Struct Chem (2014) 25:997–1004
DOI 10.1007/s11224-014-0419-1

ORIGINAL RESEARCH

Direct QSPR: the most efficient way of predicting organic carbon/water partition coefficient (log K_{OC}) for polyhalogenated POPs

Karolina Jagiello · Anita Sosnowska · Sharnek Walker · Maciej Haranczyk · Agnieszka Gajewicz · Toru Kawai · Noriyuki Suzuki · Jerzy Leszczynski · Tomasz Puzyn

Received: 22 January 2014 / Accepted: 28 February 2014 / Published online: 23 March 2014

Abstract The organic carbon/water partition coefficient (K_{OC}) is one of the most important parameters describing partitioning of chemicals in soil/water system and measuring their relative potential mobility in soils. Because of a large number of possible compounds entering the environment, the experimental measurements of the soil sorption coefficient for all of them are virtually impossible. The alternative methods, such as quantitative structure–property relationship (QSPR techniques) have been applied to predict this important physical/chemical parameter. Most available QSPR models have been based on correlations with the *n*-octanol/water partition coefficient (K_{OW}), which enforces the requirement to conduct experiments for obtaining the K_{OW} values. In our study, we have developed a QSPR model that allows predicting logarithmic values of the organic carbon/water partition coefficient (log K_{OC}) for 1,436 chlorinated and brominated congeners of persistent organic pollutants based on the computationally calculated descriptors. Appling such approach not only reduces time, cost, and the amount of waste but also allows obtaining more realistic results.

Keywords Persistent organic pollutants · Organic carbon/water partition coefficient · QSPR · Quantum–mechanical descriptors

Electronic supplementary material The online version of this article (doi:10.1007/s11224-014-0419-1) contains supplementary material, which is available to authorized users.

K. Jagiello · A. Sosnowska · A. Gajewicz · T. Puzyn (✉)
Laboratory of Environmental Chemometrics, Institute for Environmental and Human Health Protection, Faculty of Chemistry, University of Gdansk, ul. Wita Stwosza 63, 80-308 Gdansk, Poland
e-mail: puzi@qsar.eu.org

S. Walker · J. Leszczynski
Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 JR Lynch Street, Jackson, MS 39217-0510, USA

M. Haranczyk
Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720, USA

T. Kawai · N. Suzuki
National Institute for Environmental Studies, Research Center for Environmental Risk, Exposure Assessment Research Section, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan

Introduction

The occurrence of polyhalogenated persistent organic pollutants (POPs), such as chloro- and bromo-substituted biphenyls, naphthalenes, dibenzo-*p*-dioxins, dibenzofurans, and diphenyl ethers has been identified in almost all environmental compartments [1]. Due to their high lipophilicity and resistance to naturally occurring degradation processes, they are prone bioaccumulation in human and animal tissues [2]. In the organism, they are capable to induce various toxic effects, including carcinogenicity, reproductive disorders related to disrupting the hormonal system, immunotoxicity, and damages to the central and peripheral nervous systems. They are also suspected to be responsible for the increasing number of patients nowadays suffering from allergies and hypersensitivity [3, 4]. Therefore, efficient tools for comprehensive environmental risk assessment for polyhalogenated POPs are needed.

The procedure of comprehensive risk assessment requires information about the environmental transport and fate processes of a given substance. Among various

physical/chemical properties governing the environmental occurrence and transport of POPs, the most important are: water solubility, vapor pressure, and partition coefficients, i.e., *n*-octanol/water partition coefficient (K_{OW}), *n*-octanol/air partition coefficient (K_{OA}), air/water partition coefficient (K_{AW}), and organic carbon/water partition coefficient (K_{OC}) [2]. The last property (K_{OC}) is crucial for characterizing the distribution of pollutants between the solid and solution phases in soil, or between water and sediment in aquatic ecosystems [5]. Thus, soil sorption coefficient indicates whenever the chemicals undergo leaching or runoff when enter to the soil or would be immobile [6].

The accurate values characterizing the mentioned properties can be obtained experimentally. However, because of a large number of possible substitution isomers, congeners, may exist, the empirical measurements of the properties for all of them are impractical. Therefore, the only way to acquire complete physicochemical characteristics of all polyhalogenated POPs are to employ computational techniques, such as quantitative structure–property relationships (QSPR) modeling [7].

Numerous QSPR-based methods of calculating K_{OC} have already been reported [6, 8–10]. In most of them the values of organic carbon/water partition coefficient were derived from the *n*-octanol/water partition coefficient [11–13]. Thus, in fact, another experimentally measured property (log K_{OW}) has been employed as the descriptor. Gawlik et al. [14] summarized the published models into a common form (1):

$$\log K_{OC} = a \log K_{OW} + b, \quad (1)$$

where a is the regression coefficient and b is the intercept. Both a and b depend on the compounds used for fitting. The values of a and b range from 0.15 to 6.69 and from −0.78 to 2.25, respectively. However, the necessity of measuring the accurate values of K_{OW} for a large number of hydrophobic compounds in order to obtain the values of K_{OC}, makes the whole procedure less efficient, i.e., more difficult, expensive, and time-consuming.

Since the QSPR technique employing computationally calculated descriptors has been already successfully applied to predict *n*-octanol/water partition coefficient (K_{OW}) [15] the question raised whenever there is the possibility to use such descriptors to predict the organic carbon/water partition coefficient (K_{OC}). Consequently, considering that, one needs to investigate, if there is possibly a much more efficient, direct way of obtaining the values of log K_{OC}, then the scheme summarized by Gawlik et al. [14].

Therefore, our study was aimed at comparing the direct (based on computational descriptors) method of predicting log K_{OC} with the existing QSPR models utilizing the value of log K_{OW}. To perform this task, we have developed a QSPR model that predicts the organic carbon–water partition coefficients for a series of polyhalogenated POPs (polychlorinated and polybrominated benzenes, biphenyls, dibenzo-*p*-dioxins, dibenzofurans, diphenyl ethers, and naphthalenes) based on quantum–mechanical molecular descriptors. The descriptors could be obtained computationally, without performing additional experiments. The comparison resulted in practical recommendations toward the efficient environmental transport and fate modeling of polyhalogenated POPs that utilizes the values of log K_{OC} as model inputs.

Materials and methods

Predicting organic carbon/water partition coefficient (log K_{OC}) with the direct QSPR approach

At the first stage of our study, we have developed a novel QSPR model that allowed predicting the values of organic carbon/water partition coefficient directly from quantum–mechanical descriptors. The algorithm that we applied consisted of five main steps: (i) collecting experimental data and splitting them into training set (T) and validation set (V); (ii) calculating molecular descriptors; (iii) calibrating the model; (iv) internal and external validation of the model and the assessment of applicability domain; and (v) applying the model to predict the values of log K_{OC} for the compounds, for which the experimentally derived values of the coefficient have been unavailable.

The values of K_{OC} for all studied POPs derivatives were taken from the *Handbook of Physical–Chemical Properties and Environmental Fate for Organic Compounds* [16]. The experimental data have been available for 205 chlorinated or brominated POPs congeners (for details please refer to Supplementary Material). The logarithmic values of log K_{OC} ranged from 2.19 to 8.09 [16]. The compounds, for which experimental data have been available, were divided into two sets: training set and validation set. The compounds were ranked according to their endpoints (the experimentally determined values), and every forth compound was labeled as a validation compound and removed from the training set; the first and second compounds were arbitrarily included in the training set. This commonly used method produces two sets that accurately represent the data [17, 18].

In the second step of QSPR modeling, we calculated molecular descriptors (the formal, mathematical representations of a molecule) and selected the best possible combination of the descriptors to be used as independent variables in the model. We employed our algorithms and software tools for combinatorial generation of congeners and their characterization [19, 20]. Quantum–mechanical

Table 1 Symbols and definitions of all calculated molecular descriptors [25]

Symbol	Definitions of molecular descriptors	Units
*n*H	Number of hydrogen substituents	–
*n*Cl	Number of chlorine substituents	–
*n*Br	Number of bromine substituents	–
*n*A	Number of atoms in the molecule	–
MW	Molecular weight	g/mol
HOF	Standard heat of formation	kcal/mol
EE	Electronic energy	eV
Core	Core repulsion energy	eV
TE	Total energy	eV
HOF_c	Standard heat of formation in a solution represented by the conductor-like screening model (COSMO)	eV
TE_c	Total energy in a solution represented by COMSO	eV
HOMO	Energy of the highest occupied molecular orbital (HOMO)	eV
LUMO	Energy of the Lowest Unoccupied Molecular Orbital	eV
Dx	X vector of the dipole moment	Debye
Dy	Y vector of the dipole moment	Debye
Dz	Z vector of the dipole moment	Debye
Dtot	Total dipole moment	Debye
SAS	Solvent accessible surface	$Å^2$
MV	Molecular volume	$Å^3$
Q_-	Lowest negative Mulliken's partial charge on the molecule	–
Q_+	Highest positive partial charge on the molecule	–
Ahof	Polarizability derived from the heat of formation	$Å^3$
Ad	Polarizability derived from the dipole moment	$Å^3$
En	Mulliken's electronegativity	eV
Hard	Parr and Pople's absolute hardness	eV
Shift	Schuurmann MO Shift alpha	eV

descriptors were calculated at the semi-empirical PM6 level [21] in the MOPAC 2007 package [22]. PM6 method may be used in QSPR modeling for POPs, as its suitability for the performed tasks has been proved earlier [23]. We obtained a matrix of 26 molecular descriptors (Table 1) reflecting the structural variability in the studied 1,436 chlorinated and brominated POPs congeners. Then, we selected the optimal combination of the descriptors by applying hierarchical cluster analysis with the correlation ways of calculating distances between the descriptors and Ward's method of linkage [24].

The multiple linear regression (MLR) was applied as a chemometric method of modeling at the third step. We assumed that the modeled property (log K_{OC}) would be expressed as a function of molecular descriptors ($\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$,…):

$$\log K_{OC} = a_1\mathbf{x_1} + a_2\mathbf{x_2} + a_3\mathbf{x_3} + \cdots + a_n\mathbf{x_n} + b, \quad (2)$$

where a_1, a_2, a_3, …, a_n are regression coefficients and b is the intercept. Goodness-of-fit was verified by calculating determination coefficient in the training set (R^2) and the root mean square error of calibration ($RMSE_c$) (Eqs. 3 and 4).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^{n}(y_i^{obs} - \bar{y}^{obs})^2}, \quad (3)$$

$$RMSE_c = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{obs} - y_i^{pred})^2}{n}}, \quad (4)$$

where y_i^{obs} is an i th experimental value of log K_{OC}, y_i^{pred} is an i th predicted value of log K_{OC}, $\bar{y}^{obs}$ is the mean experimental value of log K_{OC} for the compounds from training set, and n indicates the number of compounds in the training set.

At the fourth step, we applied leave-one-out cross-validation method (LOO), as an internal validation technique, to evaluate robustness of the model [26, 27]. For the quantitative assessment of model's robustness, we calculated the cross-validation coefficient (Q^2_{cv}) and the root mean square of cross-validation ($RMSE_{CV}$) (Eqs. 5 and 6).

$$Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_i^{obs} - y_i^{predcv})^2}{\sum_{i=1}^{n}(y_i^{obs} - \bar{y}^{obs})^2}, \quad (5)$$

$$RMSE_{cv} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{predcv}\right)^2}{n}}, \quad (6)$$

where y_i^{obs} is an i th experimental value of log K_{OC}, y_i^{predcv} is the predicted value of log K_{OC} for an i th compound, temporarily excluded according to the leave-one-out algorithm, $\bar{y}^{obs}$ is the mean experimental value of log K_{OC} for the compounds from training set, and n indicates the number of compounds in the training set. Then, we carried out the external validation to confirm good predictive ability of the developed model. We applied the model for performing predictions of log K_{OC} for independent (external) compounds (not previously used in model's calibration). The results of external validation have been expressed in terms of Q^2_{Ext} (the external validation coefficient), and the root mean square of prediction ($RMSE_P$) [28] (Eqs. 7 and 8).

$$Q^2_{Ext} = 1 - \frac{\sum_{j=1}^{k}(y_j^{obs} - y_j^{pred})^2}{\sum_{j=1}^{k}(y_j^{obs} - \bar{y}^{obs})^2}, \quad (7)$$

$$RMSE_p = \sqrt{\frac{\sum_{j=1}^{k}(y_j^{obs} - y_j^{pred})^2}{k}}, \quad (8)$$

where y_j^{obs} is an j th experimental value of log K_{OC}, y_j^{pred} is an j th predicted value of log K_{OC}, $\bar{y}^{obs}$ is the mean

experimental value of log K_{OC} for the compounds from training set, and k indicates the number of compounds in the training set. The next integral part of the validation procedure was to clearly define the domain of applicability. In our model, applicability domain was verified with using the Williams plot [27, 28] and Insubria graph approaches [29].

In the final, fifth step, after sterling validation, the developed QSPR model was applied to predict the values of the organic carbon/water partition coefficient for the compounds, for which the experimentally measured data have been unavailable. Reliability of the predictions (related to the applicability domain) was assessed based on the leverage value and Insubria graph approach [29].

Comparing the direct method of predicting organic carbon/water partition coefficient with other methods

As mentioned in the Introduction, in most published contributions the values of log K_{OC} have been derived from another physicochemical property, i.e., n-octanol/water partition coefficient (log K_{OW}). Thus, we performed a literature search for the best available models for predicting log K_{OC}. In the next step, a comparison of the prediction efficiency between such models and the direct QSPR model developed in this study has been carried out.

In this comparison we have taken into account: (i) time required to obtain log K_{OC}, (ii) cost associated with the conducted investigations, (iii) the amount of waste arising during investigations, and iv) predictive abilities of selected approaches.

Results and discussion

Predicting organic carbon/water partition coefficient (log K_{OC}) with direct QSPR approach

The application of hierarchic cluster analysis on the matrix of quantum mechanical descriptors led to dividing descriptors into three main clusters: cluster A containing: Shift, HOMO, Q_+, Dtot, Dy, Dz, nBr, En, Hard, Dx, Q_-; cluster B containing: nA, MW, Ahof, Ad, SAS, MV, Core, nCl; and cluster C—containing: TE_c, TE, EE, HOF_c, HOF, LUMO, nH (Fig. 1).

In the variant of HCA we have applied, descriptors were grouped according to their pair correlations (descriptors highly correlated each other formed particular clusters). Thus, to avoid redundancy, we have selected one representative descriptor from each cluster. The representative descriptors were selected in a way to minimize their correlation coefficient with descriptors representing other groups. Finally, we have selected three representative descriptors: SAS, LUMO, and Dt. In the next step, we applied MLR methodology and, in effect, obtained a regression model (Eq. 9) with good predictive ability.

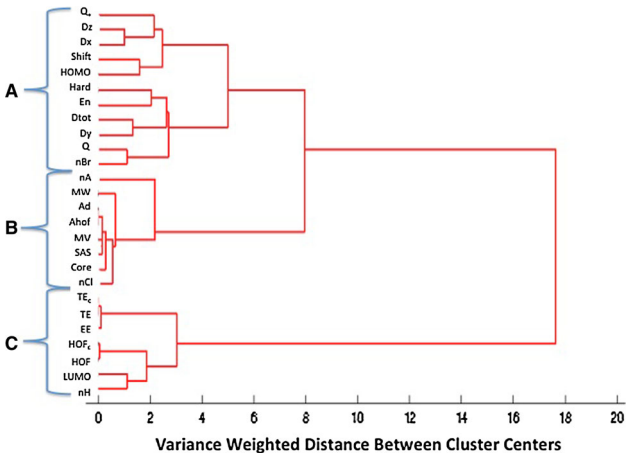

Fig. 1 Hierarchical cluster analysis of descriptors

$$\log K_{OC} = 6.03(\pm 0.01) + 0.93(\pm 0.01)\mathrm{SAS}, \quad (9)$$

$$n = 154, n_{val} = 51, F = 5712, p < 10^{-4}, R^2 = 0.97,$$
$$Q^2_{CV} = 0.97, Q^2_{Ext} = 0.97, \mathrm{RMSE_C} = 0.15,$$
$$\mathrm{RMSE_{CV}} = 0.15, \mathrm{RMSE_P} = 0.15,$$

where SAS is the solvent accessible surface area calculated at semi-empirical PM6 level, n is the number of compounds in training set, n_{val} is the number of compounds in validation set, R^2 is the determination coefficient in the training set, Q^2_{cv} is the cross-validation coefficient, Q^2_{ext} is the external validation coefficient, $\mathrm{RMSE_c}$ is the root mean square error of calibration, $\mathrm{RMSE_{cv}}$ is the root mean square of cross-validation, and $\mathrm{RMSE_p}$ is the root mean square of prediction.

Goodness-of-fit, robustness, and high predictive ability have been confirmed by the values of R^2, Q^2_{CV}, Q^2_{Ext} (close to 1) and relatively low values of the errors: $\mathrm{RMSE_C}$, $\mathrm{RMSE_{CV}}$, and $\mathrm{RMSE_P}$. Moreover, the visual correlation between observed and predicted log K_{OC} values for the training (T) and validation (V) set confirmed the good quality of the model (Fig. 2).

Since the error values ($\mathrm{RMSE_C}$, $\mathrm{RMSE_{CV}}$, and $\mathrm{RMSE_P}$) were identical and there were no significantly large residual values for the validation set displayed in Fig. 2, one can conclude that the model has not been overfitted. This means that the model predicts correctly not only for the training compounds but also for other (external) compounds.

In the next stage of validation, we have applied the leverage approach to verify applicability domain of the model. So-called the Williams plot (Fig. 3) presents the relationship between leverage values (expressing similarity

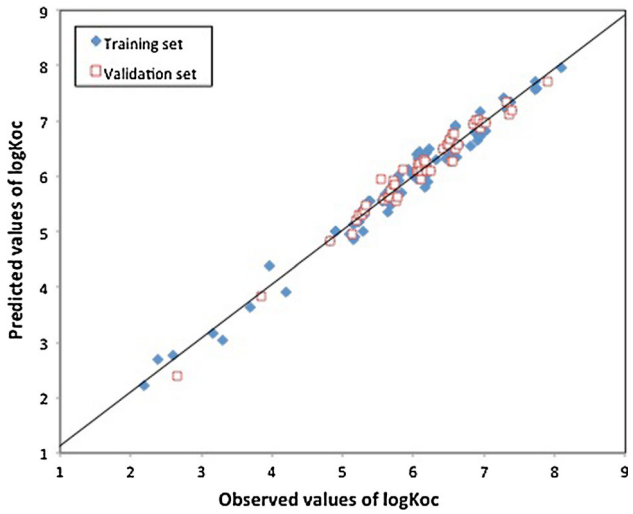

Fig. 2 Calculated versus observed values of log K_{OC}

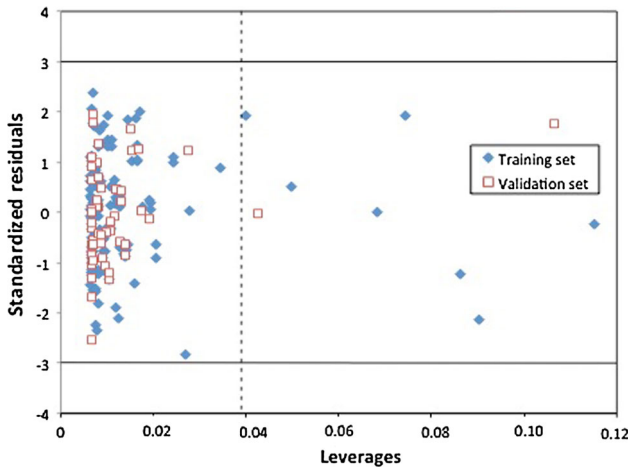

Fig. 3 Williams plot: standardized residuals versus leverages. *Solid lines* indicate ±3 SD units, *dash lines* indicates the threshold value ($h^* = 0.039$)

of a given compound to the training set) and the standardized residuals (prediction errors observed for particular compounds). Analysis of the plot confirmed that because the prediction errors for all compounds from the training and validation sets did not exceed the square area between ±3 SD units, there were not outlying predictions observed. The formal leverage (similarity) threshold value h^* was equal to 0.039. Interestingly, seven compounds from the training set were characterized by the leverages greater than the threshold value, but—simultaneously—they had small residuals. Such compounds are called "good high leverage points," and—as it has been previously demonstrated by Jaworska et al. [30]—compounds from the training set having h_i greater than h^*, stabilize the model and make it predictive for new compounds differing

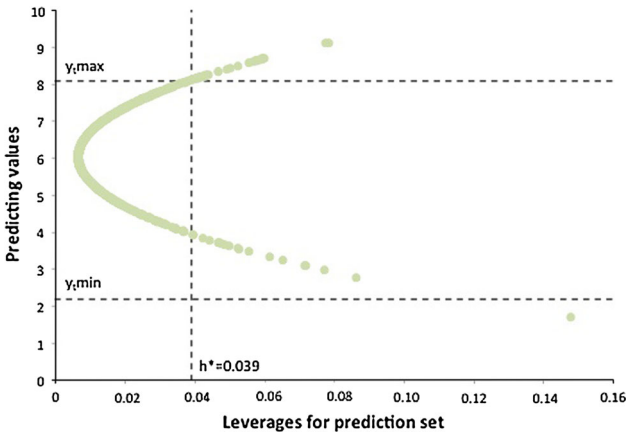

Fig. 4 Insubria graph (plot of the leverage values for the prediction set versus predicted values)

structurally from the training set. Obviously, this is the true only when the residuals observed for the training compounds are small.

Mechanistic interpretation of the developed model, according to the physicochemical theory of dissolution, was intuitive: non-polar chemicals with large solvent accessible surface area (SAS) are less soluble in water. The theory divides the dissolution process into six stages, namely: (i) breaking up solute–solute intermolecular bonds; (ii) breaking up solvent–solvent intermolecular bonds; (iii) formation of a cavity in the solvent phase large enough to accommodate solute molecule; (iv) vaporization of solute into the cavity; (v) forming solute–solvent intermolecular bonds; and (vi) reforming solvent–solvent bonds with solvent restructuring [31]. Thus, since formation of the cavity appropriate for highly halogenated, large molecules require more energy, the solubility of larger congeners is lower, when comparing with less halogenated and smaller congeners, that will simultaneously absorbed mostly by the organic carbon layer. On the other hand, the adsorption of larger molecules on the surface of organic carbon layer is more favored, because of the larger surface of possible intermolecular interactions (attractions) between the target molecules and the organic carbon layer. SAS values increase with the increasing number of halogen atoms present in the molecule and the size of the radius of the halogen substituted. The last feature differentiates chlorinated and brominated derivatives having the same number of halogen substituents, because the atomic radius of bromine atom is larger than the radius of chlorine atom. For example, the values of log K_{OC} of pentachlorobithenyls are higher than that of trichlorobiphenyls, but lower than the values of pentabromobiphenyls. Regarding environmental implications, higher values of the organic carbon/water partition coefficient for highly halogenated organic pollutants correspond with their lower ability to leaching or running off with ground water [32].

Since our QSPR model passed all validation requirements according to OECD recommendations, we have applied the model to predict the unavailable logarithmic values of log K_{OC} for 1,231 polychlorinated and polybrominated congeners. Values of log K_{OC} predicted for particular compounds are listed in the Supplementary Material. In order to verify, whether all chemicals from the prediction set (chemicals, for which experimentally determined values of log K_{OC} have been unavailable) are inside of the model domain, we applied Insubria graph [29]. The graph (Fig. 4) plots the leverages for prediction set versus predicted values. With this plot, we defined the reliable prediction zone of the model based on structural similarity to the training compounds (leverage value) and the predicted value of log K_{OC}. We assumed that the predicted results are reliable, if both conditions: $h_i < h^*$ and $y_t\text{min} < y_i^{\text{pred}} < y_t\text{max}$ have been fulfilled ($y_t$min and $y_t$max are the minimal and the maximal value of log K_{OC} in the training set). We found that about 95 % of compounds from the prediction set were located within the model's applicability domain. Compounds found to be outside the domain were: PBB-194, PBB-196, PBB-198, PBB-203, PPB-205 to PBB-209, PBDD-73 to PBDD-75, PBDE-172 to PBDE-175, PBDE-178, PBDE-180, PBDE-182, PBDE-186, PBDE-189 to PBDE-199, PBDE-201 to PBDE-209, PBDF-135, PCDE-209, and CBz-00. For these chemicals, the predictions are less reliable because the values of log K_{OC} have been extrapolated.

Comparing the direct method of predicting organic carbon/water partition coefficient (log K_{OC}) with other methods

Many other contributions related to the prediction of log K_{OC} has been published so far [5, 6, 9, 11–13]. Methods of the prediction proposed in majority of them can be classified as "indirect" ones, because they are based on the correlation of log K_{OC} with another environmentally relevant parameter—log K_{OW} partition coefficient, which has to be either determined experimentally or calculated first [10–12, 33]. In the following paragraph, we present the results of a simple comparison between the results of the predictions by using our (direct) model and predictions by the other available (indirect) models.

We selected indirect models, originally proposed by Gerstl and Mingelgrin [11] and by Karickhoff [12] to compare them with our (direct) QSPR model.

The comparison has been performed according to the simple scheme (Fig. 5), taking into account three possible strategies of predicting log K_{OC}:

- log K_{OC}^{I} calculated according to newly developed QSPR model (direct method presented in this work),
- log K_{OC}^{II} calculated according to the equations proposed by Gerstl and Mingelgrin [11] (Eq. 10) and by Karickhoff [12] (Eq. 11) with use of the experimentally derived values of *n*-octanol/water partition coefficient (indirect method):

$$\log K_{OC}{}^{IIA} = 0.762 \log K_{OW}{}^{exp} + 1.051, \quad (10)$$

$$\log K_{OC}{}^{IIIA} = 0.762 \log K_{OW}{}^{pred.} + 1.051, \quad (11)$$

- log K_{OC}^{III} calculated according to the equations proposed by Gerstl and Mingelgrin [11] (Eq. 12) and by Karickhoff [12] (Eq. 13) with use of the predicted values of the *n*-octanol/water partition coefficient. The log K_{OW} values were predicted using one of our previously built QSPR models [15] (indirect method)

$$\log K_{OC}{}^{IIIA} = 0.762 \log K_{OW}{}^{pred.} + 1.051, \quad (12)$$

$$\log K_{OC}{}^{IIIB} = 0.989 \log K_{OW}{}^{pred.} - 0.346. \quad (13)$$

Statistical comparison of the results (predicted values of log K_{OC}), obtained with the three methods, has been performed with use of a test set containing 41 compounds, for which we were able to find the experimental values of both partition coefficients: log K_{OC}, and log K_{OW}. Thus, we investigated differences between the experimental and predicted values of log K_{OC} with pairwise *t* Student's test for each of the three strategies.

The values of $p > 0.05$ (Table 2) indicate that the results from each of the compared models differ significantly from the results obtained experimentally. Therefore, all presented calculation schemes might be applied to predict log K_{OC} partition coefficient for POPs. However, based on the lowest mean residual value (Table 2) one can assume that the QSPR model developed in this work (log K_{OW}^{I}) enables obtaining the most reliable results. The worst prediction ability characterized log K_{OW}^{III}—the scheme, in which the value of log K_{OW} was predicted with another QSPR model as a descriptor.

Therefore, more generally, we recommend using direct QSPR models such as the one we have developed in this

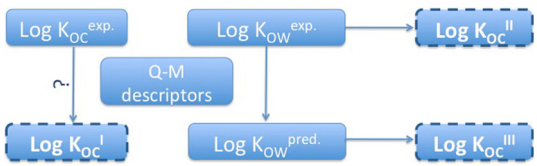

Fig. 5 Three schemes of predicting log K_{OC}: log K_{OC}^{I}—values predicted using newly developed QSPR model (direct method); log K_{OC}^{II}—values predicted using the experimental values of log K_{OW} (indirect methods); log K_{OC}^{III}—values predicted using the predicted values of log K_{OW} (indirect method)

Table 2 Comparison between the residuals derived from different schemes of predicting log K_{OC} with the observed values of log K_{OC} (the pairwise Student's t test)

Statistics	Model				
	K_{OC}^{I}	K_{OC}^{IIA}	K_{OC}^{IIB}	K_{OC}^{IIIA}	K_{OC}^{IIIB}
Mean residual	0.018	0.041	0.089	0.098	0.197
Standard deviation of residuals	0.162	1.353	1.465	1.496	1.501
Test statistic ($t_{kr} = 2.021$)	0.718	0.194	0.388	0.419	0.839
p value	0.477	0.847	0.700	0.677	0.406

contribution. Another advantage is that the application of the model that predicts the log K_{OC} value of chloro- and bromo-analogs of POPs directly from a quantum mechanical descriptor is independent on the availability of other experimental data (i.e., experimentally derived values of log K_{OW}). Since Baker et al. [34–36] observed that the correlation log K_{OC}/log K_{OW} tend to be specific only for chemicals with log $K_{OW} < 6$ searching for alternative ways of predicting of K_{OC} is reasonable and justified. The authors have demonstrated that at least for 18 POP species having log K_{OW} values in the range 6–7, these correlation is very low, measured by $R^2 = 0.294$ [36]. Application of this approach for such chemicals will lead to increased error with prediction of soil sorption coefficient. Thus, using direct model does not only prevent making possible systematic errors and mistakes during the experiments and mathematical conversions but also reduces time, cost associated with experimental research, and the amount of waste arising during such studies. Furthermore, the advantage of using computationally obtained descriptors is that they can be calculated also for not yet synthesized compounds. Thus, partition coefficients can be predicted for novel unknown and untested compounds.

It should be mentioned here that similar direct models have already been developed by other authors. Gramatica et al. [6, 9] reviewed most recently published QSPR models for predicting log K_{OC}. These models differ not only by descriptors used but also by size and composition of the training set (thus, its applicability) and predictive abilities. Moreover, many of them, as the authors note, are verified only in the case of their goodness-of-fit, while their predictive power for compounds not previously used for training is not known [6]. Therefore, applications of such improperly validated models are disputable. Gramatica et al. [9] proposed a series of QSPR models of K_{OC} for a wide and highly heterogeneous data set of 643 nonionic organic chemicals that fulfill all OECD recommendations [7]. The developed models have very good stability, robustness, and predictivity. Moreover, their applicability domains have been clearly described, according to the golden QSPR standards. However, the advantage of QSPR model presented within this study is that it includes only one descriptor. Moreover, the descriptor utilized in our model is very intuitive in mechanistic interpretation.

Conclusions

In our contribution, we have developed a QSPR model for predicting the organic carbon/water partition coefficient for 1,436 polychlorinated and polybrominated congeners of benzens, biphenyls, dibenzo-*p*-diozins, dibenzofurans, diphenylethetrs, and naphtalenes. The model is based on a single molecular descriptor (solvent accessible surface—SAS) that can be simply calculated exclusively from the characteristic of chemical structure. We have observed that the values of log K_{OC} increase with the increasing SAS that is related to the increasing number of halogen substituents. In addition, since brominated congeners are characterized by higher surface comparing with their chlorinated analogs, their log K_{OC} partition coefficients are also higher. This significantly differentiates mobility of chlorinated and brominated POPs in the environment.

The QSPR model fulfills all five OECD recommendations related to the validation procedure: it has satisfactory statistics of goodness-of-fit, robustness, and predictive ability. Applicability domain of the model covers majority of the studied chemicals.

Finally, we have compared the predictions of our direct QSPR model with the values of log K_{OC} predicted using other models based on the *n*-octanol/water partition coefficient. We have demonstrated that the estimation of log K_{OC} of chloro- and bromo-analogs of POPs with the direct QSPR leads to more reliable results than in case of application and other available methods. In addition, the application of our model is possible whenever the values of the other coefficient (log K_{OW}) are even do not known, without necessity of performing additional time-consuming and expensive experiments.

Acknowledgments This work was supported by Japan Society for the Promotion of Science (JSPS) and the Polish Academy of Science (PAN) under the Bilateral Joint Research Project, and by JSPS Grants-in-Aid for Young Scientists (B) No. 25871087. The authors (K. J., A. S., A. G. and T. P.) thank to the Polish Ministry of Science and Higher Education (grant no. DS 530-8180-D202-3) and the Foundation for Polish Science (FOCUS 2010 Programme) for the financial support. This research was supported in part (to M. H.) by the U. S. Department of Energy under contract DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DEAC02-05CH11231.

References

1. Yang G, Zhang X, Wang Z, Liu H, Ju X (2006) Estimation of theaqueous solubility (lgSw) of all polychlorinated dibenzo-furans (PCDF) and polychlorinated dibenzo-*p*-dioxins (PCDD) congeners by density functional theory. J Mol Struc: THEOCHEM 766:25–33
2. UNEP (2001) Stockholm convention on persistence organic pollutants. United Nations Environment Programme, Geneve, Switzerland
3. Blankenship AL, Kannan K, Villalobos SA, Villeneuve DL, Falandysz J, Imagawa T, Jakbsson E, Giesy JP (2000) Relative potencies of indyvidual polychlorinated naphthalenes and Halowax mixtures to induce Ah receptor-madiated responses. Environ Sci Technol 34:3153–3158
4. Villeneuve DL, Kannan K, Khim JS, Falandysz J, Nikiforov VA, Blankenship AL, Giesy JP (2000) Relative potencies of indyvidual polychlorinated naphthalenes to induce dioxin-like responses in fish and mammalian in vitro bioassays. Arch Environ Contam Toxicol 39:273–281
5. Kahn I, Fara D, Karelson M, Maran U (2005) QSPR treatment of the soil sorption coefficients for organic pollutans. J Chem Inf Model 45:94–105
6. Gramatica P, Corradi M, Cossonni V (2000) Modelling an prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. Chemosphere 41:763–777
7. OECD (2004) OECD Principles for the validation, for regulatory purposes, of (Quantittative) Structure Activity Relationship models, 37thJoint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology. Paris, France, Organisation for Economic Co-Operation and Development
8. Doucette WJ (2003) Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. Environ Toxicol Chem 22:1771–1788
9. Gramatica P, Giani E, Papa E (2007) Statistical external validation and consensus modeling: a QSPR case study for K_{OC} prediction. J Mol Graphics Modell 25:755–766
10. Sabljic A, Gusten H, Verhaar H (1995) QSAR modeling of soil sorption—improvements and systematics of log K_{OC} vs log K_{OW} correlations. Chemosphere 31:4489–4514
11. Gerstl Z, Mingelgrin U (1984) Sorption of organic substances by soils and sediments. J Environ Sci Health 19:297–312
12. Karickhoff SW (1981) Semi-empirical estimation of sorption of hydrophobic pollutants on natural sediments and soils. Chemosphere 8:833–846
13. Szabo G, Prosser SL, Bulman A (1990) Determination of the adsorption coefficient (K∼) of some aromatics for soil by RP-HPLC on two immobilized humic acid phases. Chemosphere 21:777–778
14. Gawlik BM, Sotiriou N, Feicht EA, Schulte-Hostede S, Kettrup A (1997) Alternatives for the determination of the soil adsorption coefficient, K_{OC} of non-ionicorganic compounds—a review. Chemosphere 34:2525–2551
15. Puzyn T, Suzuki N, Haranczyk M (2008) How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine? Environ Sci Technol 42:5189–5195
16. Mackay D, Shiu WY, Ma K-C, Lee SC (2007) Physical-chemical properties and environmental fate for organic chemicals. Taylor & Francis, Boca Raton
17. Hewitt M, Cronin MT, Madden JC, Rowe PH, Johnson C, Obi A, Enoch SJ (2007) Consensus QSAR models: do the benefits outweigh the complexity? J Chem Inf Model 47:1460–1468
18. Puzyn T, Mostrąg-Szlichtyng A, Gajewicz A, Skrzyński M, Worth PA (2011) Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. Struct Chem 22:795–804
19. Haranczyk M, Puzyn T, Sadowski P (2008) ConGENER—a tool for modeling of the congeneric sets of environmental pollutants. QSAR Comb Sci 27:826–833
20. Haranczyk M, Urbaszek P, Ng EG, Puzyn T (2012) Combinatorial × computational × cheminformatics approach to characterization of congeneric libraries of organic pollutants. J Chem Inf Model 52:2902–2909
21. Steward JJP (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. J Mol Modell 13:1173–1213
22. Stewart JJP (2009) MOPAC2009. Stewart computational chemistry Available from: http://openmopac.net/MOPAC2009.html. Accessed 14 April 2009
23. Puzyn T, Suzuki N, Haranczyk M, Rak J (2008) Calculation of quantum-mechanical descriptors for QSPR at the DFT level: is it necessary? J Chem Inf Model 48:1174–1180
24. Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244
25. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH Verlag, Weinheim
26. OECD (2007) Guidance Document on the Validation of (Quantitative) StructureeActivity Relationships [QSAR] Models. Organisation for Economic Co-operation and Development, Paris, France
27. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77
28. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701
29. Gramatica P, Cassani S, Roy PP, Kovarich S, Wei YC, Papa E (2012) QSAR modeling is not "push a button and find a correlation": a case study of acute toxicity of (benzo-)triazoles on algae. Mol Inform 31:817–835
30. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. ALTA 33:445–459
31. Puzyn T, Gajewicz A, Rybacka A, Haranczyk M (2011) Global versus local QSPR models for persistent organic pollutants: balancing between predictivity and economy. Struct Chem 22:873–884
32. Cleveland CB (1996) Mobility assesament of agrichemicals: current laboratory methodology and suggestion for future directions. Weed Technol 10:157–168
33. Seth R, Mackay D, Munckle J (1999) Estimating the organic carbon partition coefficient and its variability for hydrophobic chemicals. Environ Sci Technol 33:2390–2394
34. Baker JR, Mihelcic JR, Luehrs DC, Hickey JP (1997) Evaluation of estimation methods for organic carbon normalized sorption coefficient. Water Environ Res 69:136–145
35. Baker JR, Mihelcic JR, Shea E (2000) Estimating K_{OC} for persistent organic pollutants: limitation of correlations with K_{OW}. Chemosphere 41:813–817
36. Baker JR, Mihelcic JR, Sabljic A (2001) Reliable QSAR for estimating KOC for persistent organic pollutants: correlation with molecular connectivity indices. Chemosphere 45:213–221