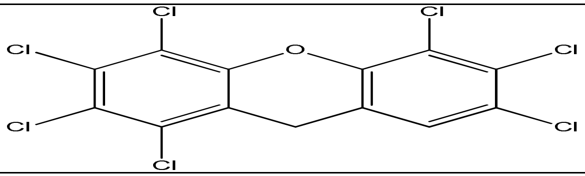
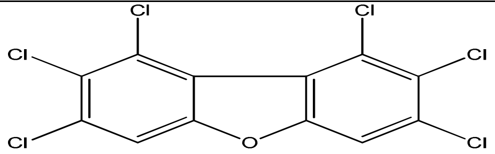
The chemical structures and experimental octanol-air partition coefficient in logarithmic scale (K_{OA}) of some persistent organic pollutant minimum were taken from literature [8, 4, 10]. The notation, structure and K_{OA} value for each member of the data set are presented in Table 1 below.

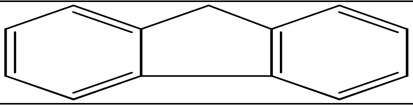
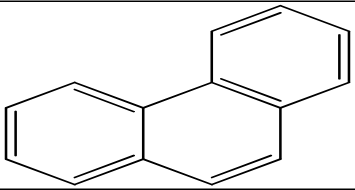
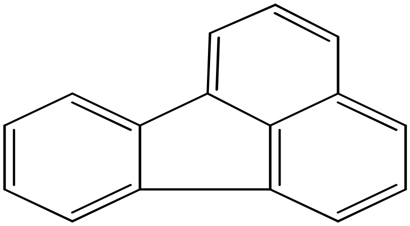
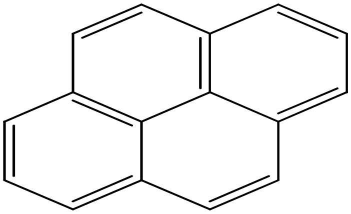
TABLE I: Experimental K_{OA} values of the selected POPs

S/n	Molecular Structure	LogK_{OA}
C1		6.82
C2		7.34
C3		7.85

C4		7.93
C5		7.80
C6		7.94
C7		8.22
C8		8.64
C9		8.90
C10		8.00
C11		9.80
C12		9.76

C13		9.52
C14		8.89
C15		10.12
C16		11.31
C17		8.27
C18		8.27
C19		9.02
C20		9.76
C21		10.51

C22		11.26
C23		5.20
C24		5.94
C25		7.62
C26		8.36
C27		9.11
C28		9.86
C29		10.61
C30		11.35
C31		12.10
C32		9.73

C33		6.79
C34		7.57
C35		8.88
C36		8.80

Geometry optimization

Geometry optimization has to do with the technique that tries to find the conformation of minimum energy of the molecule.The molecular structure of each compound in the data set was drawn with Chemdraw ultra V12.0 and saved as *cdx file. Calculations were performed using the molecular modeling program SPARTAN'14 V1.1.0 on H.P 650 computer system (Intel Pentium), 2.43GHz processor, 4GB ram size on Microsoft windows 7 Ultimate operating system.

The computational method invoked for calculating geometries in the present caseis termed a "cascade method" by Hehre [11] because of its use of molecular mechanicsas precursor for the more accurate semi-empirical methods. The attractiveness of the method lies in its ability to make calculations less computationally taxing by relegating initial geometry calculationsto less computationally intensive (and possibly more inaccurate) methods. In this method, the initial calculations, which may be initialized in a geometry far from that of equilibrium, are performed by those methods requiring less computational effort, allowing equilibrium geometries to be "honed in on" in later stages, leaving the refining to the more accurate and computationally intensive theories.

The molecules were first pre-optimized with the molecular mechanics procedure included in Spartan'14 V1.1.0 software and the resulting geometries were further refined by means of Semi-empirical (pm3). The lowest energy structure was used for each molecule to calculate their physicochemical properties (molecular descriptor).

Descriptor calculation

The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment [12].*Padel descriptor* tool kit was used to calculate the descriptors of the optimized molecules.

Training and Test set

The training set encompasses the molecules used in model development while the test set is made up of molecules not used in building the model, they are used in the external validation of the model. The data set for K_{OA} of the selected POPs was split into training and test set. At least 70% of the data set was used as training set and the rest as test set in line with the optimum splitting pattern of data set in QSAR study [13]. Consequently, the data set of 36 complexes was split into 25 training set and 11 test set. The training set was used to generate the model while the test set was used to evaluate its prediction abilities. The selection of training and test set was done using the Random Selection method.

Learning process

In this process, the correlation between the observed K_{OA} of the selected POPs and the calculated descriptors was obtained via correlation analysis using the Microsoft excel package in Microsoft office 2013. Pearson's correlation matrix was used as a qualitative model, in order to select the suitable descriptors for regression analysis. The selected descriptors were subjected to regression analysis with the experimentally determined octanol-air partition coefficienton logarithmic scale (pK_{OA}) as the dependent variable and the selected descriptors as the independent variables using Genetic function approximation (GFA) method in Material studio software. To develop the optimization model, 25 samples were included in the training set. The number of descriptors in the regression equation was set to 5, and Population and Generation were set to 1,000 and 5,000, respectively. The number of top equations returned was 5. Mutation probability was 0.1, and the smoothing parameter was 0.5. The models were scored based on Friedman’s LOF.

It is a distinctive characteristic of GFA that it could create a population of models rather than a single model. GFA algorithm, selecting the basis functions genetically, developed better models than those made using stepwise regression methods. And then, the models were estimated using the “lack of fit” (LOF), which was measured using a slight variation of the original Friedman formula, so that best model received the best fitness score [14].

In Materials Studio, LOF is measured using a slight variation of the original Friedman formula [15]. The revised formula is:

$$\text{LOF} = \text{SSE} / (1 - \frac{C+dp}{M})^2 \tag{1}$$

Where SSE is the sum of squares of errors, c is the number of terms in the model, other than the constant term, d is a user-defined smoothing parameter, p is the total number of descriptors contained in all model terms (ignoring the constant term) and M is the number of samples in the training set. Unlike the commonly used least squares measure, the LOF measure cannot always be reduced by adding more terms to the regression model. While the new term may reduce the SSE, it also increases the values of c and p, which tends to increase the LOF score. Thus, adding a new term may reduce the SSE, but actually increases the LOF score. By limiting the tendency to simply add more terms, the LOF measure resists over fitting better than the SSE measure (Materials Studio 5.0 Manual).The significant regression is given by F-test, and the higher the value, the better the model [16].

Model Validation

A reliable validation procedure is required in order to confirm the existence of chance correlations as well as ascertaining the fitting ability, stability, reliability and predictive ability of the developed models. The validation parameters of the optimum model were compared with the standards shown in table 2 below.

Table 2: Validation metrics for a generally acceptable QSAR model

S/n	Metric symbol	Name	Threshold
1	R^2	Coefficient of determination	≥ 0.6
2	Q^2	LOO cross validation coefficient	< 0.5
3	$R^2_{pred.}$	External test set’s coefficient of determination	≥ 0.6
4	$R^2 - Q^2$	Difference between R^2 and Q^2	≤ 0.3
5	F value	Variation ratio	High
6	$r^2 – r_0^2 / r^2$	Golbraikh and Tropsha condition	< 0.1
7	$r^2 – r'^2_0 / r^2$	Golbraikh and Tropsha condition	< 0.1
8	K and K'	Intercept	$0.85 \leq k$ or $k' \leq 1.15$

Source: Roy et al.; Ravinchandranet al.; Golbraikh and Tropsha [18, 19, 20]

2.6.1 Internal validation parameters

R^2 (the square of the correlation coefficient): describes the fraction of the total variation attributed to the model. The closer the value of R^2 is to 1.0, the better the regression equation explains the Y variable. R^2 is the most commonly used internal validation indicator and is expressed as follows:

$$R^2 = 1 - \frac{\sum(Yobs - Ypred)^2}{\sum(Yobs - Ytraining)^2} \qquad (2)$$

Where, Yobs; Ypred ;Ytraining are the experimental property, the predicted property and the mean experimental property of the samples in the training set, respectively .

Adjusted R^2 (R^2_{adj}): R^2 value varies directly with the increase in number of regressors i.e. descriptors, thus, R^2 cannot be a useful measure for the goodness of model fit. Therefore, R^2 is adjusted for the number of explanatory variables in the model. The adjusted R^2 is defined as:

$$R^2_{adj} = 1\text{-}\ (1 - R^2)\frac{n-1}{n-p-1} = \frac{(n-1)R^2 - P}{n-p+1} \qquad (3)$$

Where p = number of independent variables in the model.
(Brandon-Vaughn and Orr, 2015).

Q^2 (Leave one out cross validation coefficient): The LOO cross validated coefficient (Q^2) is given by;

$$Q^2 = 1 - \frac{\sum(Yp - Y)^2}{\sum(Y - Ym)^2} \qquad (4)$$

Where Yp and Y represent the predicted and observed activity respectively of the training set and $Y_m$the mean activity value of the training set [17].

Variance Ratio (F): this parameter is used to judge the overall significance of the regression coefficient. It is the ratio of regression mean square to deviations mean square defined as:

$$F = \frac{\sum(Ycal - Ym)^2}{p} \Bigg/ \frac{\sum(Yobs - Ycal)^2}{N-P-1} \qquad (5)$$

WhereY_{obs} stands for the observed response value, while Y_{calc} isthe model-derived calculated response and $Y_m$is the average of the observed response values.The F value has two degrees of freedom: p, N − p − 1. The computed F value ofa model should be significant at $p < 0.05$. A high F value is an indication that the regression coefficients are significant [18].

Standard error of estimate (s): Low standard error of estimate is an indication of a good model. It is defined as follows:

$$S = \sqrt{\frac{(Yobs - Ycal)^2}{N-P-1}} \qquad (6)$$

Its degree of freedom is N-p-1 [18].

2.6.2 Metrics for external validation

External validation of QSAR model is necessary to order to ensure the predictability and applicability of the developed QSAR model for the prediction of untested molecules.

Predictive R^2 ($R^2_{\ pred.}$):R^2 pred is termed the predictive R^2 of a development model and is an important parameter that is used to test the external predictive ability of a QSAR model. The predicted R^2 value is calculated as follows;

$$R^2_{pred.} = 1 - \frac{\sum[Yobs(test) - Ypred(test)]^2}{\sum[Yobs(test) - Ym(training)]^2} \qquad (7)$$

$Y_{pred(test)}$ and $Y_{obs(test)}$ indicate predicted and observed activity values respectively of the test set compounds and $Y_{m(training)}$ indicates mean activity value of the training set [19].

Golbraikh and Tropsha's criteria: according to Golbraikh and Tropsha, models are considered satisfactory, if all the following conditions are met.

(a) $R^2_{test} > 0.5$
(b) $r^2 - r_0^2 / r^2 < 0.1$
(c) $r^2 - r'^2_0 / r^2 < 0.1$
(d) $0.85 \leq k \leq 1.15$
(e) $0.85 \leq k' \leq 1.15$

Parameters r^2 and $r_0{}^2$ are the squared correlation coefficients between the observed and predicted values of the compounds with and without intercept, respectively. The parameter r'^2_0bears the same meaning but uses the reversed axes. K is the intercept of the plot of the observed and predicted values of the compounds and K' the reversed axes intercept [20].

RESULTS AND DISCUSSION

Table 3: GFA derived QSAR models for the K_{OA} of the selected POPs

Model	Equation	Definition of terms
1.	*p*KOA= 1.201087752 * X99 - 0.421961296 * X148 - 0.259510614 * X316 + 0.070956322 * X781 + 0.931435567 * X810 - 0.673649058	X99 : CW : VP-3 X148 : ET : nssCH2 X316 : LG : SsCl X781 : ADJ : WPSA-3 X810 : AEM : MOMI-R
2.	*p*KOA= 1.201087752 * X99 - 0.259510614 * X316 - 5.485496838 * X639 + 0.070956322 * X781 + 0.931435567 * X810 - 0.673649058	X99 : CW : VP-3 X316 : LG : SsCl X639 : XX : HybRatio X781 : ADJ : WPSA-3 X810 : AEM : MOMI-R
3.	*p*KOA = 0.971655315 * X99 - 7.881157260 * X639 - 4.131893428 * X745 + 1.029703269 * X810 + 0.085465727	X99 : CW : VP-3 X639 : XX : HybRatio X745 : ABZ : RotBtFrac X810 : AEM : MOMI-R
4.	*p*KOA = 0.971655315 * X99 - 0.606242867 * X148 - 4.131893428 * X745 + 1.029703269 * X810 + 0.085465727	X99 : CW : VP-3 X148 : ET : nssCH2 X745 : ABZ : RotBtFrac X810 : AEM : MOMI-R
5.	*p*KOA = 1.210319413 * X99 - 0.259275231 * X316 - 0.388137940 * X504 + 0.069010864 * X781 + 0.924947849 * X810 - 0.658474833	X99 : CW : VP-3 X316 : LG : SsCl X504 : SM : maxssCH2 X781 : ADJ : WPSA-3 X810 : AEM : MOMI-R

Table 4: Validation Parameters of the models

S/n	Parameters	Model 1	Model 2	model 3	Model 4	Model 5
1	Friedman LOF	0.06573200	0.0657320	0.06622200	0.06622200	0.06717800
2	R-squared	0.98891300	0.9889130	0.98633500	0.98633500	0.98866900
3	Adjusted R-squared	0.98599500	0.9859950	0.98360200	0.98360200	0.98568700
4	Cross validated R-squared	0.98272800	0.9827280	0.98014600	0.98014600	0.98235100
5	Significant Regression	Yes	Yes	Yes	Yes	Yes
6	Significance-of-regression F-value	338.929759	338.92976	360.905975	360.905975	331.549784
7	Critical SOR F-value (95%)	2.76172000	2.7617200	2.91676100	2.91676100	2.76172000
8	Replicate points	0	0	0	0	0
9	Computed experimental error	0.00000000	0.0000000	0.00000000	0.00000000	0.00000000
10	Min expt. error for non-significant LOF (95%)	0.15242700	0.1524270	0.16577700	0.16577700	0.15409500

The GFA algorithm makes use of a population of many models rather than generating a single model. The models are scored using Friedman's "lack of fit" (LOF) measure as the evaluation function [15, 21] as well as other validation parameters as shown in Table 4 above. Based on statistical significance, model 1 is selected as the optimization model for predicting the octanol-air partition coefficient of POPs because it has the least LOF score and minimum experimental error, highest R-squared, adjusted R-squared, Cross validated R-squared and F-value.

Table 5: Detailed definition of descriptors

S/n	Descriptor symbol	Definition
1	nssCH2	Count of atom-type E-State: -CH2-
2	MOMI-R	Radius of gyration
3	WPSA-3	Total molecular surface area / 1000
4	RotBFrac	Fraction of rotatable bonds, excluding terminal bonds
5	maxssCH2	Maximum atom-type E-State: -CH2-
6	SsCl	Sum of atom-type E-State: -Cl
7	VP-3	Valence path, order 3
8	HybRatio	Fraction of sp3 carbons to sp2 carbons

Plot of actual pK_{OA} against predicted pK_{OA}

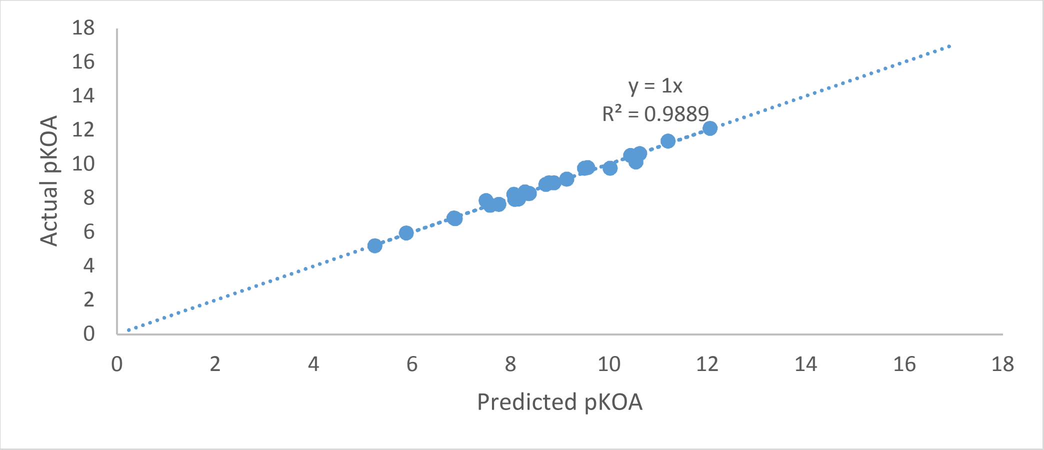


Figure 1

Residual plot of model 1

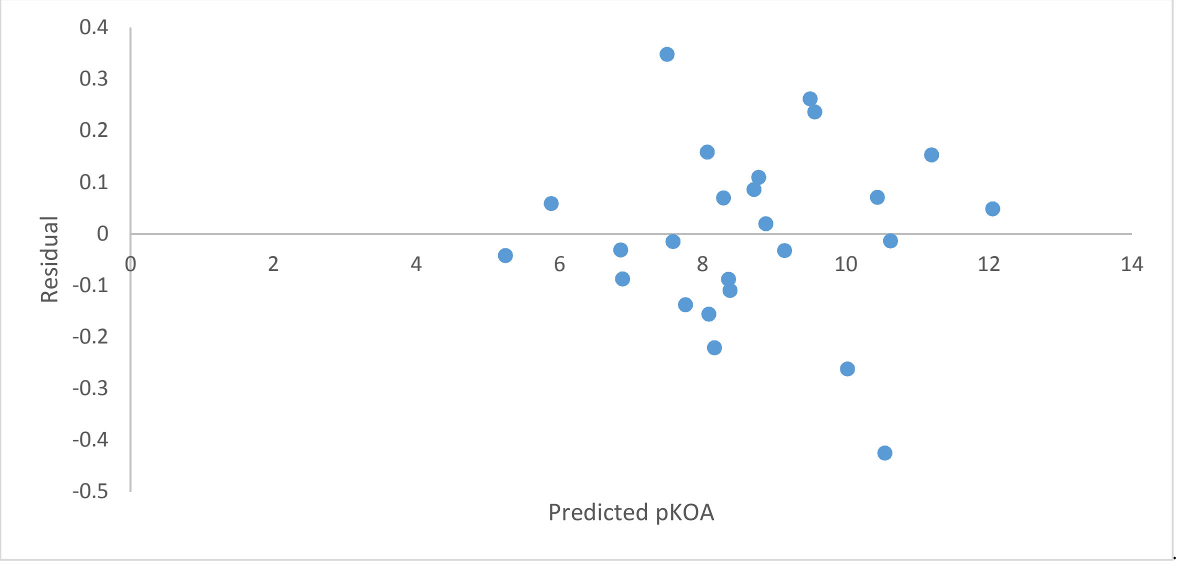


Figure 2

Table 6a: External validation of Model 1

Test set	ActualpK$_{OA}$	VP-3	nssCH2	SsCl	WPSA-3	MOMI-R	predicted pK$_{OA}$	Residual
C2	7.32	2.610648	0	1.901844	5.943832	5.441	7.45828	-0.13828
C5	7.8	3.129618	0	2.977422	5.486272	5.783878	8.08937	-0.28937
C8	8.64	3.938833	0	4.114013	5.359172	5.999731	8.95839	-0.3184
C10	8	4.298418	0	5.081722	7.208836	6.549077	9.78216	-1.78216
C13	9.52	4.379821	0	6.088428	6.043044	6.690994	9.66810	-0.1481
C16	11.31	5.998251	0	8.460297	5.754481	7.137775	11.3921	-0.08214
C19	9.02	4.179329	0	5.262004	5.907394	5.768978	8.77335	0.246647
C22	11.26	5.888305	0	8.761588	5.204924	6.599153	10.6412	0.618789
C28	9.86	4.25195	1	3.828033	6.317323	7.003276	9.98942	-0.12942
C32	9.73	5.246311	0	5.874562	7.504219	6.672433	10.8508	-1.1208
C35	8.88	3.412108	0	0	7.636583	5.285526	1.52498	-0.62499

Table 6b: External validation of Model 1

Cpds	Yobs(test)	Ym(traing)	Ypred(test)	(Yobs-Ypred)^2	(Yobs-Ym)^2
C2	7.32	8.66	7.45828	0.019121	1.7956
C5	7.8	8.66	8.089373	0.083737	0.7396
C8	8.64	8.66	8.958396	0.101376	0.0004
C10	8	8.66	9.782161	3.176098	0.4356
C13	9.52	8.66	9.668104	0.021935	0.7396
C16	11.31	8.66	11.39214	0.006748	7.0225
C19	9.02	8.66	8.773353	0.060835	0.1296
C22	11.26	8.66	10.64121	0.382899	6.76
C28	9.86	8.66	9.989424	0.016751	1.44
C32	9.73	8.66	10.8508	1.25619	1.1449
C35	8.88	8.66	8.889819	9.64E-05	0.0484
				$\sum$ = 5.125786	$\sum$=20.2562

The predicted R^2 value for the test set compounds was calculated using the formulae in equation 5.
Thus, $R^2_{pred.} = 1 - (\frac{5.125786}{20.2562}) = 0.7471$

Table 7: Golbraikh and Tropsha external validation parameters for model 1

s/n	parameter	value
1	r^2	0.7532
2	r'^2_0	0.7524
3	r_0^2	0.6924
4	k	1.027
5	K$'$	0.9694

Based on the parameters above;
$r^2 - r_0^2 / r^2 = \frac{0.7532-0.6924}{0.7521} = 0.081$
$r^2 - r'^2_0 / r^2 = \frac{0.7532-0.7524}{0.7521} = 0.001$

Table 8: Comparison of Yobs (training) and Ypred.(training) of model 1

Cpd.	Yobs.	Ypred.	residual
C1	6.82000000	6.851482	-0.14349800
C3	7.85000000	7.502049	0.03331700
C4	7.93000000	8.085627	0.06259100
C6	7.94000000	8.160617	0.04847500
C7	8.22000000	8.061669	-0.15624100
C9	8.90000000	8.880193	-0.25105100
C11	9.80000000	9.563873	-0.38314700
C12	9.76000000	10.02188	0.41793900
C14	8.89000000	8.78033	0.04602700
C15	10.12000000	10.54484	-0.08025300
C17	8.27000000	8.379892	-0.14597800
C18	8.27000000	8.357791	-0.08451000
C20	9.76000000	9.498029	0.06949700
C21	10.51000000	10.43915	0.26338100
C23	5.20000000	5.241971	-0.06403100
C24	5.94000000	5.88138	-0.05385800
C25	7.62000000	7.757617	-0.05937800
C26	8.36000000	8.290443	-0.02746400
C27	9.11000000	9.142405	-0.13035200
C29	10.61000000	10.6239	-0.01335500
C30	11.35000000	11.19715	0.03552000
C31	12.10000000	12.05141	0.23615000
C33	6.79000000	6.877068	0.05682700
C34	7.57000000	7.584947	-0.29004100
C36	8.80000000	8.71428	0.26850600

Table 9: Variance Inflation Factor (VIF) Statistic for the Descriptors in Model 1

S/n	Dependent Variable	R^2	VIF
1	VP-3	0.86	7.14
2	nssCH2	0.49	1.96
3	sscl	0.82	5.56
4	WPSA-3	0.28	1.39
5	Momi-R	0.81	5.26

3.3 Euclidean based applicability domain for the optimum QSAR model

The theoretical region in the chemical space constructed byboth the model descriptors and modeled response is termed applicability domain (AD). It plays a crucial role for assessing the uncertainty in the prediction of a particular compound based on how similar it is to the compounds employed to construct the QSAR model. The concept of AD is very important to QSAR studies considering the fact that it is unfeasible to predict the whole universe of compounds using a single QSAR model [22]. Euclidean AD is based on distance scores calculated by the Euclideandistance norms. At first, normalized mean distance score for training set compounds are calculated and these values ranges from 0 to 1(0=least diverse, 1=most diverse training set compound). Then normalized mean distance score for test set are calculated, and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain [23]. The Euclidean based applicability domain for the test and trainig set compounds of the optimum QSAR model (model 1) is shown in tables 9a and 9b respectively.

Table 10a: Euclidean based applicability domain for test set compounds

Cpd.	Distance Score	Mean Distance	Normalized Mean Distance
C1	87.55	3.502	0.181
C3	78.052	3.122	0.046
C4	76.039	3.042	0.018
C6	84.09	3.364	0.132
C7	89.416	3.577	0.207
C9	143.46	5.738	0.973
C11	80.735	3.229	0.084
C12	148.808	5.952	1
C14	80.902	3.236	0.087
C15	99.375	3.975	0.348
C17	117.87	4.715	0.611
C18	87.55	3.502	0.181

Table 10b: Euclidean based applicability domain for training set compounds

Cpd.	Distance Score	Mean Distance	Normalized Mean Distance
C1	101.776	4.071	0.382
C3	94.672	3.787	0.282
C4	84.131	3.365	0.132
C6	80.655	3.226	0.083
C7	98.031	3.921	0.329
C9	84.848	3.394	0.142
C11	78.907	3.156	0.058
C12	92.282	3.691	0.248
C14	96.151	3.846	0.303
C15	111.823	4.473	0.525
C17	74.794	2.992	0
C18	76.277	3.051	0.021
C20	96.148	3.846	0.303
C21	145.347	5.814	1
C23	114.941	4.598	0.569
C24	132.129	5.285	0.813
C25	102.551	4.102	0.393
C26	88.143	3.526	0.189
C27	79.64	3.186	0.069
C29	88.504	3.54	0.194
C30	99.909	3.996	0.356
C31	122.89	4.916	0.682
C33	118.914	4.757	0.625
C34	116.109	4.644	0.586
C36	114.436	4.577	0.562

Tables 3, 4, and 5 give the GFA derived QSAR models for predicting the octanol-air partition coefficient of some selected POPs, validation parameters of the models, and detailed definition of the descriptors used in the models respectively. Based on the validation parameters, the penta-parametric model (model 1) was selected as the optimization model for predicting the octanol-air partition coefficient of POPs. The Genetic Function Algorithm derived QSAR model is good agreement with the threshold shown in Table 2 as R^2 = 0.9889, R^2_{adj} = 0.9860, Q^2 = 0.9827,$R^2_{pred.}$ = 0.7471 and the Golbraikh and Tropsha criteria (Table 7) are also met. The predictability of model 1is evidenced by the low residual values observed in Table 8 which gives the comparison of observed and predictedoctanol-air partition coefficient of the molecules. Also, the plot of predicted pK_{OA}against observed pK_{OA}shown in Figure 1 indicates that the model is well trained and it predicts well the pK_{OA} of the compounds. Furthermore, the plot of observed pK_{OA} versus residual pK_{OA} (Figure 2) indicates that there was no systemic error in model development as the propagation of residuals was observed on both sides of zero [24].

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

$$\text{VIF} = {}^{1}\!/_{1-R^2} \qquad (4)$$

Where R^2 is the correlation coefficient of the multiple regression between the variables within the model. 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 [25, 26]. The corresponding VIF values of the five descriptors used in the optimization model (Model 1) are presented in Table 9. From this table, all the variables have VIF values of less than 10 indicating that the obtained model has statistical significance, and the descriptors were found to be reasonably orthogonal.

The applicability domain of the optimization model (model 1) was also defined for test set (Table 10a) and training set (Table 10b) compounds using Euclidean based approach. The results showed that all the compounds fall within the applicability domain of the model as their normalized mean distance score fall within the range of 0 and 1.

The result of the QSAR modelling indicated the predominance of the descriptors; nssCH2 (Count of atom-type E-State: -CH2-), MOMI-R (Radius of gyration), WPSA-3 (Total molecular surface area / 1000), SsCl (Sum of atom-type E-State: -Cl), VP-3 (Valence path, order 3).

SsCl and nssCH2 are atom-type E-state indices proposed as molecular descriptors encoding topological and electronic information related to particular atom types in the molecule [27]. The negative correlation of the descriptors; SsCl and nssCH2 in the model implies that the K_{OA} varies inversely with the values of these descriptors. Radius of gyration (MOMI-R) on the other hand is a size descriptor based on the distribution of atomic masses in a molecule. It is a measure of molecular compactness for long-chain molecules and, specifically, small values are obtained when most of the atoms are close to the center of mass [28].VP-3 (valence path, order 3) accounts for the presence of the heteroatom as well as double and triple bonds present in the compound [29] while WPSA-3 describes the surface area of the molecule. The positive coefficient of the descriptors; VP-3, WPSA-3 and MOMI-Rindicated that the magnitude of the K_{OA} of these compounds increases with increase in the values of these descriptors.

CONCLUSION

The generated QSAR models, performed to explore the structural requirements controlling the observed octanol-air partition coefficient of POPs, hinted that this property is predominantly affected bynssCH2 (Count of atom-type E-State: -CH2-), MOMI-R (Radius of gyration), WPSA-3 (Total molecular surface area / 1000), SsCl (Sum of atom-type E-State: -Cl), VP-3 (Valence path, order 3). The robustness and applicability of QSAR equation has been established by internal and external validation techniques. It is envisaged that this validated process of risk assessment provided by this model will help evaluate the impact of both existing chemicals (POPs) and those which will be produced in the future.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this paper. Also, they declare that this paper or part of it has not been published elsewhere**.**

CONTRIBUTION OF THE AUTHORS

This work was carried out in collaboration among all authors. Authors JPA and AU designed the study and wrote the protocol. Authors JPA, OCC, and ANS did the literature search and performed the statistical analysis. Authors JPA, AO and HS wrote the first draft of the manuscript. All authors read and approved the final manuscript.

REFERENCES

[1] United Nations Industrial Development Organisation (UNIDO). Retrieved 20th October, **2015** from http://www.unido.org/en/who-we-are/unido-in-brief.html.

[2] A. J. Sweetman, K. Prevedouros, N. Farrar, F. Jaward, K. C. Jones.*Lancaster University*; **2015**, (EPG 1/3/169).

[3] H. W. Vallack, D. J. Bakker , I. Brandt , E. B. M. Lunden , A. Brouwer , K. R. Bull , C. Gough, R. Guardans, I. Holoubek, B. Jansson, R. Koch, J. Kuylenstierna, A. Lecloux, D. Mackay, P. McCutcheon, P. Mocarelli, R. D.F. Taalman (1998). *Environ.Toxicol.Pharmacol.*; **1998**; 6:143–175.

[4] T. Harner. Application of octanol-air partition coefficients. PhD. Dissertation, University of Toronto, **1998**.
[5] F.Wania,D. Mackay. *Environ. Sci. Technol.* **1996**;30: 390-396.
[6] S. Paterson, D. Mackay,E. Bacci,D. Calamari**.***Environ. sci. Technol.* **1991**;*25*:866-871.
[7] S.L. Simonich, R.A. Hites.*Environ. Sci Technol.* **1995**; 2905-2914.
[8] X. Li, J. Chen,L. Zhang, X. Qiao, L. Huang.Dalian University of Technology, Linggong Road 2, Dalian 116024, People's Republic of China, **2006**.
[9] C. Hansch,T.P.Fujita.*J. Am Chem Soc*. **1964**; *86:* 1616-1626.
[10] N. Li, F.Wania,Y.D. Lei, G.L. Daly. University of Toronto at Scarborough, Toronto, Ontario, Canada M1C 1A4, **2003**.
[11] Hehre, W. J. Inc., Irvine, CA. **2005**.
[12] V. Consonni,R. T. Milano.University of Milano-Bicocca, PzadellaScienza 1 – 20126 Milano (Italy), **2012**.
[13] S.S. Patil.*Inter. J. Comp. Enginer. Res.*, **2011**; 2(4): 68-74.
[14] W. Wu,C. Zhang,W. Lin,Q. Chen,X. Guo,Y. Qian, Y.*PLoSONE*, **2015**; 10(3): e0119575.
[15] J.F.Friedman. Stanford University,**1990**,*Technical Report No. 102*.
[16] K.F. Khaled; N.S. Abdel-Shafi. *Int. J. Electrochem. Sci.*, **2011**; 6:4077 – 4094.
[17] B. K. Vaughn,A. Orr. Comprehensive R archive network (CRAN): http:// CRAN.R-project.org. Retrieved July 3rd, **2015**.
[18] K. Roy.*Springer Briefs in Molecular Science*. **2015**: DOI 10.1007/978-3-319-17281-1_2.
[19] V. Ravichandran,H. Rajak, A. Jain, S. Sivadasan, C. P. Varghese, R. K. Agrawal.*Inter. J. Drug DesignDiscov.*, **2011**; 2: 511-519.
[20] A. Golbraikh,A.J.Tropsha.*Mol. Graphics Mod*. **2002**, 20, 269-276.
[21] D. Rogers, A.J. Hopfinger.*J ChemInfComput Sci*. **1994**; 34: 854– 866.
[22] P. Gramatica P. *QSAR Comb Sci*. **2007**; 26:694–701.
[23] A. Pravin.DTC_EuclideanProgamme, Drug Theoretics&Cheminformatics (DTC)Laboratory, Jadavpur University, **2013**.
[24] M.J. Heravi,A. Kyani. *J. Chem. Inf. Comput. Sci.*, **2004**; 44: 1328–1335.
[25] S. Shapiro, B. Guggenheim. *Quant. Struct.-Act. Relat.*, **1998**; 17: 327–337.
[26] M. Jaiswal,P.V. Khadikar,A. Scozzafava,C.T. Supuran. *Bioorg. Med. Chem. Lett.* **2004**; 14: 3283–3290.
[27] L.B. Kier,L.H. Hall, L.H. (1999). *The Electrotopological State.* Academic Press, San Diego, **1999**.
[28] R. Todeschini, V.Consonni. Handbook of Molecular Descriptors, Wiley–VCH, Weinheim (GER), **2000**.
[29] B. Bhhatarai,P. Gramatica. *Chemical Research in Toxicology*, **2010**, 23, 528–539.