A Novel Approach to the Prediction of Biomagnification Factors Based on Molecular Structure Images

Ming Cai Zhang
Lanzhou University
Ling Zhu
Lanzhou University
Hong Lin Zhai (zhaihl@163.com)
Lanzhou University https://orcid.org/0000-0001-7088-4962
Ke Xin Bi
Lanzhou University
Bing Qiang Zhao
Lanzhou University

Research Article

Keywords: Biomagnification factors (BMF), Organochlorine pollutants, molecular structure, Tchebichef moment (TM), Quantitative structure-property relationship (QSPR)

Posted Date: August 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-803682/v1

A novel approach to the prediction of biomagnification factors based on molecular structure images

Ming Cai Zhang, Ling Zhu, Hong Lin Zhai*, Ke Xin Bi, Bing Qiang Zhao

College of Chemistry & Chemical Engineering, Lanzhou University, Lanzhou, 730000, PR China

Corresponding author

Prof. Hong Lin Zhai

Tel.: +86 931 8912596; fax: +86 931 8912582; E-mail address: zhaihl@163.com

ORCID: 0000-0001-7088-4962

Abstract

Although biomagnification factor (BMF) is an important index of pollutants in food chains, its experimental determination is quite tedious. In this contribution, as the feature information, Tchebichef moments (TMs) were calculated directly from the molecular structural images, and then stepwise regression was employed to establish the prediction model of the *logBMF*. The proposed approach was applied to the *logBMF* prediction of organochlorine pollutants, and the correlation coefficient with leave-one-out cross-validation (R_{cv}) of the obtained model was 0.96, and the root mean square error ($RMSE_p$) for the external independent test set was 0.21. Compared with traditional two-dimensional (2D) quantitative structure-property relationship (QSPR) as well as the reported method, the proposed approach was more simple, accurate and reliable. This study

* Correspondence to: Tel.: +86 931 8912596; fax: +86 931 8912582; E-mail address: zhaihl@163.com (H.L. Zhai).

not only obtained the satisfactory prediction model for organochlorine pollutants, but also provided another effective approach to QSPR research.

Keywords: Biomagnification factors (BMF); Organochlorine pollutants; molecular structure; Tchebichef moment (TM); Quantitative structure-property relationship (QSPR)

Declarations:

Acknowledgement

This research did not receive any specific grant from funding agencies.

Availability of data and materials

The datasets used are available from the published literature, and the calculation programs are provided in Supplementary Information.

Competing interests

The authors declare that they have no competing interests.

Authors Contributions

Ming Cai Zhang: Conceptualization, Methodology, Software, Writing- Original draft preparation

Ling Zhu: Software

Hong Lin Zhai: Supervision, Writing- Reviewing and Editing; Corresponding author

Ke Xin Bi and Bing Qiang Zhao: Methodology, Software, Validation

1. Introduction

With the developments of modern industries, environmental pollution has attracted more and more attention due to its influence on human health. Persistent organic pollution (POP) is one of the main factors contributing to environmental pollutions. A great many POPs are organochlorine compounds (Rosa Vilanova 2001) that easily circulate into organisms with ecosystem cycles (Jepson et al. 2016; Paul D. Jepson 2009). Organochlorine pollutants can be accumulated along the food chains due to their stability, which causes a toxicity magnification in the organisms at the top of the food chains (also called Biomagnification phenomenon) (Birgit M. Braune 1989). To assess this toxicity of POPs, biomagnification factor (BMF) was defined and calculated by the following formula (D. Mackay 2000):

$$BMF = \frac{c_B}{c_A} \quad (1)$$

where c_B is the concentration of chemical in the organism and c_A is the concentration in the organism's diet.

Although the BMFs can be determined by the experimental approaches (Charles J. Henny 2003; Serrano et al. 2008; Woodburn et al. 2013), there are very cumbersome and time consuming. It is well known that the behaviors of a molecule are closely related to its chemical structure, and quantitative structure-property relationship (QSPR) and quantitative structure-activity relationship (QSAR) become a type of the useful strategies (Dearden 2016; Gramatica 2020). After calculation a large number of molecular descriptors, the most significant descriptors related to BMF were selected by genetic algorithm and used to build the predictive model of artificial neural network (Fatemi and Baher 2009). The BMFs of polybrominated diphenyl ethers (PBDEs) were predicted by means of QSAR method

(Mansouri et al. 2012). Acceptable-by-design QSAR method to predict the dietary BMF of organic chemicals in fish, in which two kinds of variable selection methods including genetic algorithms and reshaped sequential replacement were employed (Grisoni et al. 2019). Augmented multivariate Image Analysis applied to QSPR (aug-MIA-QSPR) approach was reported to predict the BMFs of aromatic organochlorine pollutants (da Mota et al. 2017), which indicated that the image of molecular structure could provide the useful information for the BMF prediction.

Image moment firstly proposed by Hu (Hu 1962) is one of the description methods for grayscale images, and then a series of image moments have been developed such as Zernike (Teague 1980), Tchebichef (Ramakrishnan Mukundan 2001) and Krawtchouk (Yap et al. 2003) moments. Although these moments are often used to the de-noising or compression in digital image processing, several image moments have been applied to the feature extraction of target information from chemical spectra and employed to establish the analytical models owing to their powerful multi-resolution as well as good invariance (Zhai et al. 2018). As an excellent member of moment family, Tchebichef image moment (TM, also called Chebyshev moment) possesses the more advantages of feature extraction.

In this study, as the novel descriptors of chemical structures, TMs were calculated directly from the gray images of the molecular structures of organochlorine compounds, and stepwise regression was employed to establish the linear model for the prediction of BMF. The performance of the obtained model was evaluated thoroughly and rigorously. Furthermore, the results from our approach were compared with that of other methods.

2. Data and methods

2.1 Data set

The data set was derived from the literatures (da Mota et al. 2017), which consisted of 30 polychlorinated biphenyls (PCBs) congeners and 10 organochlorine pesticides (DDT, DDE, HCB, TCDF, OCDF, TCDD, H6CDD, H7CDD, OCDD and DDD). Their values of *logBMF* are listed in Table 1 as Exp. column. All of 40 samples were randomly divided into training set (30 samples) and test set (10 samples). The training set was used to establish the prediction model, and the test set was employed to evaluate the prediction capability of the obtained model as external independent sample set.

2.2 Methods

2.2.1 Images of molecular structures

The two-dimensional (2D) molecular structures of the 40 compounds were drawn in ChemBioDraw (v12) software with default conditions (Fixed Length: 1.058 cm, Line Width: 0.035 cm, Bond Spacing: 12% of length, Hash Spacing: 0.095 cm, Font: Times New Roman, Size: 12) and saved as the grayscale BMP format with the size of 303 pixels × 258 pixels under the resolution ratio of 96 DPI.

2.2.2 Calculation of Tchebichef image moments

For a given grayscale image with size of $N \times M$, the TM can be calculated using the following formula (Bayraktar et al. 2007):

$$T_{n,m} = \frac{1}{\tilde{\rho}(n,N)\tilde{\rho}(m,M)} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \tilde{t}_n(x)\tilde{t}_m(y) f(x,y) \quad (n = 0,1,2,...N-1, \quad m = 0,1,2,...M-1) \tag{2}$$

where $\tilde{t}_n(x)$ and $\tilde{t}_m(y)$ are the normalized discrete Tchebichef polynomial of degree n and m, respectively; $\tilde{\rho}(n,N)$ is the squared-norm of the normalized polynomials and $f(x,y)$ is the image intensity function. Their detail calculation programs are provided in Supplementary Information.

Thus the reconstruction of image with $T_{n,m}$ can be performed:

$$\hat{f}(x,y)=\sum_{n=0}^{nN}\sum_{m=0}^{mM}T_{n,m}\tilde{t}_n(x)\tilde{t}_m(y) \quad (3)$$

where $\hat{f}(x,y)$ is the reconstructed image, nN and mM are the maximum orders of n and m (n=0-nN, nN<N-1; m=0-mM, mM<M-1). The reconstruction error ε can be calculated:

$$\varepsilon=\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left|f(x,y)-\hat{f}(x,y)\right| \quad (4)$$

2.2.3 Modeling and evaluation

Stepwise regression was employed to establish the linear prediction model, in which TMs were regarded as the independent variables and $logBMF$ was denoted as dependent variable. The performance of obtained model was evaluated by means of various statistical parameters such as the determination coefficient (R_c), the adjusted determination coefficient (R_{adj}), root mean square error ($RMSE_c$), the correlation coefficient with leave-one-out (LOO) cross-validation (R_{cv}) and LOO root mean square error ($RMSE_{cv}$) for training set; F-test for model and t-test for the regression coefficients; the correlation coefficient of test set (R_p) and root mean square error ($RMSE_p$) for test set (Gadaleta et al. 2016).

In order to further inspect the robustness of the model, a randomized test was performed on the established model, in which models are established with invariant X-matrix and randomized Y-matrix (Mitra et al. 2010). To determine the reliable of model, ${}^cR_P{}^2$ was adopted by following corrected formula (Todeschini 2010):

$${}^cR_p^2=R\sqrt{\left(R^2-R_r^2\right)} \quad (5)$$

where R is R_c of the model and $R_r{}^2$ is the average of R^2 for the randomized model.

The predictive capability of the model can be validated by external test, and the related parameters (k, k', $r_m{}^2$, $r_m{}'^2$ and $\Delta r_m{}^2$) are defined by follows (Ojha et al. 2011; Roy et al. 2013):

$$k = \frac{\sum\left(Y_{obs} \times Y_{pred}\right)}{\sum\left(Y_{pred}\right)^2} \quad (6)$$

$$k' = \frac{\sum\left(Y_{obs} \times Y_{pred}\right)}{\sum\left(Y_{obs}\right)^2} \quad (7)$$

$$r_m^2 = r^2\left(1 - \sqrt{r^2 - r_0^2}\right) \quad (8)$$

$$r_m'^2 = r^2\left(1 - \sqrt{r^2 - r_0'^2}\right) \quad (9)$$

$$\Delta r_m^2 = \left|r_m^2 - r_m'^2\right| \quad (10)$$

Here, k and k' is the slope of experimental and predicted values respectively. Y_{obs} and Y_{pred} are the observed and predicted values, respectively. r_m^2 and $r_m'^2$ are modified r^2. r^2 and $r_0'^2$ are determination coefficients between the observed and predicted values for the least square linear regression with and without interpret. And Δr_m^2 is the absolute of the difference between them.

Meanwhile, it is necessary to discuss the applicability domain (AD) of the established model to study its scope and limitations. In this work, Williams plot was employed to calculate the applicability domain (AD) of the established model, which presents the relationship between leverage (Hat matrix) and standardized residuals, and Hat matrix could be calculated by the following relation (P. 2007):

$$H = X(X^T X)^{-1} X^T \quad (11)$$

where X is the matrix composed of descriptors in the established model and T means the transpose matrix. In general, the threshold H^* of H is equal to $3p/n$ (n is the number of training set sample and p is the established model's variables number plus one), and the standardized residuals are normally accepted within the range ±3 (Roy et al. 2015).

2.2.4 Comparison with 2D QSPR as well as the reported method

Traditional 2D QSPR method was also applied to the same data set. The molecular descriptors of the 40 samples were calculated by CODESSA (v2.63) after being optimized by HyperChem (v7.5). A

linear QSPR model was established by stepwise regression based on the training set, and used to the prediction of the test set. The obtained results were compared with that of the proposed method.

The proposed TM model was also applied to predict the *logBMF* of the samples in the five different test sets as same as the reference (da Mota et al. 2017), and the calculated results were compared with that of the method in this reference.

All calculation programs were written in M-file based on MATLAB v7.0 (Mathworks Inc. USA), and carried out with PC (CPU 3.40 GHz, RAM 16.0 GB).

3. Results and discussions

3.1 Tchebichef moments and molecular structural images

Owing to the excellent description ability with multi-resolution and invariance properties in image processing, Tchebichef image moment (TM) is an important image characteristic based on the discrete orthogonal polynomials (Ramakrishnan Mukundan 2001). What is more, no numerical approximation is needed during the calculation.

TMs with different moment orders represent different information in image according to Eq. 2, which can decompose the information of molecular structure image (multi-resolution ability). Then the important features ($T_{n,m}$) related to the *logBMF* of chemical compounds could be selected by stepwise regression to establish the prediction model. On the other hand, TMs have the relative computational stability owing to its invariance property in the image operation of shifting, scaling and rotation (Mukundan et al. 2001), which means that, for the molecular structure images, whether the different drawing scales or the positions in the canvas has little influence on the results of TM calculation. This is a significant advantage in the QSPR studies based on molecular structure images, which could

guarantee the stability of the obtained feature information.

3.2 Model and evaluation

After the TMs were directly calculated from the grayscale images of molecular structures, the maximum orders were determined as nN=28 and mM=43 according to the change of reconstruction errors (Eq. 4). Then a linear quantitative model was established by stepwise regression based on the training set, in which the TMs were the independent variables and *logBMF* was the target response variable. The values of TMs in the following model are listed in Table S1.

$$log\ BMF = -1.0732 + 92.1940 \times T_{1,1} - 6.9062 \times T_{4,14} - 10.2245 \times T_{13,9} - 5.5451 \times T_{13,12} + 14.7530 \times T_{13,27}$$

The calculated *logBMF* values of all samples are listed Table 1 and the statistical parameters of the established model are summarized in Table 2. The linear relationship between the calculated values and the experimental values are shown in Fig. 1. As can be seen from Table 2, the R_c (0.9726), R_{adj} (0.9668) and $RMSE_c$ (0.1052) were satisfactory, which indicates that the model was accurate; R_{cv} (0.9570) and $RMSE_{cv}$ (0.1320) suggested that there was not over-fitting; the p-value of F-test (2.08e-14) showed the good linear relationship between the independent variables and response variable in this model, and the results of t-test ensured that the regression coefficients had statistical significance. For the test set, R_p (0.9594) and $RMSE_p$ (0.2129) represented that the established model possessed satisfactory predictive ability. All above statistical parameters indicated that the model had high reliability and accuracy.

To investigate the robustness and reliability of the TM model, the further evaluation was carried out. For randomized test, the parameter ${}^cR_P{}^2$ is 0.5988 (more than its threshold value of 0.5), indicating that the model has not randomness and fortuitousness. For the external test, the obtained parameters (listed in Table 2) also conform to the requirements ($0.85 \le k \le 1.15$; $0.85 \le k' \le 1.15$; $r_m{}^2 \ge 0.5$; $r_m{}'^2 \ge 0.5$;

189 $\Delta r_m^2 \leq$ 0.2) (Ojha et al. 2011). Besides, Williams plot is shown Fig. 2A. As it could be observed,
190 sample **8** (1, 2, 3, 4, 6, 7, 8, 9-Octachlorodibenzofuran) of the training set and sample **1** (2378TCDF) of
191 test set are outliers with high leverage. Compared with the structures of other chemicals in dataset,
192 sample **8** may be different with others so that they are not well modeled by adopted variables. Another
193 possible reason is the sample belongs to other type chemicals. To the sample **1**, it owns the same
194 structure with sample **8** so that the model may not well predict the value of *logBMF* of it.

195 All above results and discussions indicated that the proposed method was reliable and reasonable,
196 and the established model possessed the higher robustness and prediction ability.

197 3.3 Comparison with other methods

198 3.3.1 Comparison with 2D-QSPR method

199 Based on the obtained 337 common molecular descriptors (Supporting information, Table S2), the
200 prediction model was established by stepwise regression as follows:

201 $$\begin{aligned} log\ BMF = {} & 57.2032 + 0.0105 \times X_{39} - 0.3412 \times X_{162} - 499.8468 \times X_{174} + 20.8516 \times X_{176} \\ & -42.3347 \times X_{276} - 36.4478 \times X_{327} \end{aligned}$$

202 where X_{39}, Wiener index; X_{162}, No. of occupied electronic levels; X_{174}, Avgnucleoph. react. Index for a
203 Cl atom; X_{176}, Max eletroph. react. index for a C atom; X_{276}, Avg bond order of a Cl atom; X_{327},
204 Principal moment of inertia A.

205 The calculated values of *logBMF* are also listed in Table 1. The obtained statistical parameters
206 (listed in Table 2) and Williams plot (showed in Fig.2B) illustrate that the established 2D-QSPR model
207 was robust and reliable. The comparison of statistical parameters in the Table 2 indicated that the TM
208 model was slightly better than the 2D-QSPR model the owing to its higher prediction ability, which
209 suggested the feasibility of the proposed approach.

3.3.2 Comparison with the method in reference

Compared with the best results obtained by aug-MIA-QSPR$_{color}$ method in the literature (da Mota et al. 2017), the statistical parameters (listed in Table 2) of the TM model had the more satisfied. For the five different test sets (named test **1**~**5**) used in this literature, the established TM model was also applied to predict the *logBMF* values of the samples, respectively. The obtained statistical parameters R_p^2 and $RMSE_p$ are shown in Table 3. It can be seen that the predicted results from the proposed model are significantly better than that of aug-MIA-QSPR$_{color}$ model, which demonstrates that the proposed model possesses stronger predictive ability and reliability.

4. Conclusion

The extraction and selection of features are the most important factors in QSAR/QSPR research. In this study, TM method was used to extract the feature information of molecular structure images and stepwise regression was used to select the effective feature variables and establish the linear quantitative model to predict the log*BMF* of organochlorine pollutants. The results of comprehensive evaluation and comparison indicate that the established model has satisfactory robustness and predictive ability, although it could not provide the explicit physicochemical meaning of the variables in model. This study presented that, as an effective extraction pathway of feature information, TM method is more suitable for the many QSPR/QSAR research based on molecular structure images.

References

Bayraktar B, Bernas T, Robinson JP, Rajwa B (2007) A numerical recipe for accurate image reconstruction from discrete orthogonal moments Pattern Recognition 40:659-669

Birgit M. Braune RJN (1989) Dynamics of organochlorine compounds in herring gulls: III. Tissue distribution and bioaccumulation in lake ontario gulls Environmental Toxicology and Chemistry 8:957-968

Charles J. Henny JLK, Robert A. Grove, V. Raymond Bentley and John E. Elliott (2003) Biomagnification Factors (Fish to Osprey Eggs from Willamette River, Oregon, U.S.A.) for PCDDs, PCDFs, PCBs and OC Pesticides Environmental Monitoring & Assessment 84:275-315

D. Mackay AF (2000) Bioaccumulation of persistent organic chemicals: mechanisms and models Environmental Pollution 110:375-391

da Mota EG, Duarte MH, Barigye SJ, Ramalho TC, Freitas MP (2017) Exploring MIA-QSPR's for the modeling of biomagnification factors of aromatic organochlorine pollutants Ecotoxicology and environmental safety 135:130-136 doi:10.1016/j.ecoenv.2016.09.030

Dearden JC (2016) The History and Development of Quantitative Structure-Activity Relationships (QSARs) International Journal of Quantitative Structure-Property Relationships 1:1-44 doi:10.4018/ijqspr.2016010101

Fatemi MH, Baher E (2009) A novel quantitative structure-activity relationship model for prediction of biomagnification factor of some organochlorine pollutants Molecular diversity 13:343-352 doi:10.1007/s11030-009-9121-4

Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability Domain for QSAR Models International Journal of Quantitative Structure-Property Relationships 1:45-63 doi:10.4018/ijqspr.2016010102

Gramatica P (2020) Principles of QSAR Modeling: Comments and Suggestions From Personal Experience International Journal of Quantitative Structure-Property Relationships 5:61-97 doi:10.4018/IJQSPR.20200701.oa1

Grisoni F, Consonni V, Vighi M (2019) Acceptable-by-design QSARs to predict the dietary

biomagnification of organic chemicals in fish Integrated Environmental Assessment and Management 15:51-63 doi:10.1002/ieam.4106

Hu M-K (1962) Visual pattern recognition by moment invariants IRE Transactions on Information Theory 8:179-187

Jepson PD et al. (2016) PCB pollution continues to impact populations of orcas and other dolphins in European waters Sci Rep 6:18573 doi:10.1038/srep18573

Mansouri K, Consonni V, Durjava MK, Kolar B, Oberg T, Todeschini R (2012) Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling Chemosphere 89:433-444 doi:10.1016/j.chemosphere.2012.05.081

Mitra I, Saha A, Roy K (2010) Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants Molecular Simulation 36:1067-1079 doi:10.1080/08927022.2010.503326

Mukundan R, Ong SH, Lee PA (2001) Image analysis by Tchebichef moments IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 10:1357-1364

Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm2 metrics for validation of QSPR models Chemometrics and Intelligent Laboratory Systems 107:194-205 doi:10.1016/j.chemolab.2011.03.011

P. G (2007) Principles of QSAR models validation: Internal and external QSAR & Combinatorial Science 26:694-701

Paul D. Jepson PMB, Robert Deaville, Colin R. Allchin, John R. Baker, Robin J. Law (2009) Relationships between polychlorinated biphenyls and health status in harbor porpoises (Phocoena phocoena) stranded in the United Kingdom Environmental Toxicology and Chemistry 24:238-248

Ramakrishnan Mukundan S-HO, P.A. Lee (2001) Image analysis by Tchebichef moments IEEE Transactions on Image Processing 10:1357-1364

Rosa Vilanova PFn, Carolina Martı´nez, and Joan O. Grimalt (2001) Organochlorine Pollutants in Remote Mountain Lake Waters Journal of Environmental Quality 30:1286-1295

Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN (2013) Some case studies on application of "r(m)2" metrics for judging quality of quantitative structure-activity relationship predictions:

287 emphasis on scaling of response data J Comput Chem 34:1071-1082 doi:10.1002/jcc.23231
288 Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of
289 QSAR models Chemometrics and Intelligent Laboratory Systems 145:22-29
290 doi:10.1016/j.chemolab.2015.04.013
291 Serrano R, Blanes MA, Lopez FJ (2008) Biomagnification of organochlorine pollutants in farmed
292 and wild gilthead sea bream (Sparus aurata) and stable isotope characterization of the trophic
293 chains Sci Total Environ 389:340-349 doi:10.1016/j.scitotenv.2007.09.020
294 Teague MR (1980) Image analysis via the general theory of moments Journal of the Optical Society
295 of America 70:920-930
296 Todeschini R (2010) Milano Chemometrics University of MilanoBicocca, Milano, Italy (personal
297 communication)
298 Woodburn K, Drottar K, Domoradzki J, Durham J, McNett D, Jezowski R (2013) Determination of
299 the dietary biomagnification of octamethylcyclotetrasiloxane and
300 decamethylcyclopentasiloxane with the rainbow trout (Oncorhynchus mykiss) Chemosphere
301 93:779-788 doi:10.1016/j.chemosphere.2012.10.049
302 Yap PT, Paramesran R, Ong SH (2003) Image analysis by Krawtchouk moments IEEE Trans Image
303 Process 12:1367-1377 doi:10.1109/TIP.2003.818019
304 Zhai HL, Li BQ, Chen J, Wang X, Xu ML, Liu JJ, Lu SH (2018) Chemical image moments and their
305 applications TrAC Trends in Analytical Chemistry 103:119-125
306 doi:10.1016/j.trac.2018.03.017
307
308
309

Table Caption:

Table 1 Experimental and predicted *logBMF* values of all samples

Table 2 Performance of the established models

Table 3 Comparison of the predicted results for the five test sets

Figure Caption:

Figure 1 Linear relationship between *logBMF* calculated values and experimental values

Figure 2 Williams plots. (A) TM model (H^*=0.6) (B) 2D-QSPR model (H^*=0.7)

319 **Table 1** Experimental and predicted *logBMF* values of all samples

No.	Compounds	Abbr.	*logBMF*		
			Exp.	*TM*	*2D-QSPR*
1*	2378TCDF	TCDF	-0.12	-0.10	0.31
2	hexachlorobenzene	HCB	0.32	0.31	0.34
3	3,3',4,4'-Tetrachlorobiphenyl	PCB77	0.77	0.82	0.62
4	2,4,4',5-Tetrachlorobiphenyl	PCB74	0.83	0.89	0.96
5*	2,3,4,4'-Tetrachlorobiphenyl	PCB60	0.90	0.95	0.95
6	2,2',3,4',5',6-Hexachlorobiphenyl	PCB149	0.95	0.97	1.02
7	2,2',3,3',4,5,6'-Heptachlorobiphenyl	PCB174	1.00	1.11	1.12
8	1,2,3,4,6,7,8,9-Octachlorodibenzofuran	OCDF	1.00	0.97	1.06
9*	2,3,3',4',6-Pentachlorobiphenyl	PCB110	1.04	0.96	1.31
10	2,2',4,4',5-Pentachlorobiphenyl	PCB99	1.11	1.15	1.24
11	2,2',4,5,5'-Pentachlorobiphenyl	PCB101	1.25	1.18	1.12
12	2,3,7,8-tetrachlorodibenzo-p-dioxin	TCDD	1.25	1.21	1.32
13	2,3',4,4',5-Pentachlorobiphenyl	PCB118	1.30	1.09	1.25
14	3,3',4,4',5,5'-Hexachlorobiphenyl	PCB169	1.32	1.41	1.48
15*	2,3,3',4,4'-Pentachlorobiphenyl	PCB105	1.36	1.14	1.20
16*	2,2',3,3',4,4',6-Heptachlorobiphenyl	PCB171	1.36	0.85	1.44
17	2,2',3,4,5,5'-Hexachlorobiphenyl	PCB141	1.43	1.30	1.39
18	2,2',3,4,4',5',6-Heptachlorobiphenyl	PCB183	1.43	1.52	1.30
19	2,2',3,3',4,4',5,5'-Octachlorobiphenyl	PCB194	1.43	1.55	1.62
20*	2,2',3,4,4',5,5',6-Octachlorobiphenyl	PCB203	1.43	1.46	1.47
21	3,3',4,4',5-Pentachlorobiphenyl	PCB126	1.43	1.31	1.23
22	2,2',3,4,4',5'-Hexachlorobiphenyl	PCB138	1.46	1.34	1.41
23*	2,2',4,4',5,5'-Hexachlorobiphenyl	PCB153	1.46	1.28	1.25
24	2,2',3,4',5,5'-Hexachlorobiphenyl	PCB146	1.48	1.53	1.41
25	2,2',3,3',4,5',6,6'-Octachlorobiphenyl	PCB201	1.48	1.44	1.45
26	2,2',3,3',4,5,6,6'-Octachlorobiphenyl	PCB200	1.50	1.49	1.41
27	2,2',3,3',4,5,5'-Heptachlorobiphenyl	PCB172	1.53	1.40	1.57
28*	2,2',3,4,4',5,5'-Heptachlorobiphenyl	PCB180	1.53	1.45	1.48
29	1,1-Dichloro-2,2-(4-ClC6H4)ethane	p,p'DDD	1.61	1.81	1.67
30*	Dichlorodiphenyltrichloroethane	DDT	1.92	2.02	2.37
31	1,1-Dichloro-2,2-(4-ClC6H4)ethene	p,p'-DDE	2.19	1.96	2.14
32*	1,2,3,6,7,8-Hexachlorodibenzo-p-Dioxin	H6CDD	2.44	2.15	2.11
33	1,2,3,4,6,7,8-Heptachlorodibenzo-p-Dioxin	H7CDD	2.44	2.42	2.33
34	1,2,3,4,6,7,8,9-Octachlorodibenzo-p-Dioxin	OCDD	2.49	2.58	2.48
35	2,3',4,4'-Tetrachlorobiphenyl	PCB66	0.83	1.04	0.93
36	2,2',3,5',6-Pentachlorobiphenyl	PCB95	0.83	0.88	0.83
37	2,2',3,3',4,4',5-Heptachlorobiphenyl	PCB170	1.53	1.46	1.59
38	2,3,3',4,4',5,6-Heptachlorobiphenyl	PCB190	1.53	1.50	1.68
39	2,2',3,4,4',5,6'-Heptachlorobiphenyl	PCB182	1.39	1.47	1.28

40	2,2',3,4',5,5',6-Heptachlorobiphenyl	PCB187	1.39	1.39	1.25

320 Note: The samples with asterisk (*) belong to test set while others belong to training set.

321

Table 2 Performance of the established models

Data set	Item	TM	2D-QSPR
Training set	LV_c	5	6
	R_c	0.9726	0.9732
	R_{adj}	0.9668	0.9661
	R_{cv}	0.9570	0.9545
	$RMSE_c$	0.1052	0.1040
	$RMSE_{cv}$	0.1320	0.1353
	F-test (*p*-value)	2.08E-14	1.56E-13
	R_r^2	0.5669	0.3065
	${}^cR_p^2$	0.5988	0.7790
	MAE_c	0.08	0.09
Test set	R_p	0.9594	0.9236
	$RMSE_p$	0.2129	0.2541
	k	1.0783	0.9757
	k'	0.9129	0.9952
	r_m^2	0.8567	0.7685
	$r_m'^2$	0.9161	0.6220
	Δr_m^2	0.0594	0.1465
	MAE_p	0.16	0.21

Table 3 Comparison of the predicted results for the different test sets

Test set	R_p^2		$RMSE_p$	
	TM	aug-MIA-QSPR$_{color}$	TM	aug-MIA-QSPR$_{color}$
test **1**	0.9710	0.8451	0.0792	0.2310
test **2**	0.9877	0.7808	0.0589	0.2261
test **3**	0.9844	0.8719	0.0803	0.2048
test **4**	0.9854	0.8759	0.0590	0.1957
test **5**	0.9860	0.8978	0.0690	0.2061
Average±SD	0.9829±0.0068	0.8543±0.0452	0.0693±0.0104	0.2127±0.0151

Figure 1

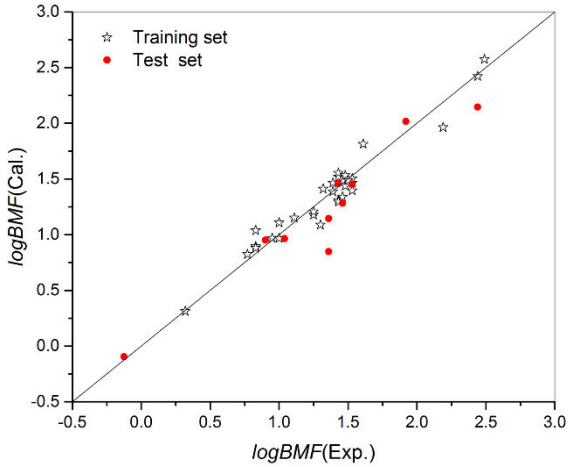
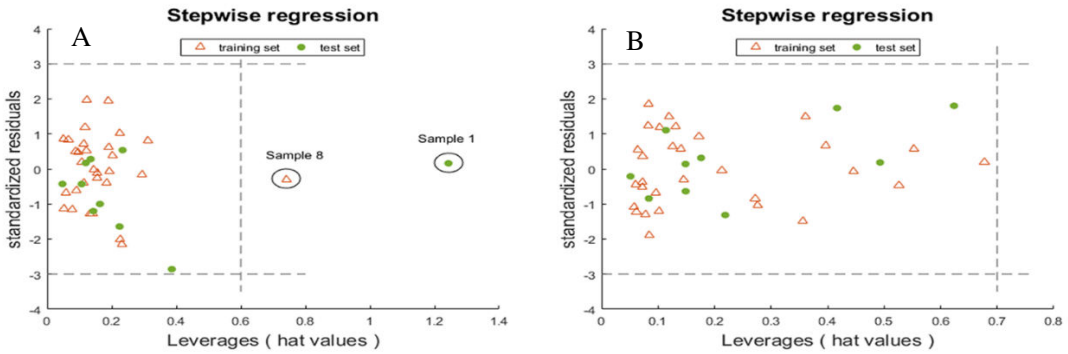


Figure 2



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx