https://doi.org/10.48048/tis.2022.3976

Comparative Study of Depuration Rate Prediction against Mussel (Elliptio complanata) using Different Chemometric Approaches

Vandana Pandey

Department of Chemistry, Kurukshetra University, Kurukshetra, Haryana 136119, India

(Corresponding author's e-mail: vandana_p71@rediffmail.com)

Received: 11 December 2020, Revised: 2 June 2021, Accepted: 11 June 2021

Abstract

Different chemometric approaches were applied to a heterogeneous dataset of persistent organic pollutants(POPs), which included polybrominated diphenyl ethers(PBDEs), polychlorinated biphenyls(PCBs) and polycyclic aromatic hydrocarbons(PAHs) with associated depuration rate constant in Mussel (Elliptio complanata), to develop robust quantitative structure-activity relationship(QSAR) models. These models were further validated for statistical significance and predictive ability by internal and external validation. Out of various methods available, genetic algorithm and principal component analysis (PCA) approaches were used to identify relevant molecular descriptors from a large descriptor pool that exhibited a strong correlation with the depuration rate constant values of the diverse dataset. Then, multiple linear regression(MLR) and artificial neural network (ANN) methods were applied to the selected descriptors to create good predictive models. Statistical comparison of 3 hybrid approaches namely, GA-MLR, GA-ANN and PCA-ANN have shown that the genetic algorithm coupled with ANN model is superior to the other 2 models (R^2train= 0.961, R^2test= 0.947, mapetest= 7.939 and rmsetest=0.128). The applicability domain of the selected models was analyzed using the Euclidean distance and leverage approach signifies that all test set compounds fall within the applicability domain of the developed regression-based models.

Keywords: Persistent organic pollutants, Chemometric methods, Genetic algorithm, Principal component analysis, Neural network

Introduction

POPsare mostly synthetic toxic chemicals that adversely affect the environment around the world due to their transportation through wind and water. Because of their lipophilic nature, they easily accumulate, persist, and transfer from one species to the next through the food chain and bioconcentrate leading to toxicologically relevant concentration. PBDEs[1], PCBs [2] and PAHs [3] are some commonly occurring POPs. PBDEs have been and are still used in textiles, flexible polyurethane foams, as well as in electric appliances and electronic devices. These hydrophobic, organobromine compounds are known carcinogens [4]. PCBs are a mixture of chlorinated hydrocarbons and have found applications in dielectrics, paint and plastics. Constant exposure to PCBscreates several eye problems, pigmentation and skin problems [5]. PAHsare a group of more than a hundred chemicals that originate from anthropogenic activities such as industrial production and petroleum applications. Many PAHs have toxic, mutagenic and/or carcinogenic properties [6]. Due to extremely low concentration levels particularly in an aquatic environment, direct monitoring of these pollutants is difficult, expensive and time-consuming [7]. One way to monitor the accumulation of pollutants in the aquatic environment is through the use of sentinel species such as mussels [8]. Mussels are considered an ideal aquatic monitor for POPs due to their ability to concentrate bio-available residue from the water. Elliptio complanata is a freshwater mussel, widely utilized to monitor contaminants in the freshwater ecosystem and the overall health of the aquatic ecosystem[9]. However, to adequately acquire and interpret bioconcentration and bioaccumulation data, it is necessary to understand toxicokinetic parameters that influence POPs bioaccumulation. Depuration or elimination rate constant (k_d) is one such toxicokinetic parameter that gives information about the time required for the polluted mussels to reach a steady state in the environment and characterizes the depuration process [10]. Calculation of depuration rate constant from experimental results is expensive and time-consuming. Therefore, to decrease the experimental cost and to fill the data gap of organic pollutants, a QSAR [11] can be used as an alternative approach. The properties of pollutants in the environment are directly linked to their respective molecular structure and derives simultaneously from various chemical properties/activities[12]. QSAR is essentially a predictive approach discovered by Hansch and Leo[13] and Free and Wilson[14] used in various fields including ecotoxicological risk assessment. It is based on developing a relationship between structural features represented by descriptors and the observed or experimental property of a molecule. There are various statistical and machine learning methods to construct a good predictive model for a given data set. MLR[14], partial least squares (PLS)[15] and PCA (principal component analysis) are some of the most common linear regression methods. MLR yields models that are simpler and easier to interpret. The general purpose of multiple regressions is to quantify the relationship between more than one independent variable and a dependent variable and construct a linear relationship between them in the form of an equation. On the other hand, ANN [16] and support vector machines (SWM)[17] represent the non-linear regression methods. ANNs are universally recognized tools where numerical variables play an important role in the solution of the problem. Thus, a large number of studies have shown that the 3-layer feed-forward neural network trained by the back-propagation algorithm [18] is a very powerful tool for deriving complex QSAR models.

An important factor to the success of QSAR modeling is the selection of a few most relevant descriptors [19] from a wide range of descriptors available. There are several methods available for optimization and selection of descriptors required for the model building process. It can be achieved by attempting to select important descriptors either through model selection [20,21](SWMLR, GA-MLR, etc.) or finding a lower-dimensional transformation that preserves most of the information present in the original set of descriptors (PCA)[22].

Genetic algorithm is a very powerful searching and optimizing method to select the variables significantly contributing to the prediction and discard the other variables by using an appropriate fitness function [23]. It uses binary bit string representation (1or 0) to encode descriptors. These strings are evaluated based on a fitness function. It is based on the principles of natural selection and evolution and uses natural rules and terminology like chromosome, population, crossover and mutation, etc. Here population can be defined as a collection of descriptors and a chromosome is simply a subset of descriptors chosen from the descriptor pool. Crossover involves swapping a portion of the chromosomes of a pair of individuals. The mutation is performed by randomly changing a part of the chromosome. PCA is also a robust statistical technique useful for reducing the redundancy present in the data set and eliminating uncorrelated variables [24]. It determines dimensions with the maximum variation that are orthogonal to each other.

The main objective of the present work was to develop a statistically robust and easily interpretable QSAR model with high external predictability by comparing results obtained from different chemometric approaches and those from previous work and experimental values. Therefore, in the present work, 3 hybrid approaches namely, GA-MLR, GA-ANN and PCA-ANN were used to generate robust and reliable QSAR models with good validation characteristics.

Materials and methods

Dataset and Softwares

The dataset used in this study has depuration rate constant values (in terms of logk_d) of a heterogeneous set of 64 POPs, selected from work published by Li *et al.*[25]. The entire dataset as well as their corresponding logk_d values are given in **Table1**. Online facilities, Pubchem, eDragon [26] and genetic algorithm 1.4tool [27] were used for SDF file preparation, descriptor generation and best subset selection respectively. PCA and ANN calculations were performed with Matlab (R2013a) software.

Generation and selection of molecular descriptors

SDF files of all these compounds were generated using Pubchem facilities and uploaded to Dragon software for descriptor generation. A large pool of descriptors was generated for each molecule including molecular properties, constitutional descriptors, topological descriptors, connectivity indices, information indices, topological charge indices, geometrical descriptors, WHIM, 3D-Morse, Getaway and RDF descriptors. For the development of a robust QSAR model, a key step is the selection of the optimal subset of variables. Genetic algorithm and PCA methods were applied for subjective feature selection after preprocessing the data set.

Genetic algorithm

For selecting an appropriate number of descriptors and to get the statistically best QSAR model, genetic algorithm was implemented after objective feature selection. In the present case, a string composed of 366 genes representing the presence and absence of a descriptor in a string was coded by 1or 0 respectively. The number of genes with a value of 1was kept relatively low to have a small set of descriptors. The operators used for this procedure were crossover and mutation. In each generation, the fitness of every individual in the population was evaluated. The chromosomes with less number of selected descriptors, a high value of fitness function (F) and a low value of lack of fit (LOF) were marked as informative chromosomes.

Principal component analysis (PCA)

To discard irrelevant and unstable information present in the large descriptor set, PCA was also applied to the entire data set after deleting constant and near-constant variables. Before implementing it, the input and target dataset were scaled. The PCA not only orthogonalizes the component of the input vectors but also orders the resulting orthogonal components so that principal components with the largest variation come 1st and eliminates those components which contribute the least to the variation in the dataset. In the present case, the first few principal components (PCs)were selected which accounted for more than 95% of overall variability. These PCs were used as input for the subsequent non-linear mapping by ANN.

Splitting and mapping of the dataset

Rational division of the data set into training and test set is a critical step in the QSAR methodology. Kennard Stone algorithm [28] was used for the rational splitting of the data set into training and test set. Linear and non-linear mapping of the refined data set was carried out by MLR and ANN methods, respectively. The advantage of MLR is its simple form and easily interpretable mathematical expression. For improvement of the performance of the model and to explore the non-linear relationship between selected independent variables and the output property, ANN was implemented. Two major components contribute to the effectiveness of a neural network in solving a particular problem: Its architecture and the algorithm by which it is trained. In this work, a fully connected, 3-layered, feed-forward ANN trained with the Levenberg-Marquardt algorithm was used.

Validation

One of the crucial aspects of QSAR modeling for the development of a reliable QSAR model is validation [29]. Validation parameters provide a thorough assessment of predictive QSAR models, thus ensuring maximum reliability, quality and efficiency of the regression models for practical purposes. The parameters, squared correlation coefficient (R^2), cross-validation(Q^2), adjusted R^2 (R^2adj), root-mean-squared error (RMSE) and scrambling (Y-Randomization) technique were used for internal validation of the models. For estimating the external predictability of proposed models, R^2(test), mape(test), rmse(test) as well as parameters recommended by Golbraikh and Tropsha and Roy *et al.* for the proposed QSAR models were calculated [30,31]. Defining the Applicability Domain (AD) is an important aspect according to Organization for Economic Co-operation and Development (OECD)[32,33] principles for model validation. In the present study, AD was verified by calculating the normalized mean Euclidean distance value for each compound as well as using the Leverage approach.

Table 1 Experimental calculated log$k_d$values and normalized mean distance values of POPs.

S.N.	Compounds	log k_d (Obs))	GA-MLR	GA-ANN	PCA-ANN	Normalized Mean distance
1	BDE-75	–1.585	–1.5462	–1.4582	–1.7464	0.195837
2	BDE-47	–1.721	–1.8271	–1.8529	–1.7854	0.572955
3	BDE-100	–2.097	–1.9627	–1.9023	–1.8644	1
4	BDE-99	-2	–1.9866	–1.9876	–1.9243	0.517048
5	BDE-153	–2.222	–2.1273	–2.0384	–2.025	0.462603
6	BDE-138	–2.046	–2.1376	–1.9485	–1.9722	0.357974
7	BDE-190	–1.886	–1.7531	–1.7175	–2.062	0.324775

S.N.	**Compounds**	**log k_d (Obs))**	**GA-MLR**	**GA-ANN**	**PCA-ANN**	**Normalized Mean distance**
8	PCB-19	–1.081	–1.2603	–1.1416	–1.0309	0.12595
9	PCB-22	–1.149	–1.3198	–1.1836	–1.1565	0.096177
10	PCB-42	–1.347	–1.4651	–1.3674	–1.3828	0.009119
11	PCB-74	–1.553	–1.5686	–1.5158	–1.4732	0.005888
12	PCB-66	–1.509	–1.5425	–1.4732	–1.4274	0.00316
13	PCB-95	–1.538	–1.7076	–1.6773	–1.6293	0.003045
14	PCB-91	–1.553	–1.6312	–1.5594	–1.6648	0.037164
15	PCB-92	–1.678	–1.8254	–1.8525	–1.7011	0.028575
16	PCB-99	–1.658	–1.7497	–1.7338	–1.7143	0.007493
17	PCB-97	–1.638	–1.7619	–1.7543	–1.6756	0.003626
18	PCB-87	–1.602	–1.7431	–1.7345	–1.6753	0
19	PCB-85	–1.721	–1.6677	–1.618	–1.6944	0.010578
20	PCB-110	–1.602	–1.7359	–1.7329	–1.7165	0.007052
21	PCB-118	–1.854	–1.8263	–1.8394	–1.759	0.022907
22	PCB-105	–1.824	–1.7504	–1.7527	–1.7368	0.006165
23	PCB-136	–1.745	–1.7772	–1.7542	–1.8063	0.204627
24	PCB-151	–2	–1.9031	–1.9844	–1.9037	0.175867
25	PCB-149	–1.921	–1.8990	–1.9318	–1.895	0.172723
26	PCB-134	–2.097	–1.8599	–1.9326	–1.8956	0.180186
27	PCB-146	–2	–2.0204	–2.0945	–1.9577	0.207131
28	PCB-141	–2.155	–2.0337	–2.1059	–1.9476	0.21503
29	PCB-130	–2.155	–1.9485	–2.0626	–1.948	0.198582
30	PCB-137	–2.222	–1.9582	–2.0242	–1.9646	0.179165
31	PCB-128	–1.721	–1.8492	–1.8986	–1.9317	0.173113
32	PCB-156	–2.046	–2.0356	–2.1168	–1.9932	0.220116
33	Naphthalene	–0.654	–0.5538	–0.62	–0.6413	0.950287
34	2-methylnaphthalene	–0.686	–0.5904	–0.6543	–0.5969	0.814141
35	1-methylnaphthalnene	–0.604	–0.5965	–0.6524	–0.709	0.809827
36	1,1'-biphenyl	–0.672	–0.6267	–0.7018	–0.718	0.669799
37	2,6-dimethylnaphthalene	–0.577	–0.6293	–0.6861	–0.5514	0.692558
38	Acenaphthylene	–0.734	–0.6783	–0.6965	–0.7672	0.686605
39	dibenzofuran	–0.635	–0.8085	–0.7566	–0.6331	0.487331
40	Acenaphthene	–0.625	–0.7027	–0.6373	–0.6158	0.755504
41	2,3,5-trimethylnaphthalene	–0.746	–0.6713	–0.7192	–0.7071	0.585927
42	9H-fluorene	–0.721	–0.6681	–0.7156	–0.6326	0.594371
43	dibenzothiophene	–0.793	–0.7249	–0.7257	–0.6747	0.570586
44	Phenanthrene	–0.768	–0.7771	–0.7718	–0.7952	0.466385
45	1-methylphenanthrene	–0.858	–0.7812	–0.8404	–0.8474	0.394472
46	Benzo[b]fluoranthene	–1.082	–1.0680	-1.1857	–1.1187	0.199553
47	Benzo[e]pyrene	–1.138	–1.0218	–1.1543	–1.1909	0.216395
48	Indeno[123cd]pyrene	–1.327	–1.1621	–1.2267	–1.2026	0.216741
49	Naphtho[1,2-b]phenanthrene	–1.163	–1.2747	–1.2891	–1.1496	0.173614
50	Benzo[ghi]perylene	–1.223	–1.3048	–1.2825	–1.2041	0.018608

S.N.	Compounds	log k_d (Obs))	GA-MLR	GA-ANN	PCA-ANN	Normalized Mean distance
51	Corronene	–1.3	–1.4109	–1.3445	–1.3155	0.105814
	Test set					
1	BDE-28*	–1.432	–1.5919	–1.5355	–1.5918	0.261885
2	PCB-138*	–2.155	–1.9259	–1.977	–1.9424	0.172519
3	PCB-179*	–2.301	–2.0331	–2.2183	–2.0405	0.47711
4	PCB-178*	–2.222	–2.1817	–2.4787	–2.0938	0.558747
5	1-methyl-9H-fluorene*	–0.903	–0.7136	–0.7556	–0.7008	0.505664
6	Anthracene*	–0.747	–0.7833	–0.7718	–0.6309	0.460829
7	Fluoranthene*	–0.901	–0.8529	–0.8872	–0.8886	0.352265
8	Pyrene*	–0.786	–0.8449	–0.8869	–0.9101	0.363409
9	Benzo[a]anthrcene*	–1.034	–1.0218	–1.1304	–1.0146	0.203054
10	Chrysene*	–1.078	–1.0263	–1.1366	–0.9846	0.204619
11	Benzo[k]fluoranthene*	–1.23	–1.1156	–1.2056	–1.0621	0.152059
12	Benzo[a]pyrene*	–1.122	–1.0825	–1.1891	–1.1087	0.174782
13	Perylene*	–1.376	–1.0431	–1.1579	–1.0852	0.175923

Results and discussion

A total of 1666 descriptors available in e-dragon software were generated online using the eDragon descriptor calculation facility. The objective feature selection procedure was performed in 3 steps. At first, descriptors with constant and near-constant values for all molecules were removed. Then in the 2nd step, descriptors with a correlation coefficient less than 0.7 with the dependent variable (logk_d) were considered redundant and removed. Finally, pairs of descriptors with a correlation coefficient greater than 0.9 were considered intercorrelated and the one having a greater correlation with logk_d value in each correlated pair was retained. Applying these 3 steps, the number of descriptors was drastically reduced from 1666 to 366. Subsequently, the entire dataset was used for subjective feature selection. Genetic algorithm and PCA approaches were applied to select the most significant descriptors. The composition of training and test sets is crucial to obtain an internally consistent model and to test its external ability of prediction using an equally representative set. Kennard-Stone is a rational method that was employed to split the entire dataset into training and test sets. As a result, 80% of the data set (51 compounds) was used as the training set and the remaining 20% (13 compounds) was used as the test set.

GA-MLR

The genetic algorithm was used to search feature space and select descriptors relevant to log k_d values and the linear models were built using MLR method with the selected descriptors from GA, called GA-MLR. The important parameters that impact the GA performance are listed as follows: Crossover probability=1, mutation probability=0.5, the initial number of equations generated=100, and the total number of iterations=100. Various validation parameters were used to check the robustness of the models. Finally, 4 indices showing high accordance with logk_d activity were selected out namely, RDF070v (RDF descriptors: Radial Distribution Function - 70weighted by van der Waals volume), Mor03m(3D-MoRSE descriptors:Signal 3/weighted by mass), Ss(Constitutional descriptor: Sum of Kier-Hall electrotopological state) and Me (Constitutional indices: Mean atomic Sanderson electronegativity (scaled on Carbon atom)). The correlation among selected descriptors was analyzed by computing the correlation matrix. No significant correlation is observed between the selected descriptors (supplementary **Table S1**). The best multivariate linear model representing a linear relationship between selected 4 descriptors and log $k_d$values of POPs in mussel Elliptio complanata is described by the following equation:

$$\log k_d = 4.02065(+/-0.8776) -0.05948(+/-0.01716)\ \mathrm{RDF070v} +0.02534(+/-0.0041)\ \mathrm{Mor03m} - 0.02651(+/-0.0035)\ \mathrm{Ss} -4.23799(+/-0.95537)\ \mathrm{Me} \quad (1)$$

N=51, R^2:0.954, R^{2}Adj:0.950, Q^2:0.9423, RMSEP=0.111, MAPE=7.30.

Y randomization technique was also performed by randomly shuffling the dependent variable while keeping the independent variable unchanged. Fifty random models were generated, which resulted in low values of average R^2 (0.093) and average Q^2(–0.131), confirming that the developed QSAR model is reliable. From all the statistical parameters, it can be seen that the proposed model is stable, robust and predictive, and was consequently used for the prediction of activities of the test set data. The prediction results obtained from the GA-MLR approach for the entire dataset are given in **Table1**.

Descriptors selected through the genetic algorithm, encode different aspects of the molecular structure. The descriptor RDF070v belongs to the 3D RDF descriptor which has a negative contribution to the log k_d value. These descriptors are based on the distance distribution in the molecule, related to the van der Waals volumes. The selection of Mor03 descriptor weighted by mass signifies the relevance of halide moiety to the depuration rate constant values. For (PBDEs) having Br moiety, the value of Mor03m descriptor is more negative as compared to its value for PAHs. The regression coefficient for this descriptor is positive and indicates that with increasing of Mor03m, logk_d value increases. Sum of Kier -Hall electrotopological state (Ss) is the sum of each atom and bond's electronic accessibility-a combination of steric and electronic effects and is related to the likelihood of electrons interacting with other molecules. This constitutional descriptor reflects the steric and electronic effects of surrounding atoms. A negative coefficient for Ss indicates that its value adversely affects the log$k_d$values of the chemicals. The descriptor Me has the highest contribution towards the depuration rate constant value as suggested by the GA-MLR model. The negative contribution of this descriptor indicates that the logk_d value of a candidate compound increases with decreasing value of Me descriptor. It is worthwhile to mention that the presences of halide atoms as well as their electronegativity in the POPs are the two important decisive factors for deciding the value of log k_d. As the number of halide atoms increases the value of logk_d decreases. It can be explained as follows: Increased number of halide atoms in POPs causes more polarization and steric hindrance, making them more lipophilic which results in restricted partitioning from lipid tissues of mussels to the aqueous phase. As the Chlorine atom is more electronegative than the bromine atom, the effect is more prominent in PCBs than in PBDEs. Moreover, planer PAHs are more electrophilic in nature, consequently, they are more hydrophilic and less lipophilic in nature, which results in more electrophilic interaction and a comparatively higher value of depuration rate constant.

GA-ANN

To explore non-linear relationships existing between logk_d values and the selected descriptors, an ANNmodeling method combined with GA was implemented. Various factors were considered including the number of nodes in the hidden layer, types of training algorithm, choice of activation function, number of epochs and learning rate. The number of nodes in the input layer was 4, as the input vectors were set of 4 descriptors selected by the GA. To optimize the number of nodes in the hidden layer, the concept of ρ as proposed by Andrea and Kalayeh [34] was used. The output consisted of the logk_d value. Levenberg-Marquardt algorithm [35] was used to train the network, as it is the fastest modern 2nd-order Hessian-based algorithm for non-linear least square optimization. The transfer function in the1st layer was tan-sigmoid and the output layer transfer function was linear. A fully connected 3-layered feed-forward with a backpropagation pattern with mean squared error (MSE) as the performance function was used. For the evaluation of the prediction power of the trained ANN, it was used to predict the log k_d values of the POPs selected in the test set. Several training sessions were conducted with the different numbers of hidden nodes ranging from 5 to 10. Finally, a network of 4-5-1 was selected and calculated log k_d values for the training and test set are present in **Table 1**. Statistical parameters for the training set are: R^2=0.961, RMSE=0.096 and MAPE=5.593.

PCA-ANN

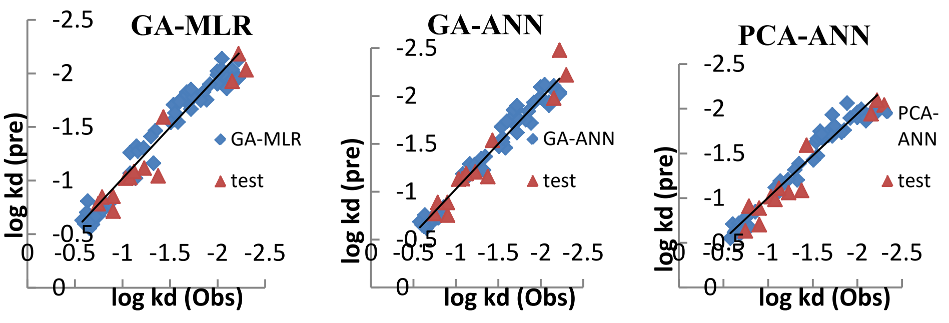
PCA is another data compression method used to group correlated variables replacing the orthogonal descriptors with a new set called principal components (PCs). The entire data set comprising 366 descriptors and corresponding output values were subjected to PCA analysis. Before implementing it, the set of molecular descriptors as well as target property was first normalized to have 0 mean and unity standard deviation and then was subjected to the PCA analysis for getting informative PCs before being introduced into the neural network. The first 7 PCs accounting for more than 98.9% of data variance in the original data matrix were selected and used as ANNinput. Therefore, in the implementation of the 3-layered ANN method, 7-x-1 architecture was applied, where x denotes the number of nodes in the hidden layer. The number of nodes in the hidden layer was optimized ranging from 3 to 7, following the concept of ρ [34]. As in GA-ANN, here also the transfer function in the 1st layer was tan-sigmoid and the output layer transfer function was linear. To optimize the number of nodes in the hidden layer, MSE value for the test set was calculated by varying the number of nodes in the hidden layer. From all ANNs, 7-5-1 architecture was selected (R^2train=0.961, RMSE(train)=0.105 and MAPE(train)=5.644), and calculated logk_d values for the entire set are presented in the **Table1**.

External validation

The statistical parameter used to analyze the prediction performance of the selected QSAR models areR^2test(coefficient of determination of external validation), RMSE(test) (root mean square error of external validation) and MAPE(test) (mean absolute percent error of prediction). The performance was also verified by computing parameters suggested by Golbraikh and Tropsha and r^2m proposed by Roy and Roy. All parameters are presented in **Table 2**. Here k and k' denote Golbraikh andTropsha slopes of the linear regression lines between the observed and the predicted activities in the external validation, R^2_0 and R'^2_0 represent Golbraikh-Tropsha absolute values the coefficients of multiple determination, r^2 and r'^2_0 are respectively the determination coefficients of the regression function, calculated using the experimental and the predicted data of the prediction set, forcing respectively the origin of the axis (r^2_0) or not (r^2). The parameter r^2m is calculated using the experimental values on the ordinate axis. The calculated values of the above-mentioned parameters are in good agreement with the proposed criteria.

It can be noted that although descriptors appearing in the GA-MLR modeling were used as input for generating a 3-layered GA-ANN model, the statistics have shown a large improvement, it is due to the capability of ANN to explore the complicated non-linear relationship between independent variables and log $k_d$values.

Figure 1 shows the correlation between experimental log$k_d$values of selected POPs against values calculated using GA-MLR (a) and GA-ANN (b) and PCA-ANN (c) methods, respectively. From the plots, it is clear that all models can predict the depuration rate constant values of selected POPs for training and test set quite well, as all points are close to the regression line, thus ensuring the presented models' predictive ability. It can be seen in **Figure 1(b)** that the data are more concentrated in the vicinity of the trend line for the GA-ANN method. Therefore, graphical representation also confirms the superiority of the GA-ANN model over the other two models. For evaluation of systematic error in the proposed models, the residual values were plotted against experimental log k_d values (**Figure S1 supplementary data**). The spread of residual around the residual=0 line indicates there is no systematic error in the proposed model. Recently, Yu [36] developed QSAR models using MLR and support vector machine method for the depuration rate prediction for a homogenous set having 63 PCB congeners using 4 molecular descriptors. In the present case, 3 hybrid regression methods, GA-MLR, GA-ANN and PCA-ANN were used to develop QSAR models for the prediction of depuration rate constant values of a heterogeneous dataset containing POPs. Li *et al.*[25] have performed a QSAR study using the PLS method on the same heterogeneous set of POPs with their depuration rate values (R^2train=0.95, R^2test=0.89and SE(test)= 0.16). Comparison of 3 approaches used in the present work and that of Li *et al.*[25] shows that the GA-ANN method outperforms all other approaches. The superiority of GA-ANN can be attributed to the non-linear mapping ability of ANN.



(a) (b) (c)

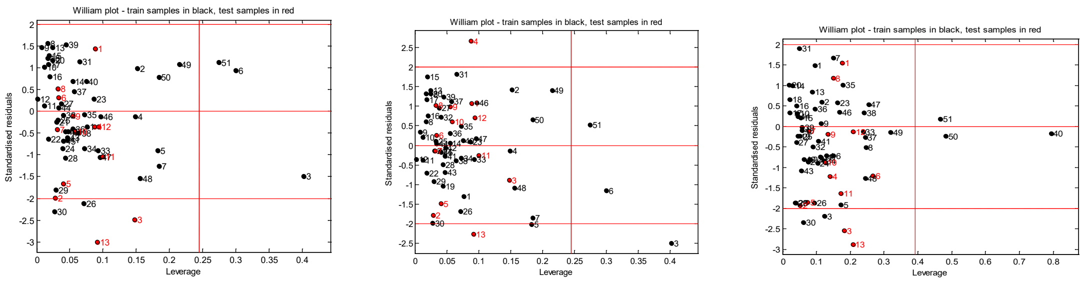
Figure1 Scatter plots of observed vs predicted log$k_d$values of POPs for models developed from GA-MLR, GA-ANN and PCA-ANN methods.

Table 2 Statistical significance and validation parameters.

Statistical parameters	Conditions	GA-MLR	GA-ANN	PCA-ANN
R^2(train)		0.954	0.961	0.961
rmse(train)		0.111	0.096	0.105
mape(train)		7.3	5.593	5.644
R^2(test)	> 0.6	0.937024	0.947977	0.938961
mse(test)		0.024824	0.016403	0.026691
rmse(test)		0.157557	0.128073	0.163373
mape(test)		8.993842	7.939984	10.74804
r_0^2		0.938	0.943	0.939
r'^2_0		0.938	0.943	0.939
K	0.85<k<1.15	1.064394	0.991871	1.073897
k'	0.85<k<1.15	0.931524	1.000189	0.923426
R_0^2		0.972705	0.999419	0.960219
R'^2_0		0.962536	1	0.956609
$r_0^2 - r'^2_0$	<0.3	0	0	0
r2m=	>0.5	0.908338	0.900828	0.939

Applicability domain

An acceptable QSAR model should possess a defined AD according to the OCED principle. AD represents the chemical space defined by the structural information extracted from the chemicals used in training set compounds in the QSAR modeling. All presented models were verified for the applicability of the domain to generate reliably predicted values of log $k_d$for the chemicals. The applicability domain of the model was analyzed by the Euclidean distance method as well as by using Williams plots (**Figure 2**). The outcomes from applicability domain analysis for the GA-MLR model, by the Euclidean distance method, are quite satisfactory within the normal distribution range and normalized mean distance values are reported in **Table 1**. The graphical representation of the results is also presented in the supplementary **Figure S2**. The Williams plot was used to visualize influential chemicals, i.e. chemicals with leverage greater than h*, as well as outlier chemicals, i.e. chemicals with standardized residual greater than ±3δ, the cut-off values. All compounds of training and test set to fall within 3 standard deviation units, indicating there is no outlier present in the models. The model is characterized by the presence of some influential training chemicals, with leverage values greater than the 'warning leverage' (h*) fixed at 3(P+1)/n. (where P represents the number of model descriptors and n denotes the number of chemicals in the training set). The Williams plots for GA-MLR and GA-ANN models show the presence of 3 training samples (3, 6 and 51) having a great influence on the models (i.e., greater than the warning leverage h*), as both models have the same descriptors for the linear and nonlinear mapping. For the PCA-ANN model, influential chemicals are 40, 50 and 51. However, the test set is within the AD for all 3 models. Therefore, predictions are reliable.



GA-MLR GA-ANN PCA-ANN

Figure 2 Williams plot of the regression-based QSAR model using the GA-MLR, GA-ANN and PCA-ANN techniques.

The derived QSAR models indicate the key factor to the success of the QSAR study is the selection of a small set of most relevant molecular descriptors used for linear and non-linear mapping. In the present case, the descriptors encoding van der Waals volume, mass, electronegativity and electronic accessibility are important contributors, controlling the depuration rate constant values of POPs.

Conclusions

In the present study 3 hybrid regression methods, GA-MLR, GA-ANNand PCA-ANN were investigated to establish quantitative-structure-activity relationships for the prediction of depuration rate constant values of a dataset containing (POPs) in mussel Elliptio complanata. The proposed models were assessed and validated using the OECD principles. Results indicate that a properly selected and trained neural network can fairly represent the dependence of the logk_d values on the selected descriptors. The optimized neural network can then simulate the complicated nonlinear relationship between the logk_d value and the descriptors. All presented models are robust enough to make an accurate and reliable prediction. Based on the results, the GA-ANN model was admittedly the best model considering the goodness of fit and predictivity parameters for the evaluation of the proposed model.

References

[1] M Athanasiadou, SN Cuadra, G Marsh, A Bergman and K Jakobsson. Polybrominated diphenyl ethers (PBDEs) andbioaccumulative hydroxylated PBDE metabolites in young humans from Managua, Nicaragua. *Environ. Health Perspect*. 2008; **116**, 400-8.

[2] SH Safe.Polychlorinated biphenyls(PCBs): Environmental impact, biochemical and toxic responses, and implications for risk assessment.*Crit. Rev. Toxicol.*1994; **24**,87-149.

[3] ALC Lima, TI Eglinton and CM Reddy. High-resolution record of pyrogenic polycyclic aromatic hydrocarbon deposition during the 20th century. *Environ. Sci. Technol.* 2003; **37**,53-61.

[4] National Toxicology Program. *14th Report on carcinogens*.National Toxicology Program, ResearchTriangle Park, North Carolina, 2014.

[5] H Iwata, M Watanabe, Y Okajima, S Tanabe, M Amano, N Miyazaki and EA Petrov. Toxicokinetics of PCDD, PCDF and coplanar PCB congeners in Baikal seals, *pusasibirica*: Age-related accumulation, maternal transfer, and hepatic sequestration. *Environ. Sci. Technol.*2004;**38**, 3505-13.

[6] MS Zedeck. Polycyclic aromatic hydrocarbons: A review. *J. Environ. Pathol. Toxicol.*1980;**3**,537-67.

[7] AVD Linde, AJ Hendriks and DTHM Sijm. Estimating biotransformation rate constants of organic chemicals from modeled and measured elimination rates.*Chemosphere* 2001; **44**, 423-35.

[8] S Uno, H Shiraishi, S Hatakeyama and A Otsuki.Uptake and depuration kinetics and BCFsof several pesticides in three species of shellfish (*Corbicula leana*, *Corbicula japonica* and*Cipangopludinachinensis*): Comparison between field and laboratory experiment. *Aquat.Toxicol*. 1997; **39**, 23-43.

[9] S O'Rourke, KG Drouillard and GDHaffner. Determination of laboratory and field elimination rates of polychlorinated biphenyls (PCBs) in the freshwater mussel, Elliptiocomplanata. *Arch. Environ. Contam. Toxicol*. 2004; **47**,74-83.

[10] KG Drouillard, S Chan, S O'Rourke, GD Haffner and RJ Letcher. Elimination of 10 polybrominated diphenyl ether(PBDE) congeners and selected polychlorinated biphenyls(PCBs) from the freshwater mussel, *Elliptio complanata*.*Chemosphere* 2007; **69**, 362-70.

[11] D Wu, XH Liu, L Wang, M Xu, T Sun, Z Yang and J Zhou. QSARs on the depuration rate constants of polycyclic aromatic hydrocarbons in *Elliptio complanata*. *QSARComb. Sci*. 2009; **28**,537-41.

[12] P Gramatica, E Papa and ASangion. QSAR modeling of cumulative environmental end-points for the prioritization of hazardous chemicals.*Environ.Sci.:Processes Impacts* 2018; **20**, 38-47.

[13] C Hansch and A Leo. *Exploring QSAR: Fundamentals and applications in chemistry and biology*. American Chemical Society, Washington, DC, 1995.

[14] SMJ Free and JW Wilson. A mathematical contribution to structure activity studies. *J. Med. Chem*. 1964; **7**, 395-9.

[15] S Schmidt, M Schindler, D Faber and J Hager. Fish early life stage toxicity prediction from acute daphnid toxicity and quantum chemistry. *SAR QSAR Environ. Res*. 2021; **32**, 151-74.

[16] L PA Chiari, APD Silva, AAD Oliveira, CF Lipinski, KM Honorio and ABFD Silva.Drug design of new sigma-1 antagonists against neuropathic pain: A QSAR study using partial least squares and artificial neural networks. *J. Mol. Struct*. 2021; **1223**, 129156.

[17] S Zheng, C Liand GWei. QSAR modeling for reaction rate constants of e_{aq}^{-} with diverse organic compounds in water. *Environ. Sci.: Water Res. Technol*. 2020; **6**, 1931-8.

[18] DE Rumelhart, GE Hinton and RJ Williams. Learning representations by back-propagating errors. *Nature* 1986; **323**, 533-6.

[19] R Todeschini and V Consonni. *Molecular descriptors for chemoinformatics*.Wiley-VCH, Weinheim, Germany, 2009.

[20] J Zupan and M Novic.General type of a uniform and reversible representation of chemical structures. *Anal. Chim. Acta* 1997; **348**, 409-18.

[21] BT Hoffman, T Kopajtic, JL Katz and AH Newman. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of Molconn Z descriptors. *J. Med. Chem*. 2000; **43**, 4151-9.

[22] M Shahlaei, A Fassihi and L Saghaie. Application of PC-ANN and PC-LS-SVM in QSAR of CCR1antagonist compounds: A comparative study.*Eur. J. Med. Chem*. 2010; **45**, 1572-82.

[23] MMitchell. *An introduction to genetic algorithms*. MIT Press, Cambridge, MA, 1996, p. 205-18.

[24] R Vendrame, RS Braga, Y Takahata and DS Galvao.Structure-activity relationship studies of carcinogenic activity of polycyclic aromatic hydrocarbons using calculated molecular descriptors with principal component analysis and neural network methods. *J. Chem. Inf. Comput. Sci*. 1999; **39**, 1094-104.

[25] F Li, X Liu, L Zhang, L You, H Wu, X Li, J Zhao and L Yu. QSAR studies on the depuration rates of polycyclic aromatic hydrocarbons, polybrominated diphenyl ethers and polychlorinated biphenyls in mussels (Elliptio complanata). *SAR QSAR Environ. Res*. 2011; **22**, 561-73.

[26] IV Tetko, J Gasteiger, R Todeschini, A Mauri, D Livingstone, P Ertl, VA Palyulin, EV Radchenko, NS Zefirov, AS Makarenko, VY Tanchuk and VV Prokopenko. Virtual computational chemistry laboratory - design and description. *J. Comput. Aided Mol. Des.* 2005; **19**, 453-63.

[27] K Roy and I Mitra. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screen*. 2011; **14**, 450-74.

[28] RW Kennard and LA Stone. Computer-aided design of experiments. *Technometrics* 1969; **11**, 137-48.

[29] W Tong, Q Xie, H Hong, L Shi, H Fang and RPerkins.Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect*. 2004; **112**, 1249-54.

[30] A Golbraikh and A Tropsha. Beware of q2! *J. Mol. Graph. Model*. 2002; **20**, 269-76.

[31] PP Roy, S Paul, I Mitra and K Roy. On two novel parameters for validation of predictive QSARmodels. *Molecules* 2009; **14**, 1660-701.

[32] F Sahigara, K Mansouri, D Ballabio, A Mauri, V Consonni and R Todeschini. Comparison of different approaches to define the applicability domain of QSAR models.*Molecules* 2012; **17**, 4791-810.

[33] Y Canizares-Carmenate, LEC Delgado, F Torrens and JA Castillo-Garit. Thorough evaluation of OECD principles in modelling of 1-[(2- hydroxyethoxy)methyl]-6-(phenylthio)thymine derivatives using QSARINS.*SAR QSAR Environ. Res.* 2020; **31**, 741-59.

[34] TA Andrea and H Kalayeh. Applications of neural networks inquantitative structure-activity relationships of dihydrofolatereductaseinhibitors. *J. Med. Chem*. 1991; **34**, 2824-36.

[35] M Jalali-Heravi, M Asadollahi-Baboli and P Shahbazikhah. QSAR study of heparanase inhibitors activity using artificial neuralnetworks and LevenbergeMarquardt algorithm. *Eur. J. Med. Chem*. 2008; **43**, 548-56.

[36] X Yu. Prediction of depuration rate constants for polychorinated biphenyl congeners. *ACS Omega* 2019; **4**, 15615-20.

Supplementary data

Table S1 Correlation matrix for the inter-correlation of selected descriptors.

	RDF070v	***Mor03m***	***Ss***	***Me***
RDF070v	1			
Mor03m	-0.59354	1		
Ss	0.610811	-0.48691	1	
Me	0.704123	-0.63778	0.831365	1

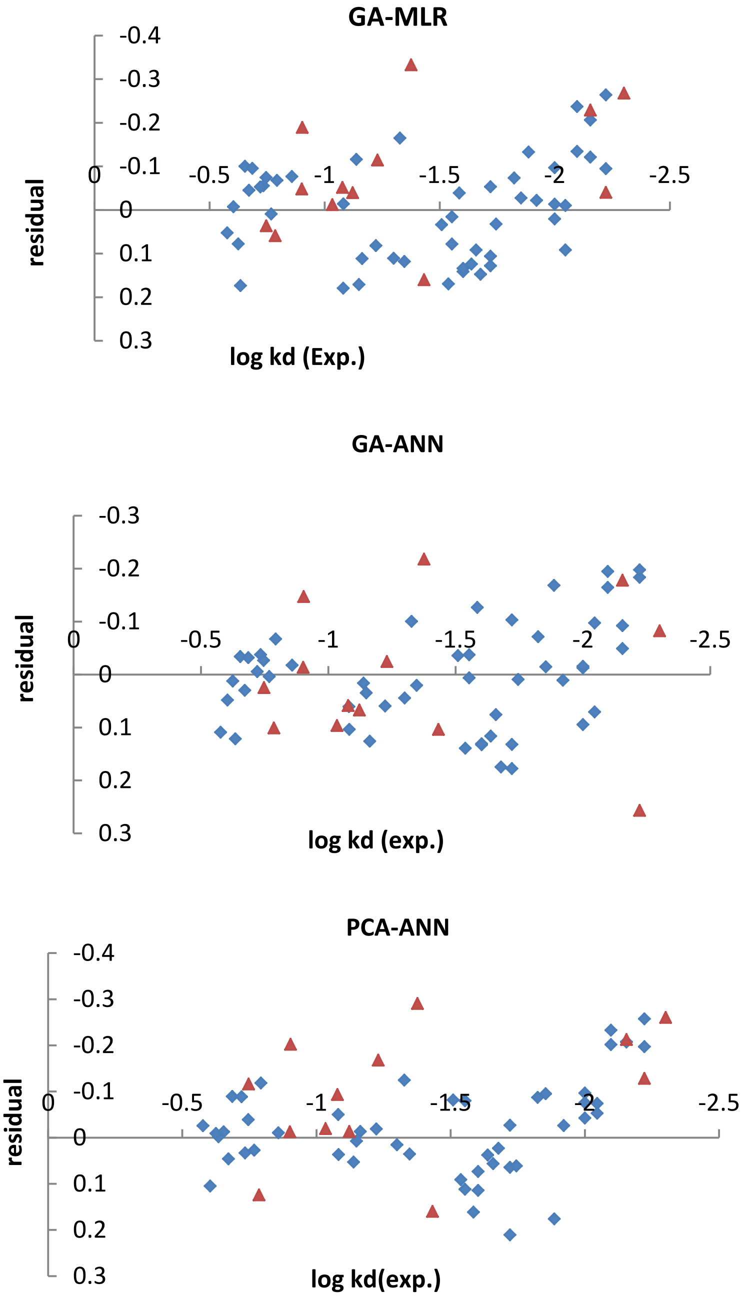


Figure S1 Residual plots of GA-MLR, GA-ANN and PCA-ANN models.

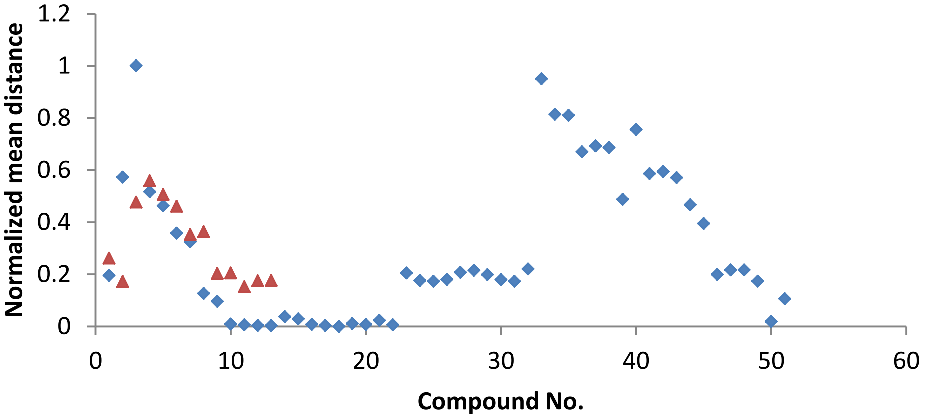


Figure S2 Graphical representation of applicability domain by Euclidean distance method for GA-MLR model.