Structural Chemistry (2023) 34:477–490
https://doi.org/10.1007/s11224-022-01992-2

ORIGINAL RESEARCH

Quantitative structure–activity relationship modeling of hydroxylated polychlorinated biphenyls as constitutive androstane receptor agonists

Lukman Kehinde Akinola[1,2] · Adamu Uzairu[1] · Gideon Adamu Shallangwa[1] · Stephen Eyije Abechi[1]

Received: 4 May 2022 / Accepted: 9 June 2022 / Published online: 17 June 2022

Abstract

Hydroxylated polychlorinated biphenyls (OH-PCBs), a series of toxic chemical compounds produced via biotic and abiotic transformation of polychlorinated biphenyls (PCBs), are known to cause endocrine disruption by interacting inappropriately with human nuclear receptors. Due to occurrence of high numbers of inactive OH-PCB congeners recorded in many experimental toxicity studies, it is pertinent to develop rapid and inexpensive QSAR models that can reliably predict the activities of OH-PCB congeners prior to experimental testing. Using a combination of genetic function approximation and multiple linear regression methods, a local QSAR model, consisting of six 2D descriptors (MATS1s, VE3_DzZ, VE1_Dzp, SpMin8_Bhv, SpMax5_Bhi, topoRadius) and two 3D descriptors (RDF95u, RDF45m), was developed from a training set of 44 OH-PCBs. Statistical parameters for fitting ($R^2 = 0.8902$, $R^2_{adj} = 0.8651$, $s = 0.2840$), cross-validation ($Q^2_{LOO} = 0.8201$, $RMSE_{CV} = 0.3242$), and Y-randomization ($cR^2_p = 0.8019$) obtained for the developed QSAR model indicate that the model is reliable, robust, and provides good fit to the data in the training set. The results of external validation carried out on 20 OH-PCBs in the test set also indicate that the developed QSAR model possessed good external predictivity and can be used to predict the agonistic activities of untested OH-PCB congeners to constitutive androstane receptor.

Keywords Agonistic activity · Constitutive androstane receptor · Hydroxylated polychlorinated biphenyls · Molecular descriptors · QSAR model

Introduction

Nuclear receptors are a class of proteins that function as ligand-activated transcription factors in human being [1–3]. They regulate the expression of specific genes that control processes such as reproduction, development, homeostasis, and metabolism upon binding to small endogenous molecules like steroid hormones, thyroid hormones, retinoid acids, fatty acids, and phospholipids [4–6]. Because of their structural similarities to endogenous ligands, some chemical compounds found in foods, cosmetics, environment, pharmaceuticals, and industrial products mimic the behaviors of endogenous ligands and interact inappropriately with nuclear receptors, thereby causing endocrine disruption [7–17]. Human exposure to endocrine disrupting chemicals has been shown in numerous studies to be associated with reduced semen quality [18–20], prostate cancer [21–24], urogenital tract abnormalities [25, 26], precocious puberty [27–29], irregular menstrual cycle [30], early menopause [31, 32], breast cancer [33–35], alteration of immune responses [36–38], diabetes [39–42], obesity [43–46], cardiovascular diseases [47], and a host of other adverse health effects.

In recent times, research efforts are being directed towards understanding the mechanisms involved in endocrine disruption by persistent organic pollutants. One of the most widely reported group of persistent organic pollutants with high endocrine disrupting potencies is polychlorinated biphenyls (PCBs) and their metabolites. Hydroxylated polychlorinated biphenyls (OH-PCBs) are a group of exogenous chemicals that are produced by oxidation of PCBs through

✉ Lukman Kehinde Akinola
lkakinola@basug.edu.ng

[1] Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria

[2] Department of Chemistry, Bauchi State University, Gadau, Nigeria

a variety of mechanisms, including metabolic transformation in living organisms and abiotic reactions with hydroxyl radicals [48]. Recently, the possible roles of nuclear receptors in mediating the endocrine disrupting effects of OH-PCBs are being vigorously studied by many research groups. Using a variety of in vitro bioassay methods, researchers have shown that some OH-PCB congeners act as agonists of estrogen receptor α (ERα) [49], estrogen receptor β (ERβ) [49], constitutive androstane receptor (CAR) [50], retinoid X receptor β (RXRβ) [51], retinoic acid receptor γ (RARγ) [51], and estrogen-related receptor γ (ERRγ) [52]. Some OH-PCB congeners have also been demonstrated in some in vitro studies to act as ERα antagonists [49], ERβ antagonists [49], androgen receptor (AR) antagonists [49], glucocorticoid receptor (GR) antagonists [49], and thyroid hormone receptor (TR) antagonists [53].

The major downside of the in vitro studies reviewed in the preceding paragraph is the high numbers of inactive OH-PCB congeners encountered in most of the reported experiments. For instance, only 22% and 38% of the OH-PCB congeners selected for investigation exhibited agonistic activities against RXRβ and RARγ, respectively [51]. Similarly, only 9%, 6%, and 30% of the OH-PCB congeners selected for investigation exhibited antagonistic activities against ERα, ERβ, and GR, respectively [49]. The low numbers of active OH-PCB congeners recorded in most of the experimental studies reviewed was due to the approach adopted by researchers in selecting OH-PCB congeners used in the experiments. In most experimental studies, the selections of chemicals used for toxicity testing were usually based on trial-and-error approach rather than on rational basis. This has resulted in waste of time and other valuable resources. In order to prioritize the choice of OH-PCB congeners for toxicity testing, it is pertinent to develop rapid and inexpensive quantitative structure–activity relationship (QSAR) models that can reliably predict the biological activities of OH-PCBs prior to experimental studies. QSAR modeling is a computational approach that establishes a correlation between the biological activities and measured or computed molecular features of a series of chemical compounds [54]. It is based on the assumption that changes in molecular structures of chemical compounds reflect corresponding changes in the observed biological activities [54]. Exposure of humans to PCBs and OH-PCBs has been

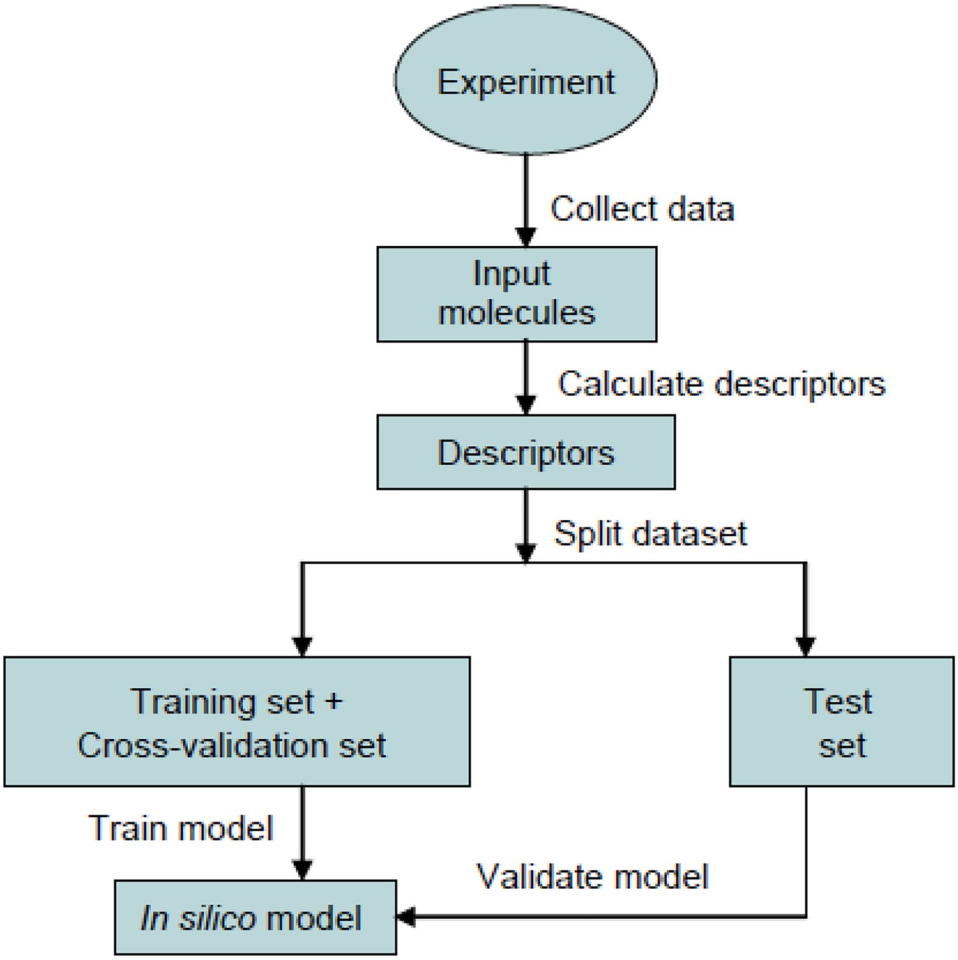

Fig. 1 Workflow showing the basic procedures for constructing and validating a QSAR model (reproduced from Wang and Hou [56])

linked to onset of obesity, diabetes, and fatty liver disease, and these metabolic disorders are thought to be mediated through activation of constitutive androstane receptor by this group of persistent organic pollutants [55]. Although QSAR models have been developed for toxicity prediction of some persistent organic pollutants, no QSAR model is currently available in the literature for the prediction of agonistic activities of OH-PCB congeners to constitutive androstane receptor. The objective of this study was to develop a QSAR model that can reliably predict the agonistic activities of untested OH-PCBs to constitutive androstane receptor using the limited experimental data available in literature.

Methods

The procedures adopted for constructing and validating the QSAR model described in this paper are summarized in the workflow displayed in Fig. 1 [56]. Details of the steps involved in these procedures are described below.

Dataset acquisition and molecular descriptor computation

The dataset used for the construction and validation of the QSAR model developed in this paper was obtained from the literature [50]. This dataset consists of structural formulae of 64 OH-PCB congeners and the experimental values of their agonistic activities to constitutive androstane receptor (Table S1 in the Supplementary Information). The agonistic activity of each OH-PCB congener in the dataset (reported as tenfold effective concentration) was measured in a reporter gene assay using yeast cells transduced with human constitutive androstane receptor [50]. This agonistic activity ($EC \times 10$) was defined as the concentration of OH-PCB in solution that produced luminescence intensity that was 10 times greater than the luminescence intensity of a blank solution [50]. Before being used as response variable in the model building step, the original activity values were converted to logarithmic scale ($log1/(EC \times 10)$). Using semi-empirical PM6 optimized molecular structures obtained with Spartan '14 program (version 1.1.4) as input [57], a total of 1875 one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D) molecular descriptors, representing the numerical information encoded within the molecular structure of each chemical compound in the dataset, were calculated using PaDEL-Descriptor software [58].

Variable elimination, molecular descriptor standardization, and dataset division

Existence of irrelevant and redundant molecular descriptors in multiple regression modeling is problematic and must

Table 1 Statistical parameters for fitting, cross-validation, and Y-randomization

Statistical parameter	Definition	Equation and terms	Equation number
R^2	Coefficient of multiple determination	$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$ $MSS = \sum_i \left(\widehat{y}_i - \overline{y}\right)^2$ $TSS = \sum_i (y_i - \overline{y})^2$ $RSS = \sum_i (y_i - \widehat{y}_i)^2$ y_i = observed dependent variable $\widehat{y}_i$ = calculated dependent variable $\overline{y}$ = mean value of the dependent variable	(1)
R^2_{adj}	Adjusted R^2	$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$ $= 1 - (1 - R^2)((n-1)/(n-p))$ n = number of objects p = number of predictor variables	(2)
s	Standard error of estimate	$s = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-p-1}}$ Symbols as above	(3)
Q^2_{LOO}	Explained variance in prediction	$Q^2_{LOO} = 1 - \frac{PRESS_{CV}}{TSS}$ $PRESS_{CV} = \sum_{i=1}^{n} (y_i - \widehat{y}_{i/i})^2$ $\widehat{y}_{i/i}$ = predicted value of the response calculated excluding the ith element from the model computation	(4)
$RMSE_{CV}$	Root-mean-square error in CV prediction	$RMSE_{CV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y}_{i/i})^2}{n}}$ Symbols as above	(5)
cR^2_p	Y-randomization parameter	$cR^2_p = R * (R^2 - (AverageR_r)^2)^{1/2}$	(6)

Table 2 Statistical parameters for external validation

Statistical parameter	Definition	Equation and terms	Equation number
R^2_{EXT}	Coefficient of determination for the prediction set	$R^2_{EXT} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$ y_i = observed dependent variable of test set $\widehat{y}_i$ = predicted dependent variable of test set $\overline{y}$ = mean value of the dependent variable of training set	(7)
R^2_0	Coefficient of determination for the prediction set on forcing to origin	$R^2_0 = 1 - \frac{\sum_i (y_i - \widehat{y}^{r0}_i)^2}{\sum_i (y_i - \overline{y})^2}$ $\widehat{y}^{r0}_i = ky_i$ k = slope	(8)
R'^2_0	Coefficient of determination for the prediction set on interchanging the axes and forcing to origin	$R'^2_0 = 1 - \frac{\sum_i (y_i - \widehat{y}^{r0}_i)^2}{\sum_i (y_i - \overline{y})^2}$ Symbols as above	(9)
k	Slope of the regression line on forcing to origin	$k = \frac{\sum_i \widehat{y}_i y_i}{\sum_i y_i^2}$ Symbols as above	(10)
Q^2_{F1}	Variance explained in external prediction	$Q^2_{F1} = 1 - \frac{PRESS_{EXT}}{SS_{EXT(\overline{y}_{TR})}}$ $SS_{EXT(\overline{y}_{TR})} = \sum_{i=1}^{next} (y_i - \overline{y}_{TR})^2$ $\overline{y}_{TR}$ = average of training observed responses	(11)
Q^2_{F2}	Variance explained in external prediction	$Q^2_{F2} = 1 - \frac{PRESS_{EXT}}{SS_{EXT(\overline{y}_{EXT})}}$ $SS_{EXT(\overline{y}_{EXT})} = \sum_{i=1}^{next} (y_i - \overline{y}_{EXT})^2$ $\overline{y}_{EXT}$ = average of external observed responses	(12)
$\overline{r^2_m}$	r^2_m metrics	$\overline{r^2_m} = (r^2_m + r'^2_m)/2$ r^2 = squared correlation value between the observed and predicted values with intercept r^2_0 = squared correlation value between the observed and predicted values without intercept r'^2_0 = squared correlation value between the observed and predicted values on interchanging the axes and without intercept $r^2_m = r^2 \times (1 - \sqrt{r^2 - r^2_0})$ $r'^2_m = r^2 \times (1 - \sqrt{r^2 - r'^2_0})$	(13)
Δr^2_m	r^2_m metrics	$\Delta r^2_m = \lvert r^2_m - r'^2_m \rvert$ Symbols as above	(14)
$RMSE_{EXT}$	Root-mean-square error in external prediction	$RMSE_{EXT} = \sqrt{\frac{\sum (y_i - \widehat{y}_i)^2}{n}}$ Symbols as above	(15)
MAE_{EXT}	Mean absolute error in external prediction	$MAE_{EXT} = \frac{\lvert y_i - \widehat{y}_i \rvert}{n}$ Symbols as above	(16)

therefore be eliminated [59]. Molecular descriptors with constant or nearly constant values (variables with low variance) were removed from the descriptors set because these variables are considered irrelevant for model building. In this paper, a descriptor is considered to have constant or nearly constant values if its variance is less than 0.0001. Multicollinearity among molecular descriptors introduces redundancy in a QSAR model since highly correlated descriptors contribute essentially the same information in the model [60]. In this paper, two molecular descriptors are considered redundant if the correlation coefficient between them exceeds 0.90. Multicolinear and low-variance descriptors were removed from the pool of 1875 descriptors calculated in the preceding section using V-WSP algorithm [61] as implemented in V-WSP tool (version 1.2) developed by Ambure et al. [62]. Pairwise correlations in a correlation matrix and variance inflation factor (VIF) were calculated and used to examine the presence or absence of multicollinearity among variables utilized in the final QSAR model [59]. After removing irrelevant and redundant descriptors from the descriptors set, the remaining descriptors were transformed into auto-scaled descriptors using standard normalization method [63]. Standardization of the original descriptors to auto-scaled descriptors (X_{ik}^{n}) was accomplished by subtracting the mean of the descriptors (μ_k) from the original descriptor values (X_{ik}) and then divided by their standard deviations (σ_K). The main reason for using standardized regression coefficients in a QSAR model is that the magnitude of the standardized regression coefficient reflects the relative contribution of each descriptor to the predicted activity in the developed model [64]. Transformation of the original descriptor to auto-scaled descriptors was done using Minitab® 18.1 [65]. Finally, the entire dataset was divided into training set (70% of the entire dataset) and test set (30% of the entire dataset) using Kennard-Stone algorithm [66–68] as implemented in Dataset Division GUI 1.2 [62]. This corresponds to 44 OH-PCBs in the training set and 20 OH-PCBs in the test set. In Table S1 (Supplementary Information), compounds assigned to the test set are marked with single asterisk to distinguish it from compounds assigned to the training set.

Variable selection and model building

Selection of predictor variables for inclusion in a regression model is one of the most crucial aspects of a QSAR study. In this paper, genetic function approximation (GFA) [54], as implemented in Materials Studio 7.0 [69], was used to select the most appropriate combination of molecular descriptors for inclusion in the QSAR model developed in this study. This was then followed by a comprehensive regression analysis using multiple linear regression (MLR) method [70] as implemented in Minitab® 18.1 [65] and Materials Studio 7.0 [69]. In this model building step, efforts were made to establish a linear relationship between the dependent variable ($log1/(EC \times 10)$) and the selected independent variables (standardized molecular descriptors) of the series of chemical compounds assigned to the training set in Table S1 (Supplementary Information).

Internal and external validation of QSAR model

The goodness-of-fit of the QSAR model developed from the application of the GFA and MLR procedures described in the preceding section was evaluated using the following statistical parameters: R^2, R^2_{adj}, and s in Eqs. (1), (2), and (3), respectively. The stability and robustness of the developed QSAR model were evaluated using leave-one-out cross-validation and Y-randomization techniques [71]. In the leave-one-out cross-validation technique, one chemical compound from the training set was omitted and a model was developed using the data of the remaining chemical compounds. The model developed was then used to predict the agonistic activity of the omitted compound. This procedure was repeated iteratively and values of Q^2_{LOO} and $RMSE_{CV}$ in Eqs. (4) and (5) respectively were calculated. In the Y-randomization procedure, the dependent variable of compounds in the training set was randomly shuffled 50 times while keeping the independent variables as they are. New MLR model was then developed after each random shuffling. Statistical parameters for the generated random models (R, R^2, and Q^2 values for random models) and the Y-randomization parameter, cR^2_p in Eq. (6), were calculated. All the statistical parameters used for evaluating the fitness and robustness of the developed QSAR model were calculated using Minitab® 18.1 [65], Materials Studio 7.0 [69], and Y-randomization tool 1.2 [62]. In addition to the statistical parameters listed in Table 1 for evaluating the goodness-of-fit and robustness of the developed model, the QSAR model constructed in this paper was also externally validated for its ability to predict the agonistic activities of untested OH-PCB congeners. This task was accom-

Table 3 Descriptions, classes, and types of molecular descriptors utilized in the developed QSAR model

Descriptor	Description	Type	Class
MATS1s	Moran autocorrelation - lag 1/weighted by I-state	2D	Autocorrelation descriptor
VE3_DzZ	Logarithmic coefficient sum of the last eigenvector from Barysz matrix/weighted by atomic number	2D	Barysz matrix descriptor
VE1_Dzp	Coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability	2D	Barysz matrix descriptor
SpMin8_Bhv	Smallest absolute eigenvalue of Burden modified matrix—n8/weighted by relative van der Waals volume	2D	Burden modified eigenvalues descriptor
SpMax5_Bhi	Largest absolute eigenvalue of Burden modified matrix—n5/weighted by relative first ionization potential	2D	Burden modified eigenvalues descriptor
topoRadius	Topological radius (minimum atom eccentricity)	2D	Topological descriptor
RDF95u	Radial distribution function—095/unweighted	3D	RDF descriptor
RDF45m	Radial distribution function—045/weighted by relative mass	3D	RDF descriptor

plished by using the compounds assigned to the test set in Table S1 (Supplementary information). The parameters used for external validation of the developed QSAR model (R^2_{EXT}, R^2_0, R'^2_0, k, Q^2_{F1}, Q^2_{F2}, $\overline{r^2_m}$, Δr^2_m, $RMSE_{EXT}$, and MAE_{EXT}) are displayed in Table 2. These external validation metrics were calculated using MLRPlusValidation 1.3 and XternalValidationPlus 1.2 tools developed by Ambure et al. [62].

Evaluation of model applicability domain

Applicability domain expresses the scope and limitation of a QSAR model by defining the range of chemical structures for which the QSAR model is considered applicable [72]. In this paper, the leverage approach was used to evaluate the applicability domain of the QSAR model developed in this study. A measure of how far a chemical compound is from the applicability domain of a QSAR model is its leverage in the original variable space, h_{ii} [64]. This measure is defined as: $h_{ii} = x_i^T(X^TX)^{-1}x_i$, where x_i is the descriptor row-vector of the query compound, and X is the $n \times p$ matrix of p model descriptor values for n training set compounds. The superscript T refers to the transpose of the matrix vector. The warning leverage h^* is generally fixed at $3k/n$, where n is the number of training compounds and k is the number of model descriptors plus one $(p+1)$ [64]. To visualize the applicability domain of the developed model, a plot of standardized residuals versus leverages (Williams plot) was constructed.

Results and discussion

The QSAR model obtained on correlating the agonistic activities ($log\ 1/(EC \times 10)$ values) of 44 OH-PCB congeners assigned to the training set in Table S1 (Supplementary information) with the standardized molecular descriptors computed for these 44 OH-PCB congeners (Table S2 in the Supplementary information) is displayed in Eq. (17). As shown in Eq. (17), a linear relationship was established between $log\ 1/(EC \times 10)$ and eight structural features of OH-PCB congeners. Table 3 shows the descriptions, classes, and types of the molecular descriptors utilized in the developed QSAR model. As shown in Table 3, the QSAR model displayed in Eq. (17) contained six 2D descriptors (MATS1s, VE3_DzZ, VE1_Dzp, SpMin8_Bhv, SpMax5_Bhi, and topoRadius), belonging to four classes of descriptors—autocorrelation descriptor, Barysz matrix descriptor, Burden modified eigenvalues descriptor, and topological descriptor [73]. Table 3 also shows that the two 3D descriptors (RDF95u and RDF45m) in the QSAR model displayed in Eq. (17) belong to the same class of descriptor—radial distribution function (RDF) descriptor [73].

Table 4 Regression coefficients and statistical significance of predictor variables utilized for QSAR model development and values of metrics used for evaluating the goodness-of-fit and robustness of the developed QSAR model

Descriptor	Regression coefficient		Statistical significance		VIF
	Value	Standard error	*t*-value	*p* value	
Constant	6.1425	0.0453	135.480	0.000	
MATS1s	0.1698	0.0741	2.291	0.028	1.134
VE3_DzZ	−0.1062	0.0497	−2.138	0.040	1.268
VE1_Dzp	0.9477	0.0638	14.852	0.000	1.929
SpMin8_Bhv	−0.2241	0.0769	−2.914	0.006	3.177
SpMax5_Bhi	−0.8122	0.1271	−6.392	0.000	7.901
topoRadius	−0.3479	0.0669	−5.197	0.000	2.350
RDF95u	−0.8045	0.0744	−10.811	0.000	3.790
RDF45m	0.4608	0.0678	6.795	0.000	2.769

$N=44$, $R^2 = 0.8902$, $R^2_{adj} = 0.8651$, $s=0.2840$, $Q^2_{LOO} = 0.8201$, $RMSE_{CV} = 0.3242$

$$\begin{aligned} \log 1/(EC \times 10) = 6.1425 &+ 0.1698\ \mathrm{MATS1s} \\ &- 0.1062\ \mathit{VE3\_DzZ} + 0.9477\ \mathrm{VE1\_Dzp} \\ &- 0.2241\ \mathrm{SpMin8\_Bhv} - 0.8122\ \mathrm{SpMax5\_Bhi} \\ &- 0.3479\ \mathrm{topoRadius} - 0.8045\ \mathrm{RDF95u} \\ &+ 0.4608\ \mathrm{RDF45m} \end{aligned} \tag{17}$$

In Table 4, the standardized regression coefficients and the statistical significance of the predictor variables utilized in the QSAR model displayed in Eq. (17)are presented. By

Table 5 Correlation matrix of the molecular descriptors utilized for QSAR model development*

	MATS1s	VE3_DzZ	VE1_Dzp	SpMin8_Bhv	SpMax5_Bhi	topoRadius	RDF95u
VE3_DzZ	−0.067						
VE1_Dzp	−0.132	0.316					
SpMin8_Bhv	−0.020	0.089	0.000				
SpMax5_Bhi	−0.073	−0.196	−0.087	−0.794			
topoRadius	−0.084	−0.002	0.117	0.139	−0.221		
RDF95u	−0.049	0.315	0.564	0.295	−0.452	−0.188	
RDF45m	−0.021	0.142	0.267	−0.332	0.473	0.208	0.099

*Numerical values in the cells are correlation coefficients examining the magnitudes and signs of the regression coefficients in the QSAR model, a broad interpretation of the model can be made [74]. While the descriptor that plays the most important role in the predictive ability of a model is identified by the magnitude of its regression coefficient in the model, the sign of this regression coefficient indicates the direction of the relationship between the descriptor and the predicted activity [74]. As shown in Table 4, the relative contributions of the eight descriptors in Eq. (17) to the predictive ability of the developed QSAR model, as can be inferred from the magnitudes of the standardized regression coefficients of the descriptors, decreased in the following order: VE1_Dzp > SpMax5_Bhi > RDF95u > RDF45m > topoRadius > SpMin8_Bhv > MATS1s > VE3_DzZ. Table 4 also shows that while three of the eight descriptors in Eq. (17) (VE1_Dzp, RDF45m, and MATS1s) made positive contribution to agonistic activities of OH-PCBs, five of the eight descriptors (VE3_DzZ, SpMin8_Bhv, topoRadius, RDF95u, and SpMax5_Bhi) made negative contribution to agonistic activities of OH-PCBs. The positive contributions of VE1_Dzp, RDF45m, and MATS1s to agonistic activities of OH-PCBs indicate that OH-PCB congener with higher values of these descriptors would be more active than OH-PCB congener with lower values of these descriptors. Conversely, the negative contributions of VE3_DzZ, SpMin8_Bhv, topoRadius, RDF95u, and SpMax5_Bhi to agonistic activities of OH-PCBs indicate that OH-PCB congener with higher values of these descriptors would be less active than OH-PCB congener with lower values of these descriptors. Broad interpretations of QSAR models, consistent with the approach used in this paper, are well documented in the literature [75–77]. The p values shown in Table 4 are all less than 0.05, indicating that the relationship between $log\ 1/(EC \times 10)$ and each of the eight descriptors in the QSAR model displayed in Eq. (17) was statistically significant [78]. This indicates that the strength of the association between $log\ 1/(EC \times 10)$ and each of the eight molecular descriptors is strong and reliable.

Variance inflation factor (VIF) and correlation matrix—two approaches used for detecting multicollinearity among molecular descriptors in a QSAR model—are presented in Tables 4 and 5, respectively. Values of VIF reported in Table 4 for the eight descriptors are 1.134 for MATS1s, 1.268 for VE3_DzZ, 1.929 for VE1_Dzp, 3.177 for SpMin8_Bhv, 7.901 for SpMax5_Bhi, 2.350 for topoRadius, 3.790 for RDF95u, and 2.769 for RDF45m. A value of VIF less than 10 obtained for each of the eight molecular descriptors indicates that the QSAR model displayed in Eq. (17) contains no multicollinearity [79]. In the correlation matrix shown in Table 5, the absolute values of the correlation coefficients ranged from 0.000 (between VE1_Dzp and SpMin8_Bhv) to 0.798 (between SpMin8_Bhv and SpMax5_Bhi). Although no generally agreed cut-off value is currently available in literature for defining lack of collinearity between two descriptors, a

Table 6 Results of Y-randomization test to check the robustness of the developed QSAR model

Model	R	R^2	Q^2
Original	0.943524	0.890237	0.820137
Random 1	0.392217	0.153834	−1.706501
Random 2	0.295415	0.087270	−0.572871
Random 3	0.555079	0.308112	−0.036926
Random 4	0.560575	0.314244	−0.150843
Random 5	0.586008	0.343406	0.005219
Random 6	0.419719	0.176164	−0.996193
Random 7	0.377767	0.142708	−0.720482
Random 8	0.434776	0.189030	−1.691117
Random 9	0.389797	0.151942	−2.734291
Random 10	0.333463	0.111198	−2.586910
Random 11	0.468751	0.219727	−6.208023
Random 12	0.391702	0.153431	−0.280631
Random 13	0.541581	0.293310	−1.298111
Random 14	0.390383	0.152399	−0.614923
Random 15	0.430798	0.185587	−1.081088
Random 16	0.517962	0.268285	−0.154168
Random 17	0.472139	0.222916	−6.215898
Random 18	0.409741	0.167888	−0.243678
Random 19	0.342970	0.117628	−1.211103
Random 20	0.372416	0.138694	−3.878376
Random 21	0.436974	0.190946	−1.143880
Random 22	0.425312	0.180890	−1.887534
Random 23	0.262843	0.069087	−0.543526
Random 24	0.465107	0.216324	−0.453787
Random 25	0.448195	0.200879	−2.382542
Random 26	0.344701	0.118819	−6.130955
Random 27	0.353124	0.124697	−1.447859
Random 28	0.460358	0.211929	−0.471742
Random 29	0.322496	0.104004	−0.849685
Random 30	0.471075	0.221912	−0.900631
Random 31	0.327244	0.107089	−0.441135
Random 32	0.508110	0.258176	−1.114144
Random 33	0.164240	0.026975	−1.960059
Random 34	0.263265	0.069308	−2.584566
Random 35	0.338144	0.114341	−0.514544
Random 36	0.397330	0.157871	−0.255736
Random 37	0.526550	0.277255	−0.284235
Random 38	0.306055	0.093670	−3.957742
Random 39	0.388422	0.150871	−0.380122
Random 40	0.540108	0.291716	−0.202617
Random 41	0.264492	0.069956	−1.107566
Random 42	0.416236	0.173253	−0.350605
Random 43	0.363125	0.131860	−0.628484
Random 44	0.612895	0.375640	−0.267751
Random 45	0.308254	0.095021	−0.432622
Random 46	0.417994	0.174719	−1.464286
Random 47	0.373363	0.139400	−0.580706
Random 48	0.254558	0.064800	−0.995991
Random 49	0.529422	0.280288	−0.098344
Random 50	0.512676	0.262837	−0.179581

Average R = 0.409719, average R^2 = 0.177046, average Q^2 = −1.327798, cR_p^2 = 0.801922

Table 7 External validation parameters for the developed QSAR model and their threshold values

Parameter	Value for developed model		Threshold value	Remark
	100% data	95% data		
R^2_{EXT}	0.7625	0.7651	$R^2 > 0.6$	Reliably predicted
$\lvert R^2_0 - R'^2_0 \rvert$	0.0593	0.0184	$\lvert R^2_0 - R'^2_0 \rvert < 0.3$	Reliably predicted
$(R^2 - R^2_0)/R^2$	0.0782	0.0618	$(R^2 - R^2_0)/R^2 < 0.1$	Reliably predicted
k	0.9552	0.9610	$0.85 \leq k \leq 1.15$	Reliably predicted
Q^2_{F1}	0.6134	0.6583	$Q^2_{F1} > 0.6$	Reliably predicted
Q^2_{F2}	0.6057	0.6479	$Q^2_{F2} > 0.6$	Reliably predicted
$\overline{r^2_m}$	0.6904	0.6950	$\overline{r^2_m} > 0.5$	Reliably predicted
Δr^2_m	0.0782	0.0819	$\Delta r^2_m < 0.2$	Reliably predicted
$RMSE_{EXT}$	0.5700	0.5457	Low	Reliably predicted
MAE_{EXT}	0.4898	0.4673	Low	Reliably predicted

squared correlation coefficient lower than 0.80 has been suggested as a threshold value for accepting lack of collinearity between two variables [63]. Detraction of QSAR model from mechanistic interpretation and deterioration of the model's statistical parameters are the two major problems encountered when multicollinearity exists among the descriptors utilized in a QSAR model [60, 80].

The values of coefficient of multiple determination (R^2) and standard error of estimate (s)—two statistical metrics used for evaluating the goodness-of-fit of a QSAR model - are shown in Table 4. The values of R^2 and s reported in Table 4 were 0.8902 and 0.2840, respectively. This R^2 value indicates that 89.02% of total variation in the response variable ($log1/(EC \times 10)$) could be accounted for by its relationship with the eight predictor variables (molecular descriptors) utilized in the developed QSAR model. Value of $R^2 >$ 0.6, as obtained in this paper, suggests that the QSAR model displayed in Eq. (17) provides a good fit to the data used in

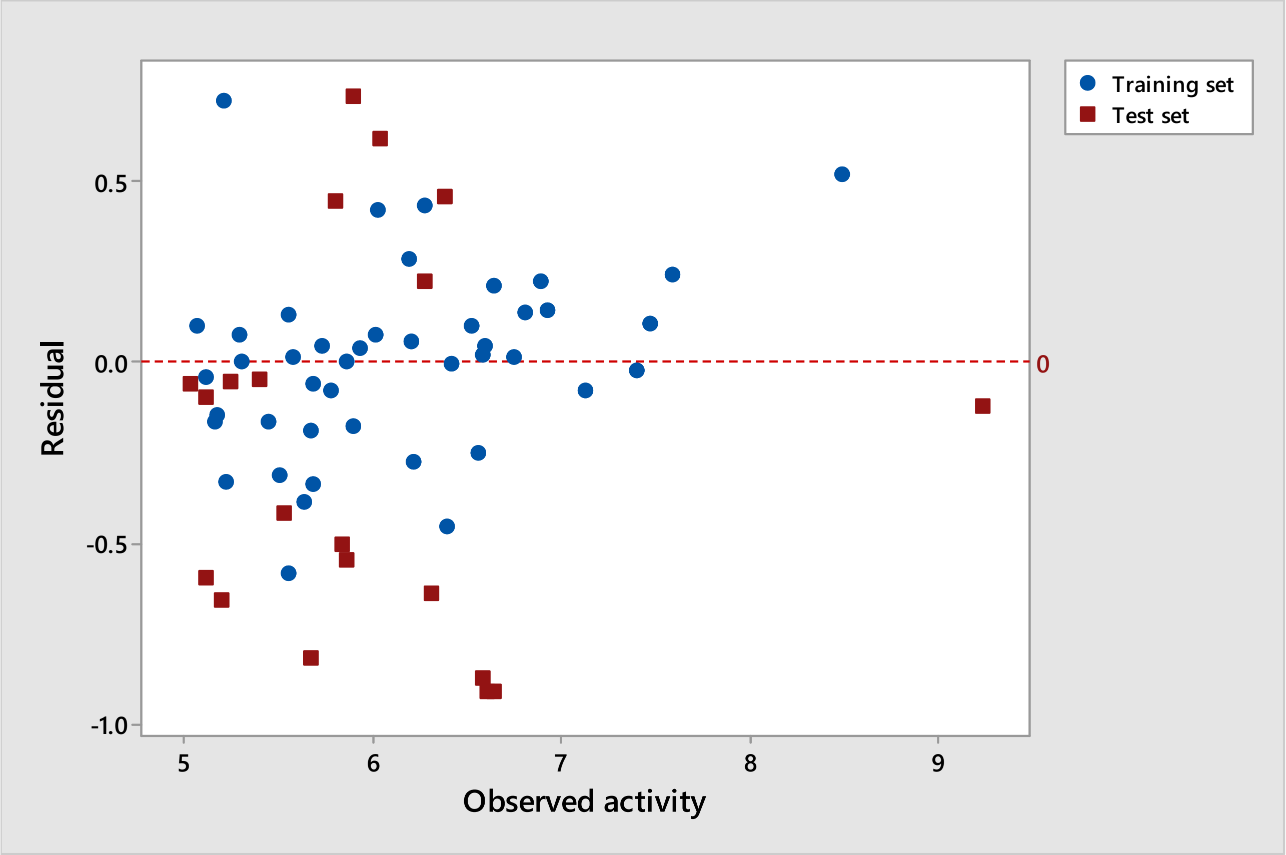

Fig. 2 Plot of residuals versus observed activities

building the model [81]. The standard error of estimate (*s*) measures the dispersion of the observed values from the regression line [64]. The low value of *s* reported in Table 4 indicates that the observed values of the agonistic activities of OH-PCBs are close to the regression line predicted by the developed QSAR model. Another statistical parameter that is used as a measure of goodness-of-fit of a QSAR model is the adjusted R^2 (R^2_{adj}). Unlike the value of R^2 which always increases regardless of whether the addition of an extra predictor variable to a QSAR model improves the model or not, value of R^2_{adj} only increases when addition of an extra predictor variable improves the model, thus eliminating the possibility of overfitting [82]. The value of R^2_{adj} reported in Table 4 for the QSAR model displayed in Eq. (17) was 0.8651. The eight predictor variables retained in the QSAR model are considered acceptable because further addition of extra independent variable to the model caused significant reduction in the reported R^2_{adj} value [82]. Furthermore, since the ratio of the number of compounds in the training set to the number of descriptors in the QSAR model displayed in Eq. (17) was at least 5:1 as suggested by Topliss and Costello [83], the risk of chance correlation in the developed QSAR model was avoided.

To assess the stability and robustness of the QSAR model displayed in Eq. (17), parameters obtained from leave-one-out cross-validation and Y-randomization techniques were used. In Table 4, values of Q^2_{LOO} and $RMSE_{CV}$—two parameters obtained from leave-one-out cross-validation technique—are presented. The QSAR model displayed in Eq. (17) is considered robust and stable because the value of Q^2_{LOO} reported in Table 4 ($Q^2_{LOO} = 0.8201$) was greater than the cut-off value ($Q^2_{LOO} > 0.5$) suggested in the literature [71, 84]. Furthermore, a difference of less than 0.3 between R^2 value and Q^2_{LOO} value, as obtained in this study, indicates that the QSAR model displayed in Eq. (17) did not suffer from overfitting [84]. The low value of root-mean-square error in cross-validation ($RMSE_{CV} = 0.3242$) reported in Table 4 suggests that the accuracy of the prediction made by the developed model was good [85]. The results of Y-randomization procedure conducted to check the stability and robustness of the QSAR model developed in this paper are shown in Table 6. The R^2 and Q^2 values reported in Table 6 for each of the 50 random models generated were significantly lower than the values of R^2 and Q^2 obtained for the original QSAR model displayed in Eq. (17) (R^2 and Q^2 values obtained for the original model were 0.8902 and 0.8201, respectively).

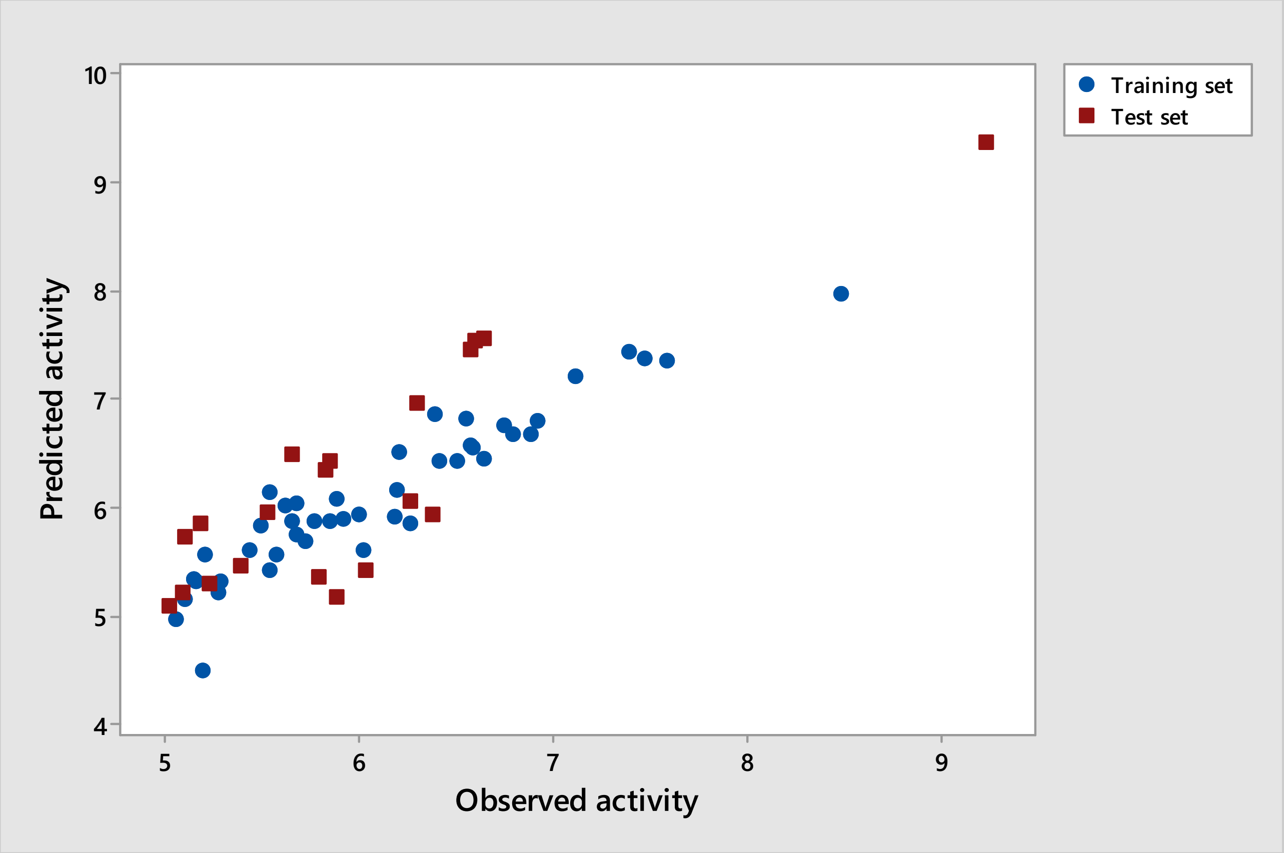

Fig. 3 Plot of predicted activities versus observed activities

The lower values of R^2 (average was 0.1770) and Q^2 (average was -1.3278) obtained for the random models ruled out the possibility of chance correlation in the original QSAR model displayed in Eq. (17) [86]. The value of cR_p^2 reported in Table 6 was 0.8019. Value of cR_p^2 higher than 0.5, as obtained in this study, indicates that the QSAR model displayed in Eq. (17) was robust and stable [86].

One of the core functions of a QSAR model is to make reliable prediction of biological activities for yet to be tested chemical compounds. Because generating an entirely new set of experimental data for the purpose of external validation is often difficult in everyday practice, the strategy suggested by QSAR experts is to divide the available dataset into two subsets—the training set and the test set [64]. In this paper, the 44 OH-PCB congeners assigned to the training set (70% of the entire dataset) were used for model building while the remaining 20 OH-PCB congeners assigned to the test set (30% of the entire dataset) were reserved for external validation of the developed model. Using the computed molecular descriptors presented in Table S3 (Supplementary information) as inputs, agonistic activities of OH-PCBs in the test set were predicted with the QSAR model displayed in Eq. (17). The values of the external validation metrics used for evaluating the predictive ability of the developed QSAR model and the threshold values used for judging the acceptability of these metrics are presented in Table 7. As shown in Table 7, the values of the external validation metrics obtained when prediction was made on chemical compounds assigned to the test set, before and after removing 5% of data with high residuals from the test set, indicate that all the validation metrics satisfied the requirements for accepting the external predictivity of the QSAR model displayed in Eq. (17) [71, 84, 87–90]. Using the values of observed activities, predicted activities and residuals shown in Tables S4 and S5 (Supplementary information), a plot of residuals versus experimental values of $log\ 1/(EC \times 10)$ (Fig. 2) and a plot of predicted values of $log\ 1/(EC \times 10)$ versus experimental values of $log\ 1/(EC \times 10)$ (Fig. 3) were constructed. The random distribution of the residuals on both sides of the horizontal zero line in Fig. 2 indicates that there was no systematic error in the QSAR model developed in this paper. Figure 3 also shows that there was a positive and strong correlation between the activities predicted by the developed QSAR model and the experimental activities obtained from literature. These results indicate that the prediction made by the QSAR model displayed in Eq. (17) was accurate and reliable.

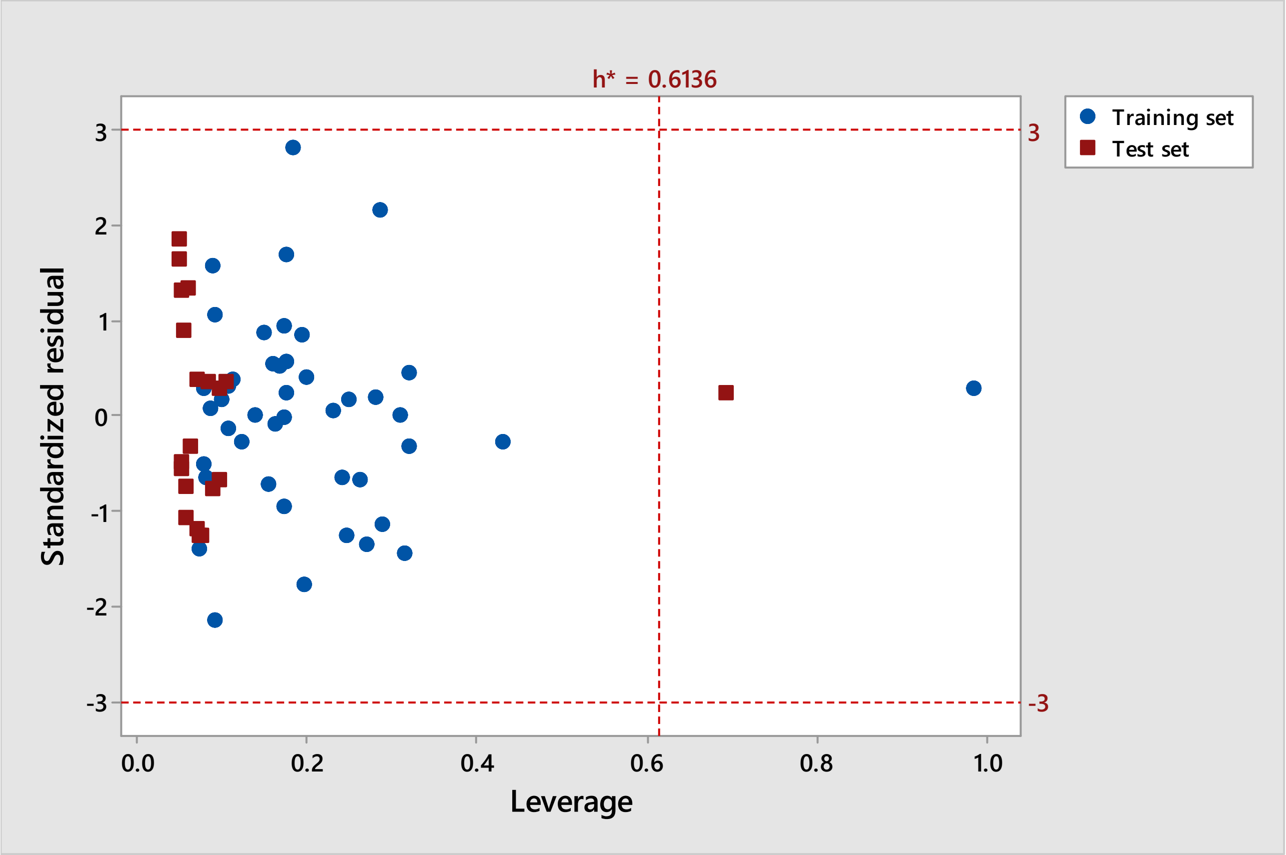

Fig. 4 Plot of standardized residuals versus leverages (Williams plot), with a warning leverage of 0.6136

Applicability domain—a concept that expresses the scope and limitation of a QSAR model by defining the range of chemical structures for which the model is considered applicable—is another crucial aspect that must be included in a QSAR study [72]. To visualize the outliers in both the descriptor space and the response space, the Williams plot shown in Fig. 4 was constructed from the values of standardized residuals and leverages presented in Tables S4 and S5 (Supplementary information). In the Williams plot displayed in Fig. 4, a compound with a standardized residual greater than three standard deviation units is considered to be a response outlier while a compound with leverage higher than the warning leverage ($h^* = 0.6136$) is considered to be a structurally influential chemical in the developed model [88]. As shown in Fig. 4, none of the 64 OH-PCBs in the dataset was response outlier but 2′,4′,6′-trichlorobiphenyl-4-ol and 2′,3,4′-trichlorobiphenyl-2-ol were found to be structurally influential chemicals because of their high leverage values. As structurally influential chemical in the training set, the compound 2′,3,4′-trichlorobiphenyl-2-ol greatly influenced the regression parameters by forcing the fitted regression line near its observed value [64]. On the other hand, the predicted response of a compound in the test set (compound 2′,4′,6′-trichlorobiphenyl-4-ol in this case) with a leverage value greater than the warning leverage may not be reliable because the prediction was probably a result of substantial extrapolation of the QSAR model [64].

Conclusion

Some congeners of hydroxylated polychlorinated biphenyls are known to cause endocrine disruption in humans by acting as nuclear receptor agonists or nuclear receptor antagonists. However, the high numbers of inactive OH-PCB congeners recorded in many experimental toxicity studies designed to measure the agonistic and antagonistic activities of OH-PCBs necessitate the need to develop QSAR models that can predict the activities of OH-PCBs prior to in vitro experiments. Use of QSAR models for large-scale screening of OH-PCB congeners for the purpose of prioritizing the congeners for further experimental testing offers the advantages of being rapid, inexpensive, and high-throughput. In this paper, a local QSAR model for predicting the agonistic activities of OH-PCBs to constitutive androstane receptor was constructed and validated using certain internal and external validation criteria. The developed QSAR model was found to be statistically reliable, robust, and possessed good external predictivity. The QSAR model developed in this paper can therefore be used by toxicologists to predict the agonistic activities of untested OH-PCBs to constitutive androstane receptor prior to experimental studies. Development of more QSAR models for the prediction of OH-PCB activities in toxicity studies involving other nuclear receptors is recommended.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11224-022-01992-2.

Author contribution All authors contributed to the study conception and design. Data collection and analysis were performed by Lukman K. Akinola. The first draft of the manuscript was written by Lukman K. Akinola and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability The supplementary data associated with this article can be found in the online version of the paper.

Code availability Not applicable.

Declarations

Competing interests The authors declare no competing interests.

References

1. Weikum ER, Liu X, Ortlund EA (2018) The nuclear receptor superfamily: a structural perspective. Protein Sci 27:1876–1892
2. Robinson-Rechavi M, Garcia HE, Laudet V (2003) The nuclear receptor superfamily. J Cell Sci 116:585–586
3. Huang P, Chandra V, Rastinejad F (2010) Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. Annu Rev Physiol 72:247–272
4. Weatherman RV, Fletterick RJ, Scanlan TS (1999) Nuclear-receptor ligands and ligand-binding domains. Annu Rev Biochem 68:559–581
5. Steinmetz ACU, Renaud JP, Moras D (2001) Binding of ligands and activation of transcription by nuclear receptors. Annu Rev Biophys Biomol Struct 30:329–359
6. Bain DL, Heneghan AF, Connaghan-Jones KD, Miura MT (2007) Nuclear receptor structure: implications for function. Annu Rev Physiol 69:201–220
7. Toporova L, Balaguer P (2020) Nuclear receptors are the major targets of endocrine disrupting chemicals. Mol Cell Endocrinol 502:110665
8. Caliman FA, Gavrilescu M (2009) Pharmaceuticals, personal care products and endocrine disrupting agents in the environment – a review. Clean: Soil, Air, Water 37:277–303
9. Zhang Z, Jia C, Hu Y, Sun L, Jiao J, Zhao L, Zhu D, Li J, Tian Y, Bai H, Li R, Hu J (2012) The estrogenic potential of salicylate esters and their possible risks in foods and cosmetics. Toxicol Lett 209(2):146–153
10. Muncke J (2009) Exposure of endocrine disrupting compounds via the food chain: is packaging a relevant source? Sci Total Environ 407(16):4549–4559
11. Koo HJ, Lee BM (2004) Estimated exposure to phthalates in cosmetics and risk assessment. J Toxicol Environ Health Part A 67:1901–1914
12. Rudel RA, Camann DE, Spengler JD, Korn LR, Brody JG (2003) Phthalates, alkylphenols, pesticides, polybrominated diphenyl ethers, other endocrine-disrupting compounds in indoor air and dust. Environ Sci Technol 37:4543–4553
13. Darbre PD (2018) Overview of air pollution and endocrine disorders. Int J Gen Med 11:191–207
14. Pojana G, Gomiero A, Jonkers N, Marcomini A (2007) Natural and synthetic endocrine disrupting compounds (EDCs) in water, sediment and biota of a coastal lagoon. Environ Int 33(7):929–936

15. Rhind SM, Kyle CE, Ruffie H, Calmettes E, Osprey M, Zhang ZL, Hamilton D, McKenzie C (2013) Short and long-term temporal changes in soil concentrations of selected endocrine disrupting compounds (EDCs) following single or multiple applications of sewage sludge to pastures. Environ Pollut 181:262–270
16. Archer E, Wolfaardt GM, van Wyk JH (2017) Pharmaceutical and personal care products (PPCPs) as endocrine disrupting contaminants (EDCs) in South African surface waters. Water SA 43:684–706
17. Loffredo E, Senesi N (2006) Fate of anthropogenic organic pollutants in soils with emphasis on adsorption/desorption processes of endocrine disruptor compounds. Pure Appl Chem 78(5):947–961
18. Zamkowska D, Karwacka A, Jurewicz J, Radwan M (2018) Environmental exposure to non-persistent endocrine disrupting chemicals and semen quality: an overview of the current epidemiological evidence. Int J Occup Med Environ Health 31(4):377–414
19. Rehman S, Usman Z, Rehman S, AlDraihem M, Rehman N, Rehman I, Ahmad G (2018) Endocrine disrupting chemicals and impact on male reproductive health. Transl Androl Urol 7(3):490–503
20. Rahban R, Nef S (2020) Regional difference in semen quality of young men: a review on the implication of environmental and lifestyle factors during fetal life and adulthood. Basic Clin Androl 30:16
21. Donato MD, Cernera G, Giovannelli P, Galasso G, Bilancio A, Migliaccio A, Castoria G (2017) Recent advances on bisphenol-A and endocrine disruptor effects on human prostate cancer. Mol Cell Endocrinol 457:35–42
22. Lacouture A, Lafront C, Peillex C, Pelletier M, Audet-Walsh E (2022) Impacts of endocrine-disrupting chemicals on prostate function and cancer. Environ Res 204:112085
23. Singh VK, Pal R, Srivastava P, Mistra G, Shukla Y, Sharma PK (2021) Exposure of androgen mimicking environmental chemicals enhances proliferation of prostate cancer (LNCaP) cells by inducing AR expression and epigenetic modification. Environ Pollut 272:116397
24. Hess-Wilson JK, Knudsen KE (2006) Endocrine disrupting compounds and prostate cancer. Cancer Lett 241:1–12
25. Fernandez MF, Olea N (2012) In: Diamanti-Kandarakis E and Gore AC (ed) Endocrine disruptors and puberty. Humana Press, New York
26. Spinder N, Bergman JEH, van Tongeren M, Boezen HM, Kromhout H, de Walle HEK (2022) Maternal occupational exposure to endocrine-disrupting chemicals and urogenital anomalies in the offspring. Hum Reprod 37(1):142–151
27. Yum T, Lee S, Kim Y (2013) Association between precocius puberty and some endocrine disruptors in human plasma. J Environ Sci Health A 48:912–917
28. Lucaccioni L, Trevisani V, Marrozzini L, Bertoncelli N, Predieri B, Lugli L, Berardi A, Lughetti L (2020) Endocrine-disrupting chemicals and their effects during female puberty: a review of current evidence. Int J Mol Sci 21:2078
29. Srilanchakon K, Thadsri T, Jantarat C, Thengyai S, Nosoognoen W, Supornsilchai V (2017) Higher phthalate concentrations are associated with precocious puberty in normal weight Thai girls. J Pediatr Endocrinol Metab 30:1293–1298
30. Varnell RR, Arnold TJ, Quandt SA, Talton JW, Chen H, Miles CM, Daniel SS, Sandberg JC, Anderson KA, Arcury TA (2021) Menstrual cycle patterns and irregularities in hired Latinx child farmworkers. J Occup Environ Med 63(1):38–43
31. Grindler NM, Allsworth JE, Macones GA, Kannan K, Roehl KA, Cooper AR (2015) Persistent organic pollutants and early menopause in U.S. women. PLoS One 10(1):e0116057
32. Scsukova S, Rollerova E, Mlynarcikova AB (2016) Impact of endocrine disrupting chemicals on the onset and development of female reproductive disorders and hormone-related cancer. Reprod Biol 16(4):243–254
33. Wam MLY, Co VA, El-Nezami H (2021) Endocrine disrupting chemicals and breast cancer: a systematic review of epidemiological studies. Crit Rev Food Sci Nutr. https://doi.org/10.1080/10408398.2021.1903382
34. Montes-Grajales D, Bernardes GJL, Olivero-Verbel JT (2016) Urban endocrine disruptors targeting breast cancer proteins. Chem Res Toxicol 29(2):150–161
35. Ben-Jonathan N (2019) In: Zhang X (ed) Estrogen receptor and breast cancer: celebrating the 60th anniversary of the discovery of ER. Humana Press, Switzerland
36. Rogers JA, Metz L, Yong VW (2013) Review: endocrine disrupting chemicals and immune responses: a focus on bisphenol-A and its potential mechanisms. Mol Immunol 53:421–430
37. Kuo CH, Yang SN, Kuo PL, Hung CH (2012) Immunomodulatory effects of environmental endocrine disrupting chemicals. Kaohsiung J Med Sci 28:S37–S42
38. Dietert RR (2015) In: Darbre PD (ed) Endocrine disruption and human health. Elsevier, Amsterdam
39. Sakkiah S, Wang T, Zou W, Wang Y, Pan B, Tong W, Hong H (2018) Endocrine disrupting chemicals mediated through binding androgen receptor are associated with diabetes mellitus. Int J Environ Res Public Health 15:25
40. Lind PM, Lind L (2018) Endocrine-disrupting chemicals and risk of diabetes: an evidence-based review. Diabetologia 61:1495–1502
41. Alonso-Magdalena P, Quesada I, Nadal A (2011) Endocrine disruptors in the etiology of type 2 diabetes mellitus. Nat Rev Endocrinol 7(6):346–353
42. Ehrlich S, Lambers D, Baccarelli A, Khoury J, Macaluso M, Ho SM (2016) Endocrine disruptors: a potential risk factor for gestational diabetes mellitus. Am J Perinatol 33(13):1313–1318
43. Darbre PD (2015) In: Darbre PD (ed) Endocrine disruption and human health. Elsevier, Amsterdam
44. Newbold RR, Padilla-Banks E, Jefferson WN, Heindel JJ (2008) Effects of endocrine disruptors on obesity. Int J Androl 31:201–208
45. Kim JT, Lee HK (2017) Childhood obesity and endocrine disrupting chemicals. Ann Pediatr Endocrinol Metab 22:219–225
46. Hatch EE, Nelson JW, Stahlhut RW, Webster TF (2010) Association of endocrine disruptors and obesity: perspectives from epidemiological studies. Int J Androl 33:324–332
47. Fu X, Xu J, Zhang R, Yu J (2020) The association between environmental endocrine disruptors and cardiovascular diseases: a systematic review and meta-analysis. Environ Res 187:109464
48. Tehrani R, Aken BV (2014) Hydroxylated polychlorinated biphenyls in the environment: sources, fate, and toxicity. Environ Sci Pollut Res Int 21:6334–6345
49. Takeuchi S, Shiraishi F, Kitamura S, Kuroki H, Jin K, Kojima H (2011) Characterization of steroid hormone receptor activities in 100 hydroxylated polychlorinated biphenyls, including congeners identified in humans. Toxicology 289:112–121
50. Kamata R, Shiraishi F, Kageyama S, Nakajima D (2015) Detection and measurement of the agonistic activities of PCBs and mono-hydroxylated PCBs to the constitutive androstane receptor using a recombinant yeast assay. Toxicol in Vitro 29:1859–1867
51. Kamata R, Nakajima D, Shiraishi F (2019) Measurement of the agonistic activities of monohydroxylated polychlorinated biphenyls at the retinoid X and retinoic acid receptors using recombinant yeast cells. Toxicol in Vitro 57:9–17

52. Cao LY, Ren XM, Guo LH (2019) Estrogen-related receptor γ is a novel target for lower-chlorinated polychlorinated biphenyls and their hydroxylated and sulfated metabolites. Environ Pollut 254:113088
53. Kitamura S, Jinno N, Suzuki T, Sugihara K, Ohta S, Kuroki H, Fujimoto N (2005) Thyroid hormone-like and estrogenic activity of hydroxylated PCBs in cell culture. Toxicology 208:377–387
54. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci 34:854–866
55. Kublbeck K, Niskanen J, Honkakoski P (2020) Metabolism-disrupting chemicals and the constitutive androstane receptor CAR. Cells 9:2306
56. Wang J, Hou T (2009) Recent advances on *in silico* ADME modeling. Annu Rep Comput Chem 5:101–127
57. Shao Y, Molnar LF, Jung Y, Kussmann J, Ochsenfeld C, Brown ST, Gilbert AT, Slipchenko LV, Levchenko SV, O'Neill DP, DiStasio RA Jr, Lochan RC, Wang T, Beran GJO, Besley NA, Herbert JM, Lin CY, van Voorhis T, Chien SH, Sodt A, Steele RP, Rassolov VA, Maslen PE, Korambath PP, Adamson RD, Austin B, Baker J, Byrd EFC, Dachsel H, Doerksen RJ, Dreuw A, Dunietz BD, Dutoi AD, Furlani TR, Gwaltney SR, Heyden A, Hirata S, Hsu CP, Kedziora G, Khalliulin RZ, Klunzinger P, Lee AM, Lee MS, Liang W, Lotan I, Nair N, Peters B, Proynov EI, Pieniazek PA, Rhee YM, Ritchie J, Rosta E, Sherrill CD, Simmonett AC, Subotnik JE, Woodcock HL III, Zhang W, Bell AT, Chakraborty AK, Chipman DM, Keil FJ, Warshel A, Hehre WJ, Schaefer HF III, Kong J, Krylov AI, Gill PMW, Head-Gordon M (2006) Advances in methods and algorithms in a modern quantum chemistry program package. Phys Chem Chem Phys 8:3172–3191
58. Yap CW (2011) PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474
59. Livingstone DJ, Salt DW (2005) Variable selection—spoilt for choice? Rev Comput Chem 21:287–348
60. Dearden JC, Cronin MTD, Kaiser KLE (2009) How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ Res 20:241–266
61. Ballabio D, Consonni V, Mauri A, Claeys-Bruno M, Sergent M, Todeschini R (2014) A novel variable reduction method adapted from space-filling designs. Chemom Intell Lab Syst 136:147–154
62. Ambure P, Aher RB, Gajewicz A, Puzyn T, Roy K (2015) "NanoBRIDGES" software: open access tools to perform QSAR and nano-QSAR modeling. Chemom Intell Lab Syst 147:1–13
63. Dastmalchi S, Hamzeh-Mivehroud M, Sokouti B (2018) Quantitative structure-activity relationship: a practical approach. CRC Press, Boca Raton
64. Gramatica P (2013) In: Reisfeld B, Mayeno AN (ed) Computational toxicology, Vol. 2. Springer, New York
65. Minitab Statistical Software (2017) Minitab 18.1. State College, PA: Minitab, Inc
66. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137–148
67. Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, Tropsha A (2012) Does rational selection of training and test sets improve the outcome of QSAR modeling? J Chem Inf Model 52:2570–2578
68. Snarey M, Terrett NK, Willett P, Wilton DJ (1997) Comparison of algorithms for dissimilarity-based compound selection. J Mol Graph Model 15:372–385
69. BIOVIA, Dassault Systemes (2001) Materials Studio 7.0, San Diego: Dassault Systemes
70. Liu P, Long W (2009) Current mathematical methods used in QSAR/QSPR studies. Int J Mol Sci 10:1978–1989
71. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77
72. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim 33(2):155–173
73. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. WILEY-VCH, Weinheim
74. Guha R (2008) On the interpretation and interpretability of quantitative structure-activity relationship models. J Comput Aided Mol Des 22:857–871
75. Ravichandran V, Jain PK, Mourya VK, Agrawal RK (2007) QSAR study on some arylsulfonamides as anti-HIV agents. Med Chem Res 16:342–351
76. Papa E, Dearden JC, Gramatica P (2007) Linear QSAR regression for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. Chemosphere 67:351–358
77. Wang C, Wei Z, Wang L, Sun P, Wang Z (2015) Assessment of bromine-based ionic liquid toxicity toward aquatic organisms and QSAR analysis. Ecotoxicol Environ Saf 115:112–118
78. Miyashita Y, Li Z, Sasaki S (1993) Chemical pattern recognition and multivariate analysis for QSAR studies. Trends Anal Chem 12(2):50–60
79. Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT (2004) Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorg Med Chem Lett 14:3283–3290
80. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden JC, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57(12):4977–5010
81. Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK (2011) Validation of QSAR models—strategies and importance. Int J Drug Des Discov 2(3):511–519
82. Muhammad U, Uzairu A, Arthur DE (2018) Review on: quantitative structure activity relationship (QSAR) modeling. J Anal Pharm Res 7(2):240–242
83. Topliss JG, Costello RJ (1972) Chance correlations in structure-activity studies using multiple regression analysis. J Med Chem 15(10):1066–1068
84. Kiralj R, Ferreira MMC (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J Braz Chem Soc 20(4):770–787
85. Sheridan RP (2012) Three useful dimensions for domain applicability in QSAR models using random forest. J Chem Inf Model 52:814–823
86. Roy PP, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. Molecules 14:1660–1701

87. Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predictive ability by external validation techniques. J Chemom 24:194–201
88. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26(5):694–701
89. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb Chem High Throughput Screen 14:450–474
90. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.