Journal of Computational Methods in Molecular Design, 2015, 5 (4):129-141

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

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Prediction of Henry's law constant of polycyclic aromatic hydrocarbons through quantitative structure property relationship modelling

John Philip Ameji[1*], Hambali Umar Hambali[2] and AlisiIkechukwu Ogadimma[3]

[1]Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria
[2]Department of Chemical Engineering, Ahmadu Bello University, Zaria, Nigeria
[3]Federal University Dutsinma, Kastina State, Nigeria

ABSTRACT

Polycyclic aromatic hydrocarbons are toxic, carcinogenic and are widely distributed in the environment. Accurate prediction of their aqueous Henry's law constant will be of immense help to environmental scientists in determining the fate of these chemicals in the environment. In this study, a Genetic function approximation (GFA)-QSPR analysis of some selected poly aromatic hydrocarbons (PAHs) was performed using different molecular descriptors. Five models for predicting the HLC of PAHs were generated. A seven parameter model with R^2 *= 0.996,* R^2_{adj} *= 0.994,* Q^2 *= 0.989,* R^2 *-* Q^2 *= 0.007,* $R^2_{pred.}$ *= 0.758,* $r^2 - r_0^2 / r^2$ *=0.00,* $r^2 - r'^2_0 / r^2$ *= 0.00, K = 0.998, K' = 1 was selected as the optimization model based on statistical significance. The Euclidean based applicability domain for training and test set compounds hinted that all the compounds fell within the applicability domain of the optimum QSPR model. The molecular descriptors; BCUTp-1l, ETA_Beta_ns, nFRing, topoDiameter, DPSA-2, LOBMIN, WD.mass were found to have profound influence on the predictive ability of the model. It is envisioned that the model will found excellent application in the prediction of Henry's law constant of poly aromatic hydrocarbons that fall within its applicability domain.*

Keywords: PAHs, QSPR, descriptors, Henry's law constant, GFA.

INTRODUCTION

Polycyclic aromatic hydrocarbons(PAHs) are a group of chemicals that are formed during the incomplete burning of coal, oil, gas, wood, garbage, or other organic substances, such as tobacco and charbroiled meat. PAHs are majorly derived from two processes: petro genic and pyrogenic processes. The petro genic part derives from oil- and drilling activities, including oil disasters, spills, and pollution from industrial sites, refineries, and most importantly traffic exhaust emissions, while the pyrogenic part derives from fires, forest fires, volcanic eruptions, and incineration. PAHs have long degradation periods, and recent studies show high accumulated concentrations in soil, aquatic, and atmospheric environments [1-4]. A few PAHs are used in medicines, production of dyes, plastics, and pesticides. Others are important constituents of asphalt used in road construction, crude oil, coal, coal tar pitch, creosote, and roofing tar [5].

Polycyclic aromatic hydrocarbons are globally distributed environmental contaminants with issues related to their known toxic and bio accumulative characteristics [6]. In humans, health risks associated with PAHs exposure

include cancer, DNA damage, tumor, reproductive defects, and damage to the skin, body fluids, and the immune system [5, 7-9].

For elucidating the environmental dynamics of PAHs, it is important to have sufficient data on the compound's air–water partition coefficient. This knowledge is very important for elucidating where the compounds tend to accumulate as well as evaluating the rate of transfer between the two phases. Conventionally, these rates are expressed as the products of kinetic constant such as mass transfer coefficient and the degree of departure from equilibrium which exist between the two phases. Elucidating the direction and the rate transfer of PAHs thus requires accurate values for the Henry's constant [10].According to Henry's law, the equilibrium ratio between the abundances in the gas phase and in the aqueous phase is constant for a dilute solution [11].

Henry's law constant (HLC) is a measure of the concentration of a chemical in air over its concentration in water. A PAH with a high HLC will volatilize from water into air and be distributed over a large area. Chemicals with a low HLC tend to persist in water and may be adsorbed onto soil. The HLC value is an integral part in calculating the volatility of a chemical.

Henry's Law constant for a chemical is generally expressed in one of two ways [12]:

$$HLC = \frac{Concentration\ in\ gas\ phase}{Concentration\ in\ liquid\ phase} \quad (1)$$

$$HLC = \frac{Liquid\ vapour\ pressure}{Chemical\ solubility} \quad (2)$$

Chemicals with a high HLC tend to volatilize from water and be distributed in the atmosphere. A chemical with a low HLC will tend to accumulate in water and soil, rather than volatilize. This can be an environmental concern since the accumulation of chemicals in water can have adverse effects upon living organisms [12].

As a result of the enormous number of chemicals of actual and potential concern, the difficulties and cost of experimentaldeterminations, and scientific interest in elucidating the fundamental molecular determinants of physical-chemicalproperties, considerable effort has been devoted to generating quantitative structure-property relationships (QSPRs) models. This concept is based on observations of linear free-energy relationships, and usually takes the form of a plot or regression of the property of interest as a function of an appropriate molecular descriptor which can be calculated using only a knowledge of molecular structure or a readily accessible molecular property [10].

The aim of this research is to build statistically robust quantitative structure property relationship models for predicting the Henry's law constant of poly aromatic hydrocarbons.

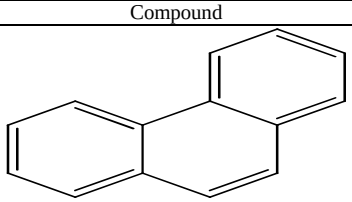
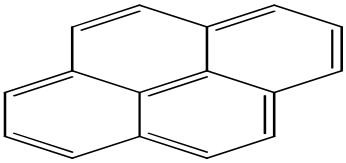
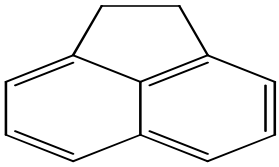
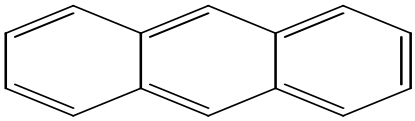
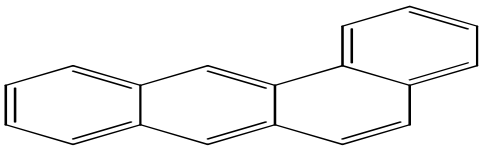
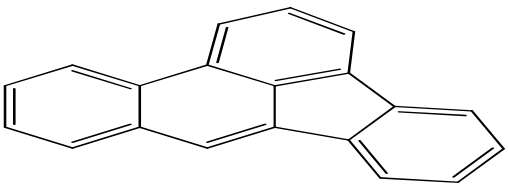
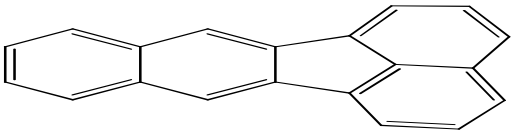
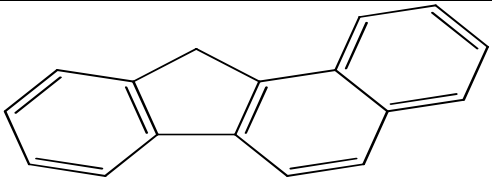
MATERIALS AND METHODS

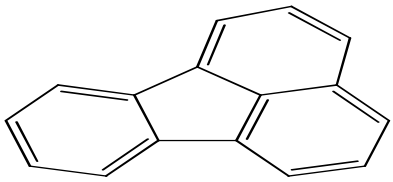
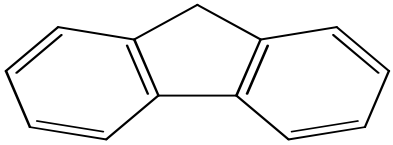
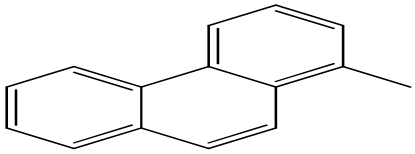
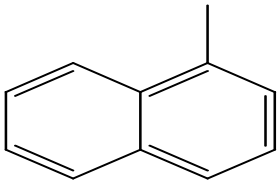
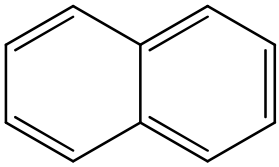
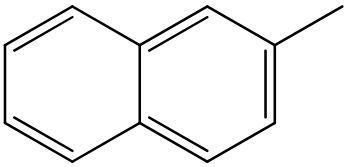
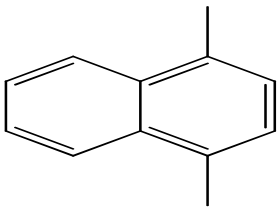
The compounds in the data set were optimized using SPARTAN'14 V1.1.0 molecular modelling software on H.P 650 computer system (Intel Pentium), 2.43GHz processor, 4GB ram size on Microsoft windows 7 Ultimate operating system. The molecules were pre-optimized with the molecular mechanics procedure included in Spartan'14 V1.1.0 software and the resulting geometries were further refined by means of Semi-empirical (pm3) method. This is termed the "Cascade method" [13]. This method of geometry optimization was used because it is less computationally taxing by relegating initial geometry calculations to less computationally intensive (and possibly more inaccurate) method.

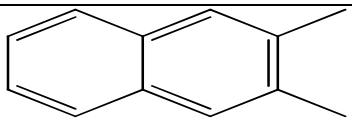
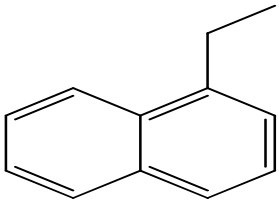
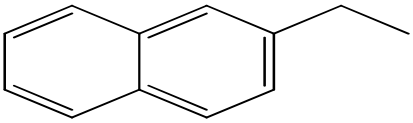
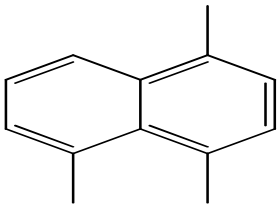
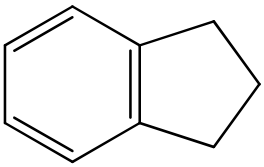
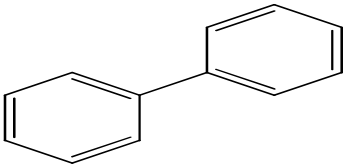
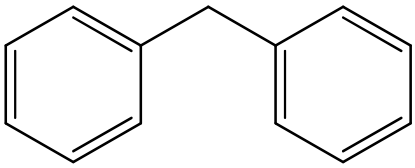
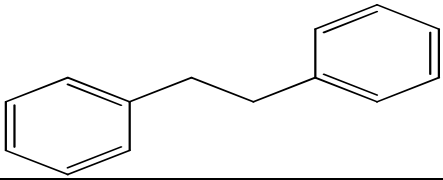
Data set

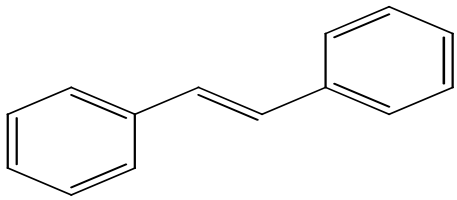
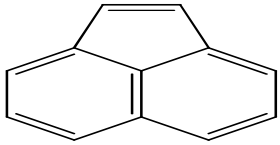
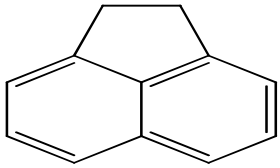
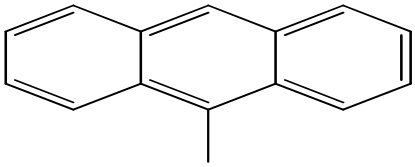
A data set comprising of series of 27PAHs with their experimental Henry's law constant values (HLC) values expressed on a logarithmic scale as pHLCwere taken from literature [10]for this study. 20 PAHs were used as training set for building the models while the remaining 7 were used as test set for external validation of the most statistically significant QSPR model. The notation, structure, HLC and pHLC of the compounds are shown in Table 1 below.

Table 1: PAHs with their Experimental HLC

S/n	Compound	HLC	pHLC
C1	 Phenanthrene	3.24	0.51
C2	 Pyrene	0.919	-0.04
C3	 Acenaphthene	12.17	1.09
C4	 Anthracene	3.961	0.6
C5	 Benz[a]anthracene	1.22	0.09
C6	 Benz[a]anthracene	0.051	-1.29
C7	 Benzo[k]fluoranthene	0.044	-1.36
C8	 Benzo[a]flourene	2.70	0.43

C9	 Flouranthene	0.957	-0.02
C10	 Flourene	7.87	0.9
C11	 1-methyl phenanthrene	5.00	0.7
C12	 1-Methyl naphthalene	44.89	1.65
C13	 Naphthalene	43.00	1.63
C14	 1-Methyl naphthalene	51.19	1.71
C15	 1,4-Dimethyl naphthalene	31.11	1.49
C16		62.49	1.8

	 2,3-Dimethyl naphthalene		
C17	 1-ethyl naphthalene	38.82	1.59
C18	 2-ethyl naphthalene	78.11	1.89
C19	 1, 4, 5-Trimethyl naphthalene	55.21	1.74
C20	 Indan	232.8	2.37
C21	 Biphenyl	28.64	1.46
C22	 Diphenylmethane	0.93	-0.032
C23		19.63	1.29

	Bibenzyl		
C24	 Trans-stilbene	40.40	1.61
C25	 Acenaphthylene	8.396	0.92
C26	 Acenaphthene	12.17	1.09
C27	 9-Methyl-anthracene	1.65	0.22

BUILDING OF QSPR MODEL

Genetic function approximation (GFA) method in Material studio software was used in the building of the models. The experimentally determined HLCon logarithmic scale (pHLC) as the dependent variable and the computed descriptors as the independent variables. In generating the GFA optimum QSPR models, 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 statistical significance of the generated models were assessed based on Friedman's LOF and the optimum model was selected based on this parameter.

In Materials Studio, LOF is measured using a slight variation of the original Friedman formula [14]. The revised formula is:

$$LOF = SSE / (1 - \frac{C+dp}{M})^2 \quad (3)$$

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.

Model Validation

Model validation was carried out in order to assess therobustness, fitting ability, stability, reliability and predictive ability of the developed models. The best GFA derived model obtained based on the model with the least lack of fit (LOF) score was subjected to both internal and external validation techniques and its validation parameters were compared with the minimum recommended standards for acceptable QSAR model shown in Table 2.

Internal validation was done using the data that created the model. The QSAR models were internally validated using the methods of least squares fit (R^2), cross validation coefficient (Q^2), adjusted R^2(R^2adj) and its confidence interval of all regression model at 95% significant level (α value). External validation on the other hand was performed on the basis of predictions of activities of molecules not used in the models using parameters as external test set's coefficient of determination($R^2_{pred.}$)[14].

Internal validation parameters

This validation is done using the data that created the model. The various internal validation parameters invoked in this study are presented thus;

R^2 (the square of the correlation coefficient):is the proportion of variability in a data set that is accounted for by a statistical model.It 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:

$$R^2 = 1 - \frac{\sum(Yobs-Ypred)^2}{\sum(Yobs-Ytraining)^2} \quad (4)$$

Where, Yobs; Ypred; Ytraining are the experimental property, the predicted property and the mean experimental property of the samples in the training set, respectively [15].

Adjusted R^2 (R^2_{adj}):is a modification of R-square that adjusts for the number of terms in a model. *R*-square always increases when a new term is added to a model, but adjusted R-square increases only if the new term improves the model more than would be expected by chance.The adjusted R^2 is defined as:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} = \frac{(n-1)R^2-P}{n-p+1} \quad (5)$$

Where p = number of independent variables in the model [16].

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

$$Q^2 = 1 - \frac{\sum(Yp-Y)^2}{\sum(Y-Ym)^2} \quad (6)$$

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[16].

.

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:

$$F = \frac{\frac{\sum(Ycal-Ym)^2}{p}}{\frac{\sum(Yobs-Ycal)^2}{N-P-1}} \quad (7)$$

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 [17].

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

$$S = \sqrt{\frac{(Yobs-Ycal)^2}{N-P-1}} \quad (8)$$

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

Leave one out cross validation (LOOCV): in this cross validation approach, the model is repeatedly refit leaving out a single observation and then used to derive a prediction for the left-out observation**.** For the model to have an excellent prediction ability, Q^2 must be > 0.5 and$R^2 - Q^2$ value should not exceed 0.3. The equation for CV is:

$$Q^2 = 1 - \frac{PRESS}{\sum(Yi - Ym)^2} \quad (9)$$

$$PRESS = \sum (Ypred, i - Yi) \quad (10)$$

Q^2 = LOOCV cross validation coefficient, R^2 = coefficient of determination.

$Yi$is the data value(s) not used to construct the CV model, PRESS is the predictive residual sum of the squares, Ym = mean of the experimental bioactivity (pMIC), $Ypred, i$ is the predicted Yi[15].

Metrics for external validation

External validation of QSAR model is performed in order to ensure the predictability and applicability of the developed QSAR model for the prediction of untested molecules. The various external validation metrics used in this work are highlighted thus:

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;

$$R^2_{pred.} = 1 - \frac{\sum[Yobs(test) - Ypred(test)]^2}{\sum[Yobs(test) - Ym(training)]^2} \quad (11)$$

$Y_{pred(test)}$ 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 [15].

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) $(R^2 - R'_0{}^2 / R^2) < 0.1$
(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 $R'_0{}^2$bears the same meaning with R_0^2but 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 [18].

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	$r^2 - r'_0{}^2 / r^2$	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

QSAR Study Results and Discussion

The best performing QSAR models for predicting the Henry's law constant of PAHs are represented by models 1, 2, 3, 4 and 5. The models, their internal validation parameters and detailed definition of the descriptors in the

modelsare shown in Tables 3, 4 and 5 respectively. Based on the model with the least LOF score and best statistical significance, model 1 was chosen as the optimization model for predicting the HLC of PAHs.

Table 3: GFA derived QSPR models for the pHLC of PAHs

Model	Equation	Definition of terms
1.	$pHLC = 0.212788074 * X36 - 0.186230525 * X611 - 0.089014980 * X686 + 0.250943018 * X748 - 0.002217301 * X771 - 0.290285405 * X803 - 1.202739805 * X839 + 2.258788057$	X36 : BCUTp-1l X611 : ETA_Beta_ns X686 : nFRing X748 : topoDiameter X771 : DPSA-2 X803 : LOBMIN X839 : WD.mass
2.	$pHLC = 0.193503331 * X36 - 0.136428149 * X47 - 0.086104163 * X686 + 0.226182560 * X748 - 0.002550314 * X771 - 0.308283129 * X803 - 1.142537556 * X839 + 2.215240549$	X36 : BCUTp-1l X47 : nBondsM X686 : nFRing X748 : topoDiameter X771 : DPSA-2 X803 : LOBMIN X839 : WD.mass
3.	$pHLC = -0.136428149 * X7 + 0.193503331 * X36 - 0.086104163 * X686 + 0.226182560 * X748 - 0.002550314 * X771 - 0.308283129 * X803 - 1.142537556 * X839 + 2.215240549$	X7 : nAromBond X36 : BCUTp-1l X686 : nFRing X748 : topoDiameter X771 : DPSA-2 X803 : LOBMIN X839 : WD.mass
4.	$pHLC = 0.260929211 * X36 - 0.202419232 * X611 - 0.095435731 * X686 + 0.226364818 * X748 - 0.019758863 * X772 - 0.287694041 * X803 - 1.226705548 * X839 - 0.047663266 * X841 + 2.290859517$	X36 : AS : BCUTp-1l X611 : WW : ETA_Beta_ns X686 : ZT : nFRing X748 : ACD : topoDiameter X772 : ADB : DPSA-3 X803 : AEG : LOBMIN X839 : AFQ : WD.mass X841 : AFS : Wlambda2.volume
5.	$pHLC = -0.153603875 * X7 + 0.244607053 * X36 - 0.091109132 * X686 + 0.198559583 * X748 - 0.021832375 * X772 - 0.308995886 * X803 - 1.158389036 * X839 - 0.055049268 * X841 + 2.258707794$	X7 : J : nAromBond X36 : AS : BCUTp-1l X686 : ZT : nFRing X748 : ACD : topoDiameter X772 : ADB : DPSA-3 X803 : AEG : LOBMIN X839 : AFQ : WD.mass X841 : AFS : Wlambda2.volume

Table 4: Validation Parameters of the models

S/n	Parameters	Model 1	Model 2	Model 3	Model 4	Model 5
1	Friedman LOF	0.016	0.017	0.017	0.017	0.018
2	R-squared	0.996	0.996	0.996	0.997	0.997
3	Adjusted R-squared	0.994	0.994	0.994	0.995	0.995
4	Cross validated R-squared	0.989	0.988	0.988	0.993	0.992
5	Significant Regression	Yes	Yes	Yes	Yes	Yes
6	Significance-of-regression F-value	464.238	445.326	445.326	486.361	475.867
7	Critical SOR F-value (95%)	2.919	2.919	2.919	2.953	2.953
8	Replicate points	0	0	0	0	0
9	Computed experimental error	0	0	0	0	0
10	Min expt. error for non-significant LOF (95%)	0.059	0.060	0.060	0.053	0.054

Table 5: Detailed definition of descriptors

S/n	Descriptor symbol	Definition
1	nAromBond	Number of aromatic bonds
2	BCUTp-1l	nhigh lowest polarizability weighted BCUTS
3	nBondsM	Total number of bonds that have bond order greater than one
4	ETA_Beta_ns	A measure of electron-richness of the molecule
5	LOBMIN	The L/B ratio for the rotation that results in the minimum area
6	nFRing	Number of fused rings
7	topoDiameter	Topological diameter (maximum atom eccentricity)
8	DPSA-2	Difference of FPSA-2 and PNSA-2
9	Wlambda2.volume	Directional WHIM, weighted by van der Waals volumes
10	WD.mass	Non-directional WHIM, weighted by atomic masses

Where PNSA-2 = Partial negative surface area * total negative charge on the molecule
FPSA-2 = PPSA-2 / total molecular surface area
PPSA-2 = Partial positive surface area * total positive charge on the molecule

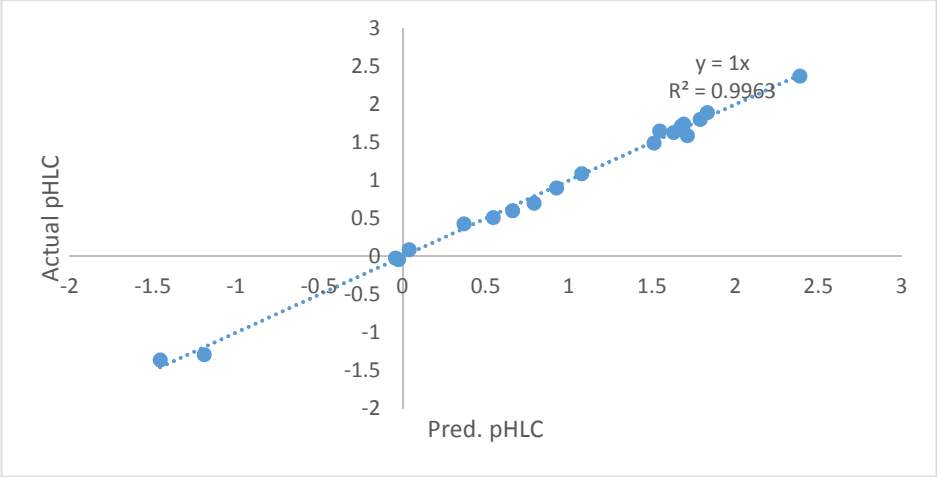


Figure 1:Plot of actual pHLC against predicted pHLC

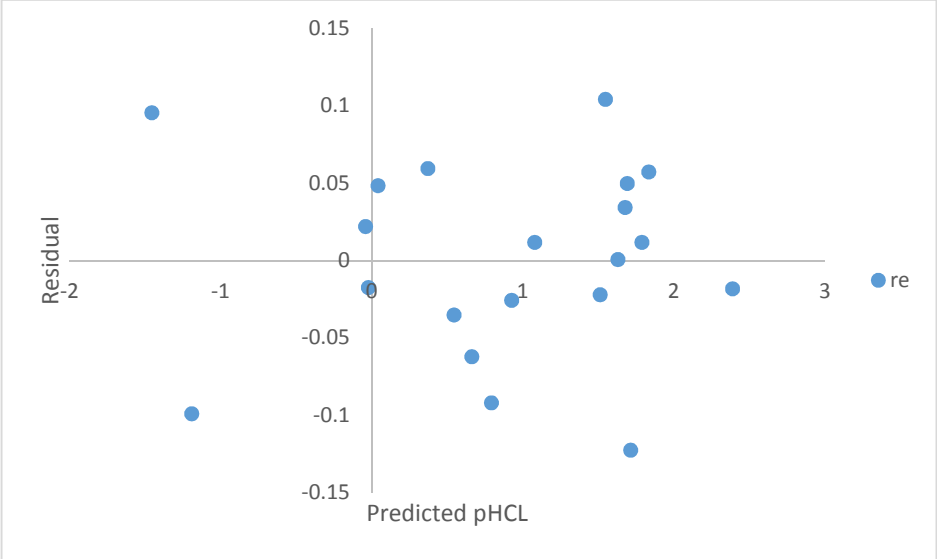


Figure 2: Residual plot of model 1

Table 6: Comparison of Yobs (training) and Ypred.(training) of model 1

Compound	Y_{obs}	Y_{pred}	Residual
C1	0.510	0.545	-0.035
C2	-0.040	-0.023	-0.017
C3	1.090	1.078	0.012
C4	0.600	0.662	-0.062
C5	0.090	0.042	0.048
C6	-1.290	-1.191	-0.099
C7	-1.360	-1.456	0.096
C8	0.430	0.371	0.059
C9	-0.020	-0.042	0.022
C10	0.900	0.926	-0.026
C11	0.700	0.792	-0.092
C12	1.650	1.546	0.104
C13	1.630	1.629	0.001
C14	1.710	1.676	0.034
C15	1.490	1.512	-0.022
C16	1.800	1.788	0.012
C17	1.590	1.713	-0.123
C18	1.890	1.833	0.057
C19	1.740	1.690	0.050
C20	2.370	2.388	-0.018

Table 7: External validation of Model 1

Test cpd	Y_{obs}	BCUTp-1l	ETA_Beta_ns	nFRing	Topo Diameter	DPSA-2	LOBMIN	WD.mass
C21	1.255	6.823	6	1	4	103.44	1.172	0.457
C22	1.415	6.637	12	0	7	122.84	2.375	0.579
C23	1.362	6.597	12	0	8	159.21	1.762	0.547
C24	1.623	6.577	12	0	9	197.88	2.395	0.675
C25	1.204	6.598	13.5	0	9	193.11	2.113	0.599
C26	1.322	7.363	11.5	4	5	120.112	2.55	0.612
C27	1.362	7.321	10	4	5	119.75	2.781	0.449

Test comp.	Y_{pred}	Ym	$(Y_{obs}-Y_{pred})^2$	$(Y_{obs}-Y_m)^2$
C21	1.885448	0.874	0.397465	0.145161
C22	1.566714	0.874	0.023017	0.292681
C23	1.41137	0.874	0.002437	0.238144
C24	1.640206	0.874	0.000296	0.561001
C25	1.223617	0.874	0.000385	0.1089
C26	1.329702	0.874	5.93E-05	0.200704
C27	1.271236	0.874	0.008238	0.238144
			$\sum$ = 0.431898	$\sum$= 1.784735

But from equation 11, $R^2_{pred.} = 1 - \frac{\sum[Yobs(test)-Ypred(test)]^2}{\sum[Yobs(test)-Ym(training)]^2}$

Thus, $R^2_{pred.} = 1 - (\frac{0.432}{1.785}) = 0.758$

Table 8: Euclidean based applicability domain for test set compounds

Test cpd.	Distance Score	Mean Distance	Normalized Mean Distance
C21	3033.555	151.678	1
C22	2683.448	134.172	0.792
C23	2045.954	102.298	0.414
C24	1583.16	79.158	0.139
C25	1626.526	81.326	0.165
C26	2732.134	136.607	0.821
C27	2738.535	136.927	0.825

Table 9: Euclidean based applicability domain for training set compounds

Training set cpd	Distance Score	Mean Distance	Normalized Mean Distance
C1	1409.442	70.472	0.036
C2	1374.093	68.705	0.015
C3	1617.893	80.895	0.159
C4	1527.582	76.379	0.106
C5	2719.743	135.987	0.814
C6	2363.332	118.167	0.602
C7	2565.25	128.262	0.722
C8	1887.493	94.375	0.32
C9	1447.933	72.397	0.058
C10	1349.493	67.475	0
C11	1589.849	79.492	0.143
C12	1425.134	71.257	0.045
C13	1521.999	76.1	0.102
C14	1405.68	70.284	0.033
C15	1358.475	67.924	0.005
C16	2080.396	104.02	0.434
C17	1820.706	91.035	0.28
C18	2112.537	105.627	0.453
C19	1660.846	83.042	0.185
C20	3033.372	151.669	1

Table 10: Golbraikh and Tropsha External Validation Parameters for the Optimum Model

s/n	parameter	value
1	r^2	0.9963
2	r'^2_0	0.9963
3	r_0^2	0.9963
4	k	0.9979
5	K'	1

Based on the parameters above;

$r^2 - r_0^2 / r^2 = \frac{0.9963-0.9963}{0.9963} = 0.000$

$r^2 - r'^2_0 / r^2 = \frac{0.9963-0.9963}{0.9963} = 0.000$

The five Genetic Function Approximation derived QSPR models are presented in Table 3. The validation parameters and detailed definition of the descriptors used in the models are presented in Tables 4 and 5 respectively. Based on the model with the least LOF score and best statistical significance, the hepta-parametric model (model 1) was selected as the optimization model for predicting the Henry's law constant of polyaromatic hydrocarbons. The validation parameters of the QSPR model is good agreement with the minimum standard shown in Table 2 as its R^2 = 0.996, R^2_{adj} = 0.994, Q^2 = 0.989,R^2_{pred}. = 0.758.The results in Table 10 also shows that Golbraikh and Tropsha criteria for robust QSPR model were also met.

The residual values of a QSPR model is the difference between the experimental or observed value and the predicted value by the model, the excellent predictability of model 1is evidenced by the low residual values observed in Table 6 which gives the comparison of observed and predicted pHLC of the molecules. Also, the plot of predictedpHLC against observed pHLC shown in Figure 1 indicates that the model is well trained and it predicts well the pHLC of the compounds. Furthermore, the plot of observed pHLC versus residual pHLC (Figure 2) indicates that there was no systemic error in model development as the propagation of residuals was observed on both sides of zero [19].

Applicability domain (AD) is the physicochemical, structural or biological space, knowledge or information on which the training set of the model has been developed. The resulting model can be reliably applicable for only those compounds which are inside this domain.It 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 [20].

The applicability domain of the optimization model (model 1) was also defined for test set (Table 8) and training set (Table 9) 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.

CONCLUSION

A highly predictive Quantitative Structure Property Relationship model was generated for the Henry's law constant of poly aromatic hydrocarbons. The molecular descriptors; BCUTp-1l, ETA_Beta_ns, nFRing, topoDiameter, DPSA-2, LOBMIN, WD.mass were found to have profound influence on the predictive ability of the optimization model. The robustness, applicability, and reliability of the optimum QSPR model has been established by various validation techniques. It is envisioned that the model will found excellent application in the prediction of Henry's law constant of poly aromatic hydrocarbons that fall within its applicability domain.

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 OCC designed the study and wrote the protocol. Author JPA, HUH, and AIO did the literature search and performed the statistical analysis. Authors JPA and AIO wrote the first draft of the manuscript. All authors read and approved the final manuscript.

REFERENCES

[1] S Manzetti, *J. Inter. Soc.PAHs*.**2013**, 33 (4): 311-330.
[2] K Nikolaou;P Masclet; G Mouvier,*Sci. Total Environ*. **1984**, 32; 103–132.
[3] C Perez; A Velando; I Munilla; M Lopez-Alonso; D Oro,*Chemosphere*, **2008**, 56, 537–547.
[4] Baek S.O., Field R.A., Goldstone M.E., Kirk P.W., Lester J.N., PerryR.*Water, Air, Soil Pollut.***1991**, 60; 279–300.
[5] U.S. Department of Health and Human Services Public Health Service Agency for Toxic Substances and Disease Registry, **1995**.
[6] Menzie, C.A., Potocki, B.B., Santodonato, J.*Environ. Sci. Technol.*, **1992**, 26; 1278–1284.
[7] Bernd S., Luís M.N., B.F. Santos, Marisa A.A.R., Mariana B. O., Isabel M. M., João A.P. C.,*Chemosphere*,**2010**, 79; 821–829.
[8] ATSDR,*Polycyclic Aromatic Hydrocarbons*, *December*, Atlanta, GA, U.S., **1990**.
[9] USEPA, *Polycyclic Aromatic Compounds Category*, EPA 260-B-01-03, Washington, DC, August **2001**.
[10] Donald M.; Wan Y S. *J. Phys. Chem*., **1981**; 10 (4),1175-1199.
[11] Sander R. *Atmos. Chem. Phys*., **2015**; 15, 4399–4981.
[12] Clark D. J. *Physico -Chemical Properties and Environmental Fate of Pesticides*: 1020 N Street, Sacramento, California 95814, **1994**.
[13] Hehre, W. J. Inc., Irvine, CA. **2005**.
[14] J.F.Friedman. Stanford University,**1990**, *Technical Report No. 102*.
[15] V Ravichandran, H Rajak, A Jain, S Sivadasan, C P Varghese, R K Agrawal. *Inter. J. Drug Des.Discov.*, **2011**; 2: 511-519.
[16] B. K. Vaughn, A. Orr. Comprehensive R archive network (CRAN): http:// CRAN.R-project.org. Retrieved July 3rd, **2015**.
[17] K. Roy. *Springer Briefs in Molecular Science*. **2015**: DOI 10.1007/978-3-319-17281-1_2.
[18] A. Golbraikh, A.J.Tropsha. *Mol. Graphics Mod*. **2002**, 20, 269-276.
[19] M.J. Heravi, A. Kyani. *J. Chem. Inf. Comput. Sci.*, **2004**; 44: 1328–1335.
[20] A. Pravin. DTC_EuclideanProgamme, Drug Theoretics&Cheminformatics (DTC) Laboratory, Jadavpur University, **2013**.