1 **Linear and nonlinear models for predicting fish bioconcentration factors for**
2 **pesticides**

3

4 Jintao Yuan ^{a,b}, Chun Xie ^c, Ting Zhang ^b, Jinfang Sun ^b, Xuejie Yuan ^c, Shuling Yu ^d,
5 Yingbiao Zhang ^e, Yunyuan Cao ^a, Xingchen Yu ^a, Xuan Yang ^a, Wu Yao ^{a,*}

6

7 ^a School of Public Health, Zhengzhou University, Zhengzhou, 450001, China

8 ^b Key Laboratory of Environmental Medicine Engineering, Ministry of Education,
9 School of Public Health, Southeast University, Nanjing, 210009, China

10 ^c Shangqiu Medical College, Shangqiu, Henan Province, 476100, China

11 ^d Key Laboratory of Natural Medicine and Immune-Engineering of Henan Province,
12 Henan University, Kaifeng, Henan, 475004, P. R. China

13 ^e Shen Zhen Prevention and Treatment Center for Occupatic, Shenzhen, Guangdong,
14 518000, China

15

16 * Corresponding author: School of Public Health, Zhengzhou University, 100 Science
17 Avenue, Zhengzhou, China, 450001. Tel.: +86-371-67781922; fax:
18 +86-371-67781922. E-mail address: yaowu@zzu.edu.cn.

19

Abstract

This work is devoted to the applications of the multiple linear regression (MLR), multilayer perceptron neural network (MLP NN) and projection pursuit regression (PPR) to quantitative structure–property relationship analysis of bioconcentration factors (BCFs) of pesticides tested on Bluegill (*Lepomis macrochirus*). Molecular descriptors of a total of 107 pesticides were calculated with the DRAGON Software and selected by inverse enhanced replacement method. Based on the selected DRAGON descriptors, a linear model was built by MLR, nonlinear models were developed using MLP NN and PPR. The robustness of the obtained models was assessed by cross-validation and external validation using test set. Outliers were also examined and deleted to improve predictive power. Comparative results revealed that PPR achieved the most accurate predictions. This study offers useful models and information for BCF prediction, risk assessment, and pesticide formulation.

Keywords: bioconcentration factor, multiple linear regression, multilayer perceptron neural network, projection pursuit regression

1. Introduction

Bioconcentration factor (BCF) is the equilibrium ratio of the concentration of a substance in an exposed organism to the concentration of a dissolved substance bioavailable in the surrounding aquatic environment (Mackay and Fraser, 2000). The European regulation on classification, labeling, and packaging requires the establishment of BCFs for all compounds to regulate chemical substances (European Commission, Environment Directorate General, 2007). For fish, BCF value facilitates the estimation of pesticide daily intake by the daily consumption of fish and by establishing safe limits for water pesticide concentration. Pesticides are used in agricultural treatments to improve crop yields (Cooper and Dobson, 2007). Nevertheless, the use of pesticides should be controlled because an important fraction of pesticides are released into the environment, which presents a potential hazard (Mnif et al., 2011; Landrigan et al., 1999). Therefore, investigating BCFs of pesticides is important.

However, experimental determination of BCF is expensive and time-consuming. Thus, a few estimation methods have been reported. Several studies have reported on the linear correlation between BCF values and n-octanol/water partition coefficient ($\log K_{ow}$) (Devillers et al., 1996; Mackay, 1982; Garg and Smith, 2014). Dimitrov et al. (2002) proposed a nonlinear model for the $\log BCF/\log K_{ow}$ relationship based on a large set of narcotic compounds to analysis the specific role of the internal water phase on the formation of the body bioconcentration. Gissi et al. (2013) integrated the most used CAESAR and Meylan models for predicting BCF of 851 compounds from

the ANTARES BCF dataset to demonstrate integrated model with more reliability of the predictions. Moreover, a chromatographic retention factor (Bermúdez-Saldaña et al., 2005) and an artificial membrane accumulation index (Fujikawa et al., 2009) have been reported. Piir et al. (2010) summarized some modeling techniques that use different compounds and algorithms. However, only one study (Jackson et al., 2009) predicted the BCF of pesticides using $\log K_{ow}$ and E-state descriptors. We propose here three alternative models for predicting BCFs of pesticides and providing useful information for risk assessment and pesticide research.

Selecting a combination of variables that produce the best result for QSAR study is also one of the most important problems. Feature selection methods such as simulated annealing, genetic algorithm, replacement method (RM), enhanced replacement method (ERM), and many more are used (González et al., 2008). Among these techniques, a modified ERM, inverse enhanced replacement method (IERM), is simple and of low computational cost (Helguera et al., 2009). Thus, IERM was employed to select the best DRAGON descriptors for QSAR models. Multiple linear regression (MLR), multilayer perceptron neural network (MLP NN) and projection pursuit regression (PPR) have attracted attention and have been extensively applied (Torrecilla et al., 2009; Goodarzi et al., 2010; Parinet et al., 2015; Yan et al., 2015; Du et al., 2008). Therefore, MLR, MLP NN and PPR were also employed in this paper.

This study was performed to develop linear and nonlinear QSAR models for predicting the $\log$ BCF of pesticides. The presented QSAR models were validated

81 using a test set and the proposed parameters by Tropsha (2010). The performances of
82 the different models were compared. In addition, the descriptors used in the QSAR
83 models were used to identify molecular characters related to the log BCF of
84 pesticides.

85

2. Materials and Methods

2.1 Data set

Molecular structures and activities of pesticides were derived from Jackson et al. (2009). One duplicate compound and a salt (the neutral form of this salt could not be optimized by the SYBYL program) were removed. A total of 107 pesticides were evaluated in this study. BCF values were measured on the same Bluegill (*Lepomis macrochirus*) following a standardized protocol (US EPA, 1996). BCF values are presented as log values in Table S1 (Table S1 in the Supplementary material). The modeled log BCF ranged from -0.92 to 4.00. The compounds were divided into training ($n=80$) and test ($n=27$) sets using DUPLEX algorithm (Snee, 1977; Puzyn et al., 2011), which allows maintenance of a comparable diversity in both sets; the latter are therefore similar in terms of representativeness (for more details, see the Supplementary material and elsewhere (Snee, 1977; Ritota et al., 2010)). The training set was used to build models, and the independent test set was used to evaluate the predictive ability of the models.

2.2 DRAGON descriptors

The three-dimensional structures of 107 pesticides were constructed using Sybyl-x 1.3. The energy of each molecule was minimized using gradient descent method and by employing Tripos force field and Gasteiger–Huckel charges. The molecular descriptors were obtained by encoding the optimized molecular structures into the DRAGON program, and 4885 different types of molecular descriptors were

calculated to describe the structural diversity of the chemicals. Then, the initial pool of descriptors was reduced by applying the DRAGON built-in variable exclusion procedure. Three types of descriptors were excluded, namely, constant (relative standard deviation < 0.001), near-constant (all values are equal except one), and highly correlated descriptors. For each pair of highly correlated descriptors, $R > 0.9$, the descriptor with the largest mean correlation coefficient with the rest of the descriptors is removed. Thus, 1230 molecular descriptors were maintained.

2.3 Variable selection

We used IERM to select the best subset of DRAGON descriptors. The RM and ERM determined the optimal subset of d ($d \ll D$) descriptors from a large group of D descriptors with a minimum standard deviation (S) (Mercader et al., 2010; Mercader et al., 2011). By contrast, IERM uses an initial set of very high S determined by an “inverse RM” (RM was used to maximize S), as follows:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^N res_i^2 \quad (1)$$

In Eq. (1), N is the number of molecules in the training set, and res_i represents the residual for molecule i . The residual is the difference between experimental and predicted properties. $S(\mathbf{d}_n)$ is a distribution in a space of $D!/ [d!(D-d)!]$ points. Full search (FS) can arrive at the global minimum by calculating $S(\mathbf{d}_n)$ for all space points, but FS is too difficult to perform if D is very large. However, IERM is more efficient than FS in reaching the global minimum. The two-step IERM was performed as follows. First, an initial set of descriptors, $\mathbf{d}_k$, was chosen by an inverse RM. One of

the descriptors, X_{ki} , was replaced iteratively by each remaining $D - d$ descriptor, and the set with the smallest value of S was retained. Second, from the resulting set, the descriptor with the greatest value of S in its coefficient was chosen and replaced using all descriptors, except the one replaced in the previous iteration. This step was repeated until the set remained unchanged. IERM variable selection was implemented in MATLAB.

2.4 MLP NN

MLP NN, which is a supervised neural network, was used to investigate the nonlinear relationship among the selected molecular descriptors and log BCF of pesticides. This network generally consists of several artificial neurons arranged in three layers (topology of NN), namely, input, hidden, and output layers. The three-layer topology with a single hidden layer is sufficient for solving similar or more complex problems (Torrecilla et al., 2009). Thus, three-layer MLP was used in this work. The descriptors of the MLR model were used as inputs for the network. The Broyden–Fletcher–Goldfarb–Shanno learning algorithm was used to develop MLP NN models. Different networks with 4–12 neurons in the hidden layer were trained to determine the optimal number of neurons in the hidden layer. Overfitting was avoided by repeating the learning process and verifying the sum-of-squares error. The sum-of-squares error is simply given by the sum of differences between the target and prediction outputs defined over the entire training set. The calculation process in each neuron in the hidden and output layers is performed by successive activation and

transfer functions. In this work, four activation functions (i.e., identity, logistic sigmoid, hyperbolic tangent, and negative exponential functions) were tested for hidden and output activation functions. We employed leave-many-out cross-validation method (LMO CV) to evaluate the robustness of the model.

2.5 PPR

Friedman and Stuetzle (1981) developed PPR, which is a powerful tool for exploring the projections of high-dimensional data into low-dimensional space, thereby overcoming the problem of dimensionality. The basic theory of PPR can be found elsewhere (Friedman and Stuetzle, 1981; Ren et al., 2007). Four parameters “nterms,” “max. terms,” “optlevel,” and “span,” that should be adjusted were introduced in this study. The parameter “nterms” controls the number of variables to be entered in the model, and “max. terms” is the maximum number of terms to choose from when building the model. Meanwhile, “optlevel” indicates the levels of optimization that differ in how thoroughly the models are refitted during this process, and “span” defines the fraction of the observations in the span of running the lines smoothly.

2.6 Variation inflation factor (VIF) and standard regression coefficient (SR)

VIF can be used to evaluate the correlation degree of each independent variable in the equation (Yuan et al., 2016). If $VIF = 1.0$, no intercorrelation exists for each variable; if VIF ranges from 1.0 to 5.0, the related equation is acceptable; and if VIF

is larger than 10.0, the regression equation is unstable and a recheck is necessary. The values of a SR can identify which independent variables have the effect on the dependent variable, irrespective of differing units of measurement. Generally, the larger the absolute SR, the more important is the descriptor. The calculation of VIF and SR can be found elsewhere (Jouan-Rimbaud et al., 1995; Curto and Pinto, 2007). VIF and SR were calculated using SPSS 18.0 software.

2.7 Model evaluation

Diagnostic statistical tools, namely, goodness of fit and prediction, were used to evaluate the performance of the models. The goodness of fit was examined by the coefficient of determination (r^2), root mean square error (RMSE), Fisher's ratio (F), and the ratio of the number of cases to the number of adjustable parameters in the model (known as the p statistics, which should be ≥ 5) (Dearden et al., 2009).

The goodness of the prediction of models was evaluated to predict some new compounds (i.e., test set).

Cross-Validations (CV) is the most commonly used techniques for internal validation. The outcome from the CV procedure is cross-validated r^2 (q^2), which is used as a criterion of both robustness and predictive ability of the model. The leave-many-out cross-validation (LMO CV) is regarded as more robust and reliable than the leave-one-out cross-validation (Gramatica, 2007). Therefore, LMO CV was used. LMO CV is a statistical method in which more than one chemical are omitted from the training set and iteratively put in test set (Gramatica, 2007). In this study,

LMO CV was performed by using five groups, and the procedure was repeated 10 times. The mean of 10 readings was given as $q^2_{\text{LMO-5}}$. The mean RMSE of readings of LMO CV was RMSECVm.

Several proposed parameters by Tropsha (Tropsha, 2010; Golbraikh et al., 2003) were used to further test the predictive ability of the developed QSAR models. According to Tropsha, a QSAR model is successful if it satisfies the following criteria:

$$q^2 > 0.5 \quad (2a)$$

$$r^2 > 0.6 \quad (2b)$$

$$(r^2 - r_0^2) / r^2 < 0.1 \text{ and } 0.85 \leq k \leq 1.15 \quad (2c)$$

or

$$(r^2 - r_0'^2) / r^2 < 0.1 \text{ and } 0.85 \leq k' \leq 1.15 \quad (2d)$$

$$|r_0^2 - r_0'^2| < 0.3 \quad (2e)$$

Here, r^2 and r_0^2 are the squared correlation coefficient values between the observed (y-axis) and predicted activity (x-axis) data with and without the intercept, respectively. The change in the axes provides the value of $r_0'^2$. Parameters k and k' indicate the slopes in the former and later cases, respectively. The calculation of these parameters can be found in the supplementary data.

2.8 Applicability domain (AD)

The AD of a QSAR model must be defined if the model is to be used for screening new compounds. The Williams plot, i.e. the plot of cross-validated

standardized residuals versus leverage values (h_i), was used. The h_i of a chemical in the original variable space is defined as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i=1, \dots, n), \quad (3)$$

where x_i is the descriptor vector of the considered compound, X is the model matrix derived from the training set descriptor values. In the Williams plot, the AD was established inside a squared area within $\pm m$ standard deviations and a “warning leverage” h^* ($m = 3$, $h^* = 3(p+1)/n$, with p as the number of model parameters and n as the number of training compounds) as stated by Gramatica (2007) and Netzeva et al., (2005). Moreover, the Williams plot can verify the presence of response outliers (*i.e.*, compounds with cross-validated standardized residuals greater than 3σ) and chemicals very structurally influential in determining model parameters (the limit of structural outliers was h^*).

3. Results and Discussion

3.1 MLR modeling

To select the set of descriptors, the linear models with the number of descriptors from one to nine were built. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the best subset size was obtained. (Yuan et al., 2009) The influences of the number of the descriptors on the r^2 and the RMSE are shown in Fig. S1. From Fig. S1, it can be seen that eight descriptors appear to be sufficient for a satisfactory model. Thus, eight DRAGON descriptors were selected by IERM and used to build an MLR model. The model and statistical parameters of regression are as follows:

$$\begin{aligned} \text{Log BCF} = & -0.752(\pm 0.136)\text{B08}[\text{C-Cl}] - 0.540(\pm 0.149)\text{B01}[\text{C-N}] - \\ & 0.898(\pm 0.167)\text{CATS2D_01_NL} + 0.103(\pm 0.014)\text{CATS2D_08_LL} + \\ & 0.344(\pm 0.040)\text{C-026} + 1.414(\pm 0.228)\text{R2u} + 0.334(\pm 0.054)\text{nCL} - 0.504(\pm 0.060)\text{O-058} \\ & - 0.613(\pm 0.522) \end{aligned} \quad (4)$$

where $N = 80$, $r^2 = 0.851$, $\text{RMSE} = 0.398$, $F_{(8,71)} = 50.492$, $p < 10^{-25}$, $q^2_{\text{LMO-5}} = 0.809$, $\text{RMSECVm} = 0.454$, and $\rho = 10$.

Moreover, multicollinearity among the descriptors in the Eq. (4) model was tested by their VIFs. SR was calculated to determine the order of importance of descriptors in Eq. (4). The VIF and SR values are listed in Table S2 (Table S2 in the Supplementary material). The VIF values (< 5) of all descriptors suggest that Eq. (4) is acceptable. The SRs in Eq. (4) were used to assign the importance to the variables of the model, and the following ranking of contributions to log BCF was achieved:

O-058 > C-026 > nCL > R2u > CATS2D_08_LL > B08[C-Cl] > CATS2D_01_NL > B01[C-N].

Then, the built model was used to predict the test set data. The r^2_{pred} ($r^2_{\text{pred}} = 0.547$) showed unsatisfactory prediction results. Thus, the outliers in the dataset were further evaluated by the Williams plot (Fig. 1). The Williams plot evidently shows two compounds (Nos. 24 and 25 in the training set) that had leverage values higher than the warning h^* value of 0.338. Thus, these values could be considered as structural outliers. In this case, the data predicted by Eq. (4) had small residuals ($<3\sigma$) for compounds 24 and 25, so these compounds are “good leverage” chemicals. According to Jaworska et al., (2005), such “good leverage” chemicals stabilize and enhance the precision of the model. Therefore, compounds 24 and 25 were still included in the training set.

The values of σ of compounds 81, 86, 95, and 97 were larger than 3σ , so these compounds were considered response outliers in the test set (Fig. 1). After removing these four outliers, the r^2_{pred} of Eq. (4) was 0.770. Establishing the factor for the failure of the model to predict these outliers accurately is challenging. The anomalous behavior of these compounds could be due to one of the following factors: (1) incorrect experimental input data; (2) selected descriptors did not capture some relevant structural features in the molecules that are absent in some molecules; or (3) the biological mechanism of the four compounds is different from that of other chemicals. For future predictions, the predicted log BCF values can be considered reliable only for chemicals that fall within the AD.

(Figure 1 here)

The ability of a model to predict the property values in the training set is enhanced as the magnitude of the F ratio increases. The large F ratio of 50.492 indicates that Eq. (4) performed an excellent prediction. The $r^2 = 0.851$ indicated good power of prediction, and $q^2_{\text{LMO-5}} = 0.809$ confirmed the robustness of the model. Several statistical parameters for the removed outlier test set are also listed in Table 1. These parameters obviously satisfy the generally accepted condition and thus demonstrate the predictive power of Eq. (4).

(Table 1 here)

Finally, Eq. (4) was used to predict log BCF. The calculated log BCF from the MLR model for the training and test sets are listed in Table S1 and shown in Fig. 2A. The points are distributed on both sides of the diagonal line, which suggests the absence of systematic error in the model development.

(Figure 2 here)

3.2 MLP NN model

Nonlinear models were then developed by submitting the selected descriptors from MLR to the three-layer MLP NN. An 8–5–1 network architecture was obtained after testing different groups of neurons in the inner layer. In the best network, the hidden function was an exponential function, and the output function was a hyperbolic tangent, with $R^2=0.866$, $\text{RMSE}=0.376$, $q^2_{\text{LMO-5}}=0.506$, and $\text{RMSECV}_{\text{m}}=0.424$ for the training set.

The “black-box” problem of ANN method, in which the exact relationship between the input and output data is difficult to determine, was resolved. Sensitivity analysis was performed to determine the order of importance of descriptors in the MLP NN model (Guha and Jurs, 2005). The differences (Rmdiffi) between the RMSE of the prediction of the complete network and RMSE were obtained when the *i*th variable was excluded from the trained network (RMSE_{*i*}). These differences were calculated as follows: Rmdiffi: $Rmdiff_i = RMSE_i - RMSE$. The most important variable was the one that resulted in the highest value of Rmdiffi. The values of Rmdiffi for the MLP NN model were calculated and are plotted in Fig. 3A. The orders of importance of the selected molecular descriptors are as follows: C-026 > CATS2D_08_LL > O-058 > R2u > nCL > B08[C-Cl] > CATS2D_01_NL > B01[C-N].

(Figure 3 here)

The built MLP NN model was used to predict the outliers previously removed from the MLR model. However, the MLP NN model still failed to accurately predict the outliers. After the outliers were omitted, the proposed MLP NN model was predictive because it satisfied the recommended conditions by Tropsha for the test set. The statistical parameters are listed in Table 1. The predictive results from the MLP NN model for the entire dataset are listed in Table S1 and are shown in Fig. 2B.

3.3 PPR model

PPR was also applied to build the non-linear model with the eight descriptors in

Section 3.1 as input variables. In this investigation, “nterms,” “max. terms,” and “optlevel” were determined as 3, 4, and 1, respectively. The parameter “span” was set to default value (0). The model yielded the following values for the training set: r^2 , 0.915; RMSE, 0.299; $q^2_{\text{LMO-5}}$, 0.722; and RMSECVm, 0.555. After the outliers were deleted from the MLR model, the proposed PPR model yielded an r^2 of 0.794 for the test set. This PPR model satisfies the recommended conditions by Tropsha (Table 1). Therefore, this PPR model was used to predict the log BCF of pesticides, and the predicted results are shown in Table S1 and Fig. 2C. Sensitivity of the PPR model was also analyzed. The Rmdiffi value of each descriptor is shown in Fig. 3C. The orders of importance of selected molecular descriptors are as follows: CATS2D_08_LL > nCL > O-058 > C-026 > R2u > B08[C-Cl] > B01[C-N] > CATS2D_01_NL.

3.4 Comparison with other model

The qualities of the three presented models were further evaluated by comparing these models from that in the literature (Jackson et al., 2009) (Table 2). MLR, MLP NN, and PPR models were used to predict log BCF of pesticides. The r^2 values of the three models in this study for the training set were all better than that reported by Jackson et al. (2009). However, the r^2 values of the three models for the test set were slightly inferior to that reported by Jackson et al. (2009). In addition, the three presented models used only eight descriptors. Thus, the V and ρ values in this study were better than previously reported results (Jackson et al., 2009). Comparative

results of the three methods indicated that PPR was the most accurate approach.

To further evaluate the quality of the presented models, another comparison was carried out, too. The log BCF values (Table 1S) were predicted by the regression-based model included in the EPI SuiteTM (<http://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>).

The r^2 for training and test sets were compared and shown in Table 2. The r^2 values of the three presented models for the training and test sets were all better than that of the regression-based model included in the EPI SuiteTM. These results indicated that the presented models provided useful tools for predicting log BCF of pesticides.

(Table 2 here)

3.5 Descriptor analysis and interpretation

Thus, the significance of the descriptors in the MLR, MLP NN, and PPR models decreased in varying orders. The difference in ordering was probably due to the fact that different approaches establish different correlations between the selected descriptors. The pie charts in Figs. 3B and 3D show that all relative contributions are close to the theoretical mean of 12.5%, although the relative contributions of eight descriptors varied (10%–15%). No evident difference was found in the relative contributions of descriptors. Therefore, each descriptor is important for predicting log BCF in both MPL NN and PPR models.

B08[C–Cl] and B01[C–N] belong to 2D atom pair descriptors, which correspond to the presence/absence of C–Cl at the topological distance of 8 and the

presence/absence of C–N at topological distance of 1. nCl, which belongs to constitutional indices, is the number of chlorine atoms. Descriptors B08[C–Cl], B01[C–N], and nCl provided information on the distribution of atomic properties along the topological structure. The weight properties indicated polarizability and polarity, which are in accordance with the results of Lu et al. (2000). CATS2D_01_NL and CATS2D_08_LL are chemically advanced template search (CATS) 2D descriptors, which indicate CATS2D negative-lipophilic at lag 01 and CATS2D lipophilic–lipophilic at lag 08, respectively. CATS2D_01_NL presented negatively charged or ionizable and lipophilic substructures. CATS2D_08_LL presented the lipophilicity of a compound. This result was consistent with the structure–activity correlation study by Veith et al. (1979), in which BCF was estimated by log P (The logarithm of the n-octanol/water partition coefficient, log P, provides indirect information on the hydrophilic or lipophilic properties of a molecule). C-026 corresponds to R–CX–R, representing the number of substituent groups bonded to the benzene ring, but excluding those bonded by a carbon atom to carbon atoms, i.e., $C_{\text{benzene ring}} - C_{\text{substituent group}}$. O-058 is the number of double-bonded oxygen atoms. Descriptors C-026 and O-058 belong to the atom-centered fragment descriptor, and these descriptors can induce higher hydrosolubility of pesticides. Descriptor R2u is classified into the Geometry, Topology, and Atom-weight Assembly molecular descriptors, which try to match 3D molecular geometry. Descriptor R2u is correlated with the steric character of a molecule.

In Eq. (4), CATS2D_08_LL had a positive coefficient to log BCF, contrary to the

negative coefficient of descriptor CATS2D_01_NL. The opposite contributions of CATS2D_08_LL and CATS2D_01_NL should be highlighted although they are related to lipophilicity. The balance of lipophilicity for reaching the biological target and exerting its action was controlled by CATS2D_08_LL and CATS2D_01_NL. Similar phenomenon was also observed in C-026 and O-058, which controls the balance of hydrosolubility. To be absorbed in fish, pesticides should penetrate a sequence of hydrophobic and hydrophilic barriers. Therefore, bioconcentration process is controlled by lipophilicity, and hydrosolubility substructure of molecules is reasonable.

In summary, polarity, lipophilicity, hydrosolubility, and steric character, as well as the R–CX–R, C – Cl, and C – N substations of molecules, play important roles in the BCF of fish.

4. Conclusion

MLR, MLP NN, and PPR models were used to develop linear and nonlinear QSAR models for predicting the log BCF of selected pesticides. The obtained models clearly demonstrated eight selected DRAGON descriptors by IERM that are strongly related to the molecular structure of pesticides. The prediction results indicate that the PPR method is the most powerful and promising tool for BCF analysis of pesticides. The selected descriptors provided insight into the structural features related to log BCF. Therefore, the presented models not only predict log BCF but also provide information for the computer-aided design of pesticides with low log BCF.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation (Grant No. 2014M550387), Postdoctoral Research Sponsorship in Henan Province (Grant No. 2013014), the Joint Fund of National Natural Science Foundation of China and Henan Province (U1404508), National Natural Science Foundation of China (NSFC, nos. 81273039 and 81273122), Scientific and Technological Key Project Education Department of Henan Province (Grant No. 14B350005), National Science Foundation for Young Scientists of china (Grant No. 81302461), Jiangsu Natural Science Foundation for young scholars (Grant No. BK2012348), Scientific Research Foundation of Henan University (Grant No. 2013YBZR039), and the open project of Key Laboratory of Environmental Medicine Engineering of Ministry of Education in Southeast University.

References

- Bermúdez-Saldaña, J.M., Escuder-Gilabert, L., Medina-Hernández, M.J., Villanueva-Camañas, R.M., Sagrado, S., 2005. Modelling bioconcentration of pesticides in fish using biopartitioning micellar chromatography. *J. Chromatogr. A* 1063, 153–160.
- Cooper, J., Dobson, H., 2007. The benefits of pesticides to mankind and the environment. *Crop Prot.* 26, 1337–1348.
- Curto, J.D., Pinto, J.C., 2007. New multicollinearity indicators in linear regression models. *Int. Stat. Rev.* 75, 114–121.
- Dearden, J.C., Cronin, M.T.D., Kaiser, K.L.E., 2009. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* 20, 241–266.
- Devillers J., Bintein, S., Domine, D., 1996. Comparison of BCF models based on log *P*. *Chemosphere* 33, 1047–1065.
- Dimitrov, S.D., Mekenyan, O.G., Walker, J.D., 2002. Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals. *SAR QSAR Environ. Res.* 13, 177–184.

445 Du, H., Wang, J., Hu, Z., Yao, X., Zhang, X., 2008. Prediction of fungicidal activities
446 of rice blast disease based on least-squares support vector machines and project
447 pursuit regression. *J. Agric. Food Chem.* 56, 10785–10792.

448

449 European Commission Environment Directorate General, 2007. REACH in Brief.
450 Brussels, Belgium.

451

452 Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. *J. Am. Stat. Assoc.*
453 76, 817–823.

454

455 Fujikawa, M., Nakao, K., Shimizu, R., Akamatsu, M., 2009. The usefulness of an
456 artificial membrane accumulation index for estimation of the bioconcentration
457 factor of organophosphorus pesticides. *Chemosphere* 74, 751–757.

458

459 Garg, R., Smith, C.J., 2014. Predicting the bioconcentration factor of highly
460 hydrophobic organic chemicals. *Food Chem. Toxicol.* 69, 252–259.

461

462 Gissi, A., Nicolotti, O., Carotti, A., Gadaleta, D., Lombardo, A., Benfenati, E., 2013.
463 Integration of QSAR models for bioconcentration suitable for REACH. *Sci. Total*
464 *Environ.* 456, 325–332.

465

466 Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.D., Lee, K.H., Tropsha, A., 2003. Rational

467 selection of training and test sets for the development of validated QSAR models. J.
 468 Comput. Aid. Mol. Des. 17, 241–253.
 469
 470 González, M.P., Terán, C., Saíz-Urra, L., Teijeira, M., 2008. Variable selection
 471 methods in QSAR: an overview. Curr. Top. Med. Chem. 8, 1606–1627.
 472
 473 Goodarzi, M., Ortiz, E.V., Coelho, L.D., Duchowicz, P.R., 2010. Linear and
 474 non-linear relationships mapping the Henry's law parameters of organic pesticides.
 475 Atmos. Environ. 44, 3179-3186.
 476
 477 Gramatica, P., 2007. Principles of QSAR models validation: internal and external.
 478 QSAR Comb. Sci. 26, 694–701.
 479
 480 Guha, R., Jurs, P.C., 2005. Interpreting computational neural network QSAR models:
 481 a measure of descriptor importance. J. Chem. Inf. Model. 45, 800–806.
 482
 483 Helguera, A.M., Duchowicz, P.R., Pérez, M.A.C., Castro, E.A., Cordeiro, M.N.D.S.,
 484 González, M.P., 2006. Application of the replacement method as novel variable
 485 selection strategy in QSAR. 1. carcinogenic potential. Chemom. Intell. Lab. Syst.
 486 81, 180–187.
 487
 488 Jackson, S.H., Cowan-Ellsberry, C.E., Thomas, G., 2009. Use of quantitative

489 structural analysis to predict fish bioconcentration factors for pesticides. *J. Agric.*
 490 *Food Chem.* 57, 958–967.

491

492 Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T., 2005. QSAR applicability
 493 domain estimation by projection of the training set in descriptor space: a review.
 494 *ATLA-Altern. Lab. Anim.* 33, 445–459.

495

496 Jouan-Rimbaud, D., Massart, W.L., Leardi R., De Noord, O.E., 1995 Genetic
 497 algorithms as a tool for wavelength selection in multivariate calibration. *Anal.*
 498 *Chem.* 67, 4295–4301.

499

500 Landrigan, P.J., Claudio, L., Markowitz, S.B., Berkowitz, G.S., Brenner, B.L., Romero,
 501 H., Wetmur, J.G., Matte, T.D., Gore, A.C., Godbold, J.H., Wolff, M.S., 1999.
 502 Pesticides and inner-city children: exposures, risks, and prevention. *Environ Health*
 503 *Perspect.* 107 (Suppl 3), 431–437.

504

505 Lu, X., Tao, S., Hu, H., Dawson, R.W., 2000. Estimation of bioconcentration factors
 506 of nonionic organic compounds in fish by molecular connectivity indices and
 507 polarity correction factors. *Chemosphere* 41, 1675–1688.

508

509 Mackay, D., Fraser, A., 2000. Bioaccumulation of persistent organic chemicals:
 510 mechanisms and models. *Environ. Pollut.* 110, 375–391.

511

512 Mackay, D., 1982. Correlation of bioconcentration factors. *Environ. Sci. Technol.* 16,
513 274–278.

514

515 Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., 2010. Replacement
516 method and enhanced replacement method versus the genetic algorithm approach
517 for the selection of molecular descriptors in QSPR/QSAR theories. *J. Chem. Inf.*
518 *Model.* 50, 1542–1548.

519

520 Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., 2011. Advances in
521 the replacement and enhanced replacement method in QSAR and QSPR theories. *J.*
522 *Chem. Inf. Model.* 51, 1575–1581.

523

524 Mnif, W., Hassine, A.I.H., Bouaziz, A., Bartegi, A., Thomas, O., Roig, B., 2011.
525 Effect of endocrine disruptor pesticides: a review. *Int. J. Environ. Res. Public*
526 *Health* 8, 2265–2303.

527

528 Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P.,
529 Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G.,
530 Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W.,
531 Stanton, D.T., van de Sandt, J.J.M., Tong, W.D., Veith, G., Yang, C.H., 2005.
532 Current status of methods for defining the applicability domain of (quantitative)

533 structure-activity relationships. *ATLA-Altern. Lab. Anim.* 33, 1–19.

534

535 Parinet, J., Julien, M., Nun, P., Robins, R.J., Remaud, G., Höhener, P., 2015.

536 Predicting equilibrium vapour pressure isotope effects by using artificial neural

537 networks or multi-linear regression – A quantitative structure property relationship

538 approach. *Chemosphere* 134, 521–527.

539

540 Piir, G., Sild, S., Roncaglioni, A., Benfenati, E., Maran, U., 2010. QSAR model for

541 the prediction of bio-concentration factor using aqueous solubility and descriptors

542 considering various electronic effects. *SAR QSAR Environ. Res.* 21, 711–729.

543

544 Puzyn, T., Mostrag-Szlichtyng, A., Gajewicz, A., Skrzyński, M., Worth, A.P., 2011.

545 Investigating the influence of data splitting on the predictive ability of

546 QSAR/QSPR models. *Struct. Chem.* 22, 795–804.

547

548 Ren, Y., Liu, H., Yao, X., Liu, M., 2007. Prediction of ozone tropospheric degradation

549 rate constants by projection pursuit regression. *Anal. Chim. Acta* 589, 150–158.

550

551 Snee, R.D., 1977. Validation of regression models: methods and examples,

552 *Technometrics*, 19, 415–428.

553

554 Torrecilla, J.S., García, J., Rojo, E., Rodríguez, F., 2009. Estimation of toxicity of

ionic liquids in leukemia rat cell line and acetylcholinesterase enzyme by principal component analysis, neural networks and multiple lineal regressions. J. Hazard. Mater. 164, 182–194.

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inf. 29, 476–488.

US EPA, 1996. ecological effects test guidelines, OPPTS 850.1730 fish BCF, http://www.epa.gov/opptsfrs/publications/OPPTS_Harmonized/850_Ecological_Effects_Test_Guidelines/Drafts/.

Veith, G.D., Defoe, D.L., Bergstedt, B.V., 1979. Measuring and estimating the bioconcentration factor of chemicals in fish. J. Fish. Res. Board Can. 36, 1040–1048.

Yan, F., Shang, Q., Xia, S., Wang, Q., Ma, P., 2015. Topological study on the toxicity of ionic liquids on *Vibrio fischeri* by the quantitative structure–activity relationship method. J. Hazard. Mater. 286, 410–415.

Yuan J., Yu S., Zhang T., Yuan X., Cao Y., Yu X., Yang X., Yao W., 2016. QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls, Ecotox. Environ. Safe. 128,

577 171–180.

578

579 Yuan, Y., Zhang, R., Hu, R., Ruan, X., 2009. Prediction of CCR5 receptor binding
580 affinity of substituted 1-(3,3-diphenylpropyl)-piperidiny amides and ureas based
581 on the heuristic method, support vector machine and projection pursuit regression.
582 Eur. J. Med. Chem. 44, 25–34.

583

584 Table 1. Some statistical parameters for the test set.

models	r^2	$(r^2 - r_0^2)/r^2$	k	$(r^2 - r_0'^2)/r^2$	k'	$ r^2 - r_0^2 $
MLR	0.770	0.013	1.093	0.020	0.876	0.010
MLP NN	0.778	0.039	1.041	0.002	0.918	0.031
PPR	0.794	0.033	1.048	0.002	0.914	0.027

585

586

587 Table 2. Statistical parameters comparison of models

References	N	Descriptor type	method	V	ρ	r^2 (training/test)
Jackson et al., 2009	93/16	$\log k_{ow}$ +E-state	MLR	16	5.8	0.80/0.88
This work	80/27	DRAGON	MLR	8	10	0.851/0.770
	80/27	DRAGON	MLP NN	8	10	0.867/0.778
	80/27	DRAGON	PPR	8	10	0.915/0.794
	80/27	-	EPI Suite	-	-	0.524/0.652

588 N, the number of compounds in the training/test set; V, number of variables in the
 589 model; ρ , compound-variables ratio

590

591

Figure captions

Figure 1. Williams plot based on the Eq. 4.

Figure 2. Plot of measured versus predicted log BCF values of the training and test set, (A) from the MLR model, (B) from the MLP NN model, (C) from the PPR model.

Figure 3. The Rmdiffi value of the each descriptor in the MLP NN (A) and the PPR (C) model; Pie chart depicting relative significance of the descriptors to the MLP NN (B) and PPR (D) model.

603

604

605

606

607