

Global versus local QSPR models for persistent organic pollutants: balancing between predictivity and economy

Tomasz Puzyn · Agnieszka Gajewicz ·
Aleksandra Rybacka · Maciej Haranczyk

Received: 7 January 2011 / Accepted: 12 February 2011 / Published online: 9 March 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Experimentally determined data on the key physicochemical parameters for halogenated congeners of persistent organic pollutants (POPs) are available only for a limited number of compounds. In the absence of experimental data, a range of computational methods can be applied to characterize those species for which experimental data is not available. One of the techniques widely used in this context is quantitative structure–property relationships (QSPR) approach. There are two ways to develop the QSPR models: using a more complex global model or fitting a simple local model that covers a specific class of chemically related compounds. The essence of the study was to investigate, if local models have significantly better explanatory and predictive ability than global models with wider applicability domains. Based on the obtained results, we concluded that whenever global models fulfill all quality recommendations by OECD, they would be applied in practice as more efficient ones in state of more time consuming procedure of modeling the particular groups of POPs one-by-one. On the contrary, local models are applicable to solve specific problems (i.e., related to only one group of POPs), when high-quality experimental

data are available for a sufficient number of training and validation compounds.

Keywords Global models · Local models · QSPR · Persistent organic pollutants

Introduction

The occurrence of polyhalogenated persistent organic pollutants (POPs), such as Cl/Br-substituted benzenes (CBz/BBz), biphenyls (PCBs/PBBs), diphenyl ethers (PCDEs/PBDEs), dibenzofurans (PCDFs/PBDFs), dibenzo-*p*-dioxins (PCDDs/PBDDs), and naphthalenes (PCNs/PBNs) in air, water, soil, and sediments has been identified as a serious environmental threat [1]. Large amounts of POPs come from various anthropogenic sources, including intentionally synthesized liquids utilized in transformers and capacitors, plasticisers, flame retardants, as well as thermal recycling of waste, domestic heating, etc. Substantial volumes of these compounds are released in effect of giant fires, as the most recent fire of the oil spill at the Deepwater Horizon platform in the Gulf of Mexico [2]. Regardless of their source, the exposure to POPs can cause a vast range of acute and chronic health effects, including mutagenic, carcinogenic, and metabolic ones. In addition, as persistent and lipophilic substances, POPs can be bioaccumulated in body and biomagnified in natural ecosystems [3].

Hence, there is an urgent need to determine physicochemical properties required to perform a comprehensive risk assessment for all POPs. Unfortunately, the number of all possible congeners (similar compounds based on the same carbon skeleton, but differ by a number of chlorine/bromine atoms and the substitution pattern) is extensive. In total, there are 1436 structurally different congeners of

Electronic supplementary material The online version of this article (doi:10.1007/s11224-011-9764-5) contains supplementary material, which is available to authorized users.

T. Puzyn (✉) · A. Gajewicz · A. Rybacka
Laboratory of Environmental Chemometrics, Faculty of
Chemistry, University of Gdańsk, Sobieskiego 18, 80-952
Gdańsk, Poland
e-mail: t.puzyn@qsar.eu.org

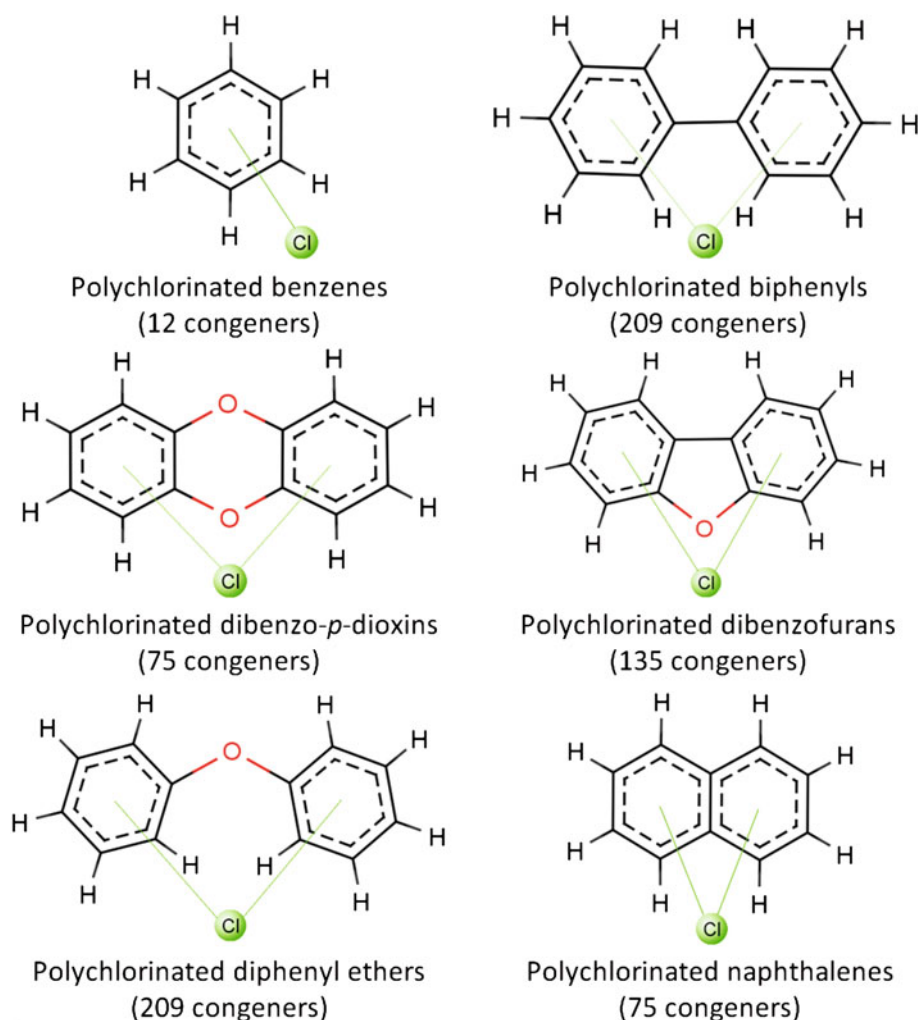
M. Haranczyk
Computational Research Division, Lawrence Berkeley National
Laboratory, One Cyclotron Road, Mail Stop 50F-1650,
Berkeley, CA 94720, USA

polychlorinated and polybrominated benzenes, biphenyls, dibenzo-*p*-dioxins, dibenzofurans, diphenyl ethers, and naphthalenes (Fig. 1). The number of possible mixed chloro- and bromo-substituted is at least one order of magnitude larger [4]. For such a large number of compounds, empirical measurement of the physicochemical properties is impossible, due to high costs and time limitations of the analytical procedures. Therefore, alternative methods for physicochemical characterization of POPs are required.

A very promising group of such methods is the quantitative structure–property relationships (QSPR) approach. QSPR is based on the assumption that each physicochemical property in a group of compounds can be expressed as a mathematical function of their chemical structure, represented by a set of so-called molecular descriptors. Thus, based on the experimental data, available only for some representatives of the group, it is possible to interpolate the lacking data for compounds, for which such data are missing, from the calculated molecular descriptors

and a suitable mathematical model [5–7]. Two possible QSPR modeling strategies have been described in the literature, namely: local and global models. Local models are restricted only to one specific class of chemically related compounds (e.g., PCBs), whereas global models are developed for a large number of structurally similar groups of compounds (e.g., PCBs, PCNs, PCDDs, PCDFs, etc.). It is widely accepted that the local models have better predictive ability in comparison with the global models [8]. However, the global models seem to be very attractive from an economic point of view, because such a modeling strategy enables to additionally save resources by predicting new data for a larger number of compounds at a time. The argument against the global modeling is that this strategy may lead to mechanistic oversimplifications and/or higher errors in the predicted data [9]. Therefore, there are two fundamental questions related to the topic. First: How significant are the differences in the results obtained using local and global QSPRs? Second, consequently: Is the reduction of the model's domain (to only one group of

Fig. 1 Chemical structures of parent molecules of benzenes, biphenyls, dibenzo-*p*-dioxins, dibenzofurans, diphenyl ethers, and naphthalenes used to construct chlorine-substituted congeners



POPs) really necessary to improve the predictive power of a QSPR model? Our study was aimed to answer both questions.

Materials and methods

Global and local QSPRs

To find the answers, we initially selected one phys/chem property and one congeneric group of POPs, namely: water solubility in 25 °C and polychlorinated naphthalenes. Then, we performed a detailed comparison between the predictions with local and global QSPRs for this group. The solubility has been selected, because it is a property, important in estimating both environmental transport and toxicokinetics after entering the body [3]. The group of PCNs (containing 75 congeners) has been selected for the case study, since the parent molecule (naphthalene) is structurally the simplest polycyclic aromatic hydrocarbon. Moreover, polychlorinated naphthalenes were, historically, the first ever intentionally synthesized POPs (between 1910s and 1980s) [10]. The global model has been developed together for 11 other groups of halogenated POPs, namely: CBzs, BBzs, PCBs, PBBs, PCDDs, PBDDs, PCDFs, PBDFs, PCDEs, PBDEs, PBNs, and PCNs (1,436 compounds in total).

We hypothesized that water solubility obtained from a local QSPR model should not substantially differ from those predicted with a global QSPR model for POPs, due to the similarity of carbon skeletons, the level of halogenation and the substitution patterns of the studied compounds. To verify, whether the hypothesis and conclusions can be extended to the other phys/chem properties and groups of POPs, we additionally performed a cross comparison between few local and global QSPRs, collected from the literature.

Development of the global QSPR model

Development of a high-quality QSPR model with good predictive ability requires reliable experimental data, on one hand, and appropriate molecular descriptors on the other one. The procedure we followed when constructing the global model included five steps:

Step 1: Experimental data collection and splitting the compounds, for which the data are available, into a training set (T) and a validation set (V)

The crucial condition that must be met to obtain a plausible QSPR model is homogeneity and high-quality of the experimental data. It is because the quality of the data

significantly influences the modeling results. Thus, no one can expect from the data predicted with the model to be better than the original data utilized to developing the model. In practice, this means that the experimental data should be obtained in a systematic way, according to the same standardized protocol [11]. This stage minimizes the risk of obtaining highly uncertain, extrapolated results from the QSPR modeling.

For the purpose of developing a global QSPR model, which quantitatively describes the relationship between the molecular structure of the halogenated POPs and water solubility ($\log S$), we collected the experimental data on water solubility originally determined at 25 °C. The values of solubility for polychlorinated biphenyls (PCBs) were taken from [12, 13], for polychlorinated dibenzo-*p*-dioxins (PCDDs) from [14], for polychlorinated dibenzofuran (PCDFs) from [14], for polychlorinated/polybrominated diphenyl ethers (PCDEs/PBDEs) from [15, 16], for polychlorinated naphthalenes (PCNs) from [17], and for polychlorinated benzenes (CBz) from [18]. The experimental data have been available for 121 halogenated congeners of POPs in total. Logarithmic values of the solubility varied between -2.58 and -10.83 [mol/dm³] (for more details, please refer to the electronic Supplementary material).

Next, the 121 congeners were sorted along with the decreasing values of water solubility. Then, every fourth compound was moved to the so-called validation set (an additional set for further external validation of the model), while the remaining compounds formed the training set (for developing the model). The application of this “three-to-one” splitting algorithm ensured that the both training and validation sets were contain the compounds evenly distributed within the range of the water solubility [19]. The splitting procedure led to a training and a validation set consisted of 91 (75%) and 30 (25%) compounds, respectively.

Step 2: Calculating molecular descriptors

Simultaneously, we combinatorially generated molecular structures of all chloro- and bromo-substituted congeners (1436 compounds) with the ConGENER [20] software package, which is based on our earlier work on characterization of combinatorially generated libraries of tautomers [21]. We utilized those structures as inputs for quantum-mechanical calculations which included two stages: (i) optimization of the molecular geometry with respect to the energy gradient and (ii) calculation of the descriptors based on the optimized geometry. The calculations have been performed at the semi-empirical level of the theory with use of PM6 method [22] in MOPAC 2009 software package [23]. We calculated the following 26 molecular descriptors: the number of atoms in the molecule (nAT), the number of

chlorine substituents (nX), the molecular weight (MW), the standard heat of formation (HOF), the electronic energy (EE), the core–core repulsion energy (Core), the total energy (TE), the total energy of the corresponding cation (TE_+), the standard heat of formation in a solution represented by the Conductor-like Screening Model, COSMO (HOF_c), the total energy in a solution represented by COMSO (TE_c), the vertical ionization potential (IP), the energy of the highest occupied molecular orbital (HOMO), the energy of the lowest unoccupied molecular orbital (LUMO), the \mathbf{X} vector of the dipole moment (D_x), the \mathbf{Y} vector of the dipole moment (D_y), the \mathbf{Z} vector of the dipole moment (D_z), the total dipole moment (D_{tot}), the solvent accessible surface (SAS), the molecular volume (MV), the lowest negative Mulliken's partial charge on the molecule (Q_-), the highest positive partial charge on the molecule (Q_+), the average polarizability derived from the heat of formation (A_{hof}), the average polarizability derived from the dipole moment (A_d), Mulliken's electronegativity (EN), Parr and Pople's absolute hardness (Hard), and Schuurmann MO shift alpha (Shift).

Step III: Calibrating and internal validation of the QSPR model

Having both, high-quality experimental data and molecular descriptors, we developed QSPR model following the golden standards and recommendations of the Organization for Economic Cooperation and Development (OECD) [24]. Regarding to the five OECD recommendations, an ideal QSPR model should be associated with:

- (i) a defined endpoint;
- (ii) an unambiguous algorithm;
- (iii) a defined applicability domain;
- (iv) appropriate measures of goodness-of-fit, robustness and predictivity;
- (v) a mechanistic interpretation, if possible.

We employed the Partial Least Squares regression combined with a genetic algorithm (GA-PLS) as the chemometric method of modeling. PLS is based on a linear transition from a large number of original descriptors to a small number of new orthogonal variables so-called "latent vectors" (LVs), being linear combinations of the original descriptors [25]. In order to select the optimal combination of the molecular descriptors to be utilized in the final QSPR model, we employed the Holland's genetic algorithm (GA) [26]. The algorithm minimizes the prediction error by searching for the most optimal combination of the descriptors. The name "genetic" came from fact that this mathematical procedure uses the rules of Darwinian theory of evolution. However, in this case, the rules are applied to "populations" and "generations" of mathematical solutions (i.e., combinations of the descriptors), not to

populations and generations of living organisms. The algorithm is controlled by a set of steering parameters. In our studies, we have specified the following ones: the size of a population: 124, the percentage of the initial terms: 40%, the maximum number of generations: 100, the percentage of convergence: 50%, the mutation rate: 0.005, double cross-over: the number of repetitions: 7. GA-PLS calculations were performed with MATLAB 7.6 [27] and PLS Toolbox 5.2 [28].

An integral part of QSPR modeling is to appropriately describe the borders of the optimum prediction space of the model. The space, so-called applicability domain (AD), is defined by the nature of the compounds included in the training set. We verified the applicability domain by use of the Williams plot, which is the plot of the leverage values versus cross-validated standardized residuals [29, 30]. The leverage value h_i for every i th compound is calculated as follows: [31] (Eq. 1):

$$h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \quad (1)$$

where \mathbf{x}_i is the vector of descriptors calculated for the considered i th compound and \mathbf{X} is the matrix of descriptors calculated for the whole training set.

The value of h_i greater than the critical one (h^*) means that the structure of a compound differs from the training set significantly and, in consequence, the compound falls outside the optimum prediction space of the model [32]. The warning value h^* is calculated according to the formula (Eq. 2):

$$h^* = \frac{3(p+1)}{n} \quad (2)$$

where p is the number of variables used in the model and n is the number of training compounds.

However, fact that $h_i > h^*$ does not always indicate that the i th training compound is an outlier. It has been shown that training compounds with high leverages and small residuals (differences between the observed and predicted values) stabilize the model and make it more precise. Such points are so-called "good leverages." Only the compounds with high leverages and residuals higher than ± 3 standard deviations units (so-called "bad leverages") destabilize the model [33].

In order to prove robustness of the model and reduce probability of the model's overfitting, we performed an internal validation [29, 34]. For this purpose, we employed the leave-one-out cross-validation (CV-LOO) algorithm, in which the same compounds were used alternating for the training and validation [30].

Goodness-of-fit (i.e., how well the model fits the data) was measured by the determination coefficient in the training set (R^2) and the root mean square error of calibration (RMSE_c)

(Eqs. 3 and 4). Whereas the quantitative assessment of the robustness was expressed by the CV-LOO determination coefficient (Q_{CV}^2), the absolute average relative deviation (AARD), and root mean square error of cross-validation ($RMSE_{CV}$) (Eqs. 3–7) [30].

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

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

$$Q_{CV}^2 = 1 - \frac{\sum_{i=1}^n (y_i^{\text{obs}} - y_i^{\text{predcv}})^2}{\sum_{i=1}^n (y_i^{\text{obs}} - \bar{y}^{\text{obs}})^2} \quad (5)$$

$$AARD = \frac{100}{n} \sum_{i=1}^n \frac{|y_i^{\text{obs}} - y_i^{\text{pred}}|}{y_i^{\text{obs}}} \quad (6)$$

$$RMSE_{CV} = \sqrt{\frac{\sum_{n=1}^n (y_i^{\text{obs}} - y_i^{\text{predcv}})^2}{n}} \quad (7)$$

where y_i^{obs} is the experimental (observed) value of the property for the i th compound, y_i^{pred} the predicted value for the i th compound, y_i^{predcv} the predicted value for the temporary excluded (cross-validated) i th compound, \bar{y}^{obs} the mean experimental value of the property in the training set, n the number of compounds in the training set.

Step IV: External validation of the developed QSPR model

To confirm the model's predictive power, we carried out the external validation based on the compounds that were not previously engaged in the model's optimization and/or calibration [30]. We utilized the external validation coefficient (Q_{Ext}^2) and the root mean square error of prediction ($RMSE_P$) (Eqs. 8 and 9) as measures of the external predictivity.

$$Q_{\text{Ext}}^2 = 1 - \frac{\sum_{j=1}^k (y_j^{\text{obs}} - y_j^{\text{pred}})^2}{\sum_{j=1}^k (y_j^{\text{obs}} - \bar{y}^{\text{obs}})^2} \quad (8)$$

$$RMSE_P = \sqrt{\frac{\sum_{j=1}^k (y_j^{\text{obs}} - y_j^{\text{pred}})^2}{k}} \quad (9)$$

where y_j^{obs} is the experimental (observed) value of the property for the j th compound, y_j^{pred} the predicted value for j th compound, \bar{y}^{obs} the mean experimental value of the property in the validation set, and k the number of compounds in the validation set.

Step V: Applying the model to predict the endpoint values for new compounds

When the QSPR model fulfills all the validation criteria, it can be applied to predict the property (i.e., water solubility) of those new compounds, for which the experimental data have not been available.

Methodology of comparing local and global QSPR models

Particular local and global models were compared each other taking into account two aspects: economy and quality of each. The number of training compounds and applicability domain of the model represented the economic aspect, whereas the measures of goodness-of-fit, robustness, and predictivity—the qualitative aspect. In addition, we employed Student's t test to verify, whether the average residuals from the predictions with local and global QSPR models differ significantly ($p < 0.05$).

Results and discussion

Comparing global and local QSPR models of water solubility

As mentioned, at first we performed a comparison between two QSPR models of water solubility ($\log S$) developed by our group. The first model was developed within this study, whereas the second QSPR was taken from one of our previous contributions.

Global QSPR model of water solubility

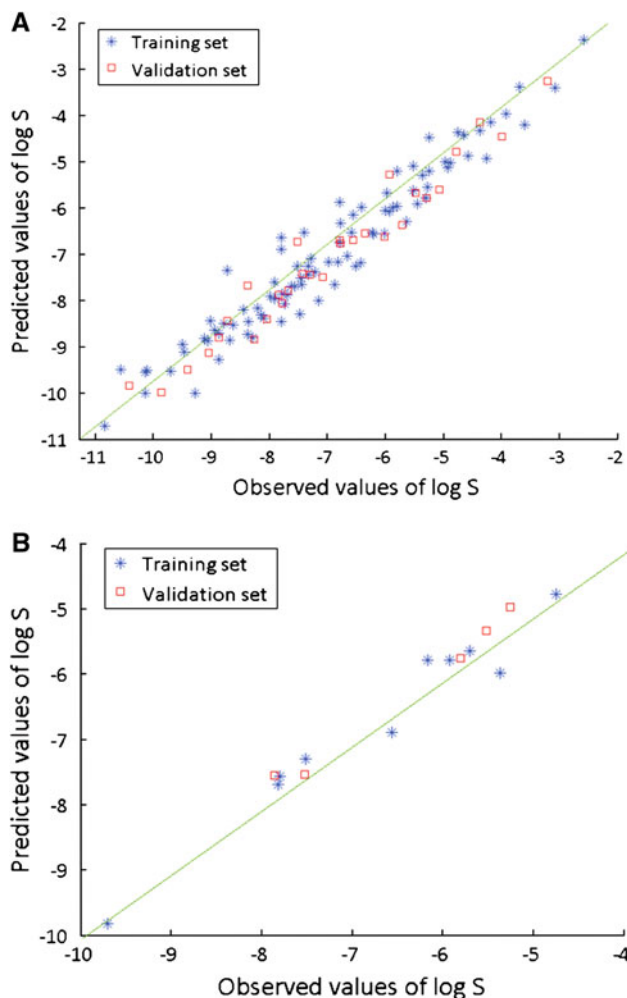
When applied the five-step procedure of QSPR, including GA-PLS method, we obtained a statistically significant ($p < 0.05$) global model, capable to successfully predict the values of $\log S$ for 1436 halogenated POPs. The model utilized three latent vectors (LVs) explaining together 95% (57% + 17% + 21%) of the total variance in the molecular descriptors and 93% (90% + 2% + 1%) of the variance in the modeled endpoint ($\log S$). Although the GA-PLS method uses orthogonal latent vectors for regression, it is also possible to derive “quasi-regression” coefficients for original descriptors (Eq. 10), keeping in mind that these coefficients cannot be individually interpreted, because they are not independent [25].

$$\log S = -0.287nAT - 0.293nX + 0.191LUMO - 0.320SAS + 0.085Q_+ + 0.126Shift \quad (10)$$

The global QSPR was characterized by the satisfactory goodness-of-fit, the robustness, and the external predictive

Table 1 Comparison of statistical parameters between local and global GA-PLS models of log *S*

Feature	Measure	Local QSPR model	Global QSPR model
Goodness-of-fit	R^2	0.96	0.93
	RMSE _C	0.28	0.46
Robustness	Q_{CV}^2	0.96	0.93
	RMSE _{CV}	0.35	0.48
Predictivity	Q_{Ext}^2	0.96	0.95
	RMSE _P	0.20	0.39

**Fig. 2** The experimentally determined values of log *S* versus the values of log *S* predicted by global (a) and local (b) QSPR models

performance (the statistical measures are summarized in Table 1). A visual correlation between the experimental and predicted values of log *S* is presented in Fig. 2a.

The model can be intuitively interpreted, according to the physicochemical theory of dissolution. The theory divides the whole 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. Thus, since formation of the cavity appropriate for highly halogenated, large molecules requires more energy, the solubility of larger congeners is lower, when comparing with less halogenated and smaller congeners. This factor is represented in the model equation (Eq. 10) by three descriptors: SAS, *nAT*, and *nX* that have a negative contribution to the solubility (i.e., the solubility increases when the solvent accessible surface, the number of atoms, and the number of halogen substituents decreases). Similarly, the descriptors that are related to electrostatic interactions (e.g., forming hydrogen bonds) between the solvent and solute and chemical reactivity, namely: LUMO, Q_+ , Shift, positively contribute the solubility. It is because the process of forming solute–solvent intermolecular bonds facilitates dissolution.

Local QSPR model of water solubility

The local model, originally calibrated only for a group of 75 polychlorinated naphthalenes, has been adapted from our previous paper [35]. It was based on eight theoretical molecular descriptors, calculated exclusively from the chemical structures at the Density Functional Theory (DFT) level with the B3LYP functional and 6-311++G(*d, p*) basis set. A combination of those eight descriptors formed one latent vector, utilized then as an independent variable to construct a one-variable GA-PLS model. The model explained 93% of the structural variance (variance in the descriptors) and 96% of the variance in log *S*. This one-variable model can be alternatively expressed in the quasi-regression form (Eq. 11):

$$\begin{aligned} \log S = & -0.109nClp1 + 0.123HOMO - 0.131Hard \\ & + 0.129E_t - 0.131SAS_w - 0.132SAV_w \\ & + 0.131DE_w - 0.129TNE_w \end{aligned} \quad (11)$$

where *nClp1* is the number of chlorine atoms in the first aromatic ring, HOMO the energy of the highest occupied molecular orbital, Hard the molecular hardness, E_t the total energy of the molecule, SAS_w the solvent accessible molecular surface area in the water, SAV_w the solvent accessible molecular volume in the water, DE_w the dispersion energy in the water, and TNE_w the total non-electrostatic energy of solvation.

High values of R^2 , Q_{CV}^2 , and Q_{Ext}^2 , as well as low values of the squared errors: RMSE_C, RMSE_{CV}, and RMSE_P (Table 1) confirmed that the model was well-fitted, robust, and demonstrated its good predictive ability. The existence of a strong linear correlation between the observed and

predicted values of $\log S$ has been graphically proved (Fig. 2b). Details on the local QSPR's development can be found in the original paper [35]. It should be mentioned, however, that the interpretation of the used descriptors is very similar to those for the global model. According to our previous contribution [35], the descriptors refer to the cavitation process (SAS_w and E_t) as well as to the dispersive (DE_w and TNE_w) and electrostatic ($nClp1$, HOMO, and Hard) interactions.

Results of the comparison

Whenever someone wants to compare two QSPR models, one usually starts from evaluating their statistical characteristics. Without doubts, the measures of goodness-of-fit, robustness, and predictivity (Table 1) favor the local QSPR. Higher correlation coefficients (R^2 , Q_{CV}^2 , and Q_{Ext}^2) and up to two times lower values of the root mean square errors for both the training and the validation sets in comparison to the global model proved that local model was more accurate and had better performance of exploring relationships between the structure and water solubility of POPs.

This conclusion is also supported by analysis of two plots (Fig. 3) presenting residuals calculated for chloronaphthalenes based on the predictions with global and local QSPRs. Note, the residuals were calculated only for 15 PCNs, for which the experimentally determined data on water solubility have been available. In case of the local model that covered a narrow calibration domain (consisted of very similar chloronaphthalene congeners only), the prediction errors were considerably lower than the prediction errors of the global model with a wider domain (all POPs). By employing Student's t test, we confirmed that the average residuals (for 15 PCNs) for both models differed significantly ($t = 4.40$, $p = 0.0006$).

Therefore, from the qualitative point of view, an application of the local model should be recommended as being more accurate and precise. However, the performance of the evaluated global model for POPs was still fairly good in comparison with other, more general QSPRs. For instance, Delaney [36] put together statistics of 10 recently published QSPR models of water solubility, calibrated on training sets containing between 150 and 2874 compounds. Then, the models' predictivity was tested on the same 21 compounds having a common chemical structure. The author found the standard errors of prediction for those 21 chemicals varied between 0.55 and 0.91 logarithmic units. Regarding that the higher residual observed for our "worse" global model for POPs was about one logarithmic unit, it may be concluded our global model predicts water solubility up to three times better than the general models reviewed by Delaney.

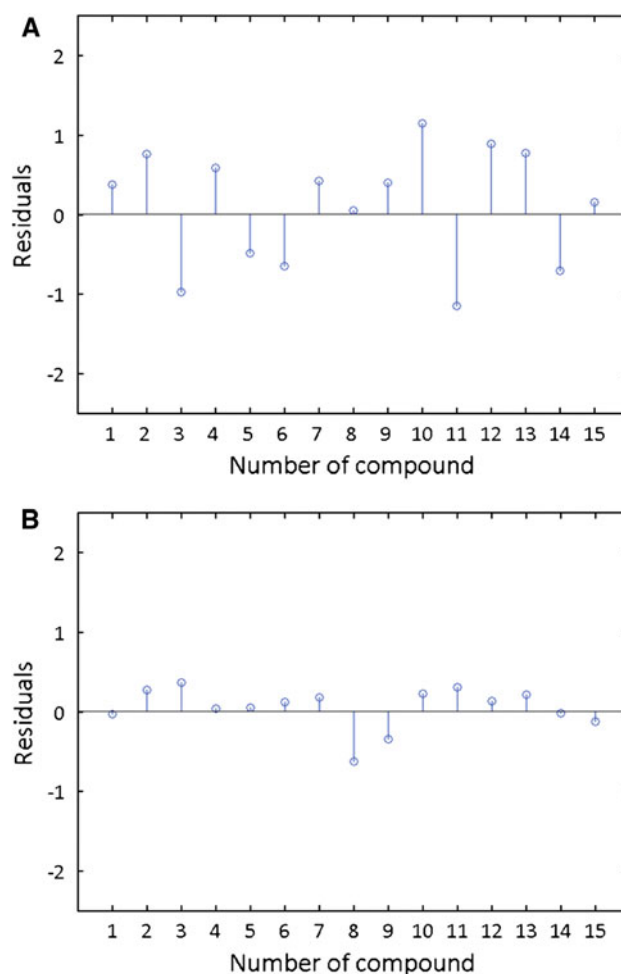


Fig. 3 Residual values (in log units) calculated for 15 chloronaphthalene congeners based on the predictions with global (a) and local (b) QSPR models

From the economical point of view, an optimal QSPR model should characterize by two features: (i) it should be based on possibly small number of training/validation compounds, without necessity to perform extensive experimental work and, simultaneously, (ii) it should ensure making predictions within a possibly wide applicability domain.

The minimal number of compounds required for developing a QSPR model is defined by the ratio between the number of descriptors and training compounds. According to the criterion proposed by Topliss and Costello [37], this ratio should be at least 5:1. The local model that utilized one variable (latent vector) has been calibrated on 10 training compounds, whereas the global model that utilized three latent vectors has been calibrated on 91 training compounds. Thus, both studied models met the criterion, since the ratios were 10:1 for the local model and 30:1 for the global model, respectively.

There are no formal requirements related to the number of validation compounds, but different authors give some recommendations, based on their experience. For instance, Gramatica [30] recommends having at least five compounds to perform an appropriate external validation. Both models fulfilled this recommendation, since the number of validation compounds was 5 for the local and 30 for the global QSPR. However, according to our experience, when the validation set is small (of about 10 compounds and less), the results of external validation could be less reliable. It is because, in such a case, the validation statistics (Q_{Ext}^2 and RMSE_P) strongly depend on the splitting algorithm. Indeed, they can significantly change, when one validation compound is replaced with another one [38]. Therefore, the external validation of our global model of $\log S$ seems to be more reliable in comparison to the external validation of the local one.

Applicability domains of both models were verified by using the Williams plots (Fig. 4). The global model has been calibrated and validated on congeners of CBzs (10 training and 2 validation compounds), PCDEs (25 training and 6 validation compounds), PBDEs (6 training and 3 validation compounds), PCBs (24 training and 10 validation compounds), PCDDs (11 training and 4 validation compounds), PCDFs (6 training and 2 validation compounds), and PCNs (9 training and 2 validation compounds). Water solubility of all validation compounds was predicted with the residuals lower than the critical threshold values (0 ± 3 standard deviations). This means the model can be successfully applied for predicting the values of $\log S$ for all seven groups of POPs listed above. Interestingly, three compounds from the training set (Fig. 4a) had the leverage values higher than the critical one ($h^* = 0.14$). The compounds are perchlorinated benzene (CBz-12), perchlorinated naphalene (PCN-75), and perchlorinated biphenyl (PCB-209). But, simultaneously, their residuals were low. This suggests the model is well stabilized by the existence of so-called “good leverage points.” In addition, the model is probably capable to perform reliable predictions for the compounds not differing substantially from the training set, but formally situated outside of the applicability domain. The last conclusion, however, should be confirmed by an additional testing with an additional validation set of compounds that have high leverage values. In a similar way, low residuals and leverage values for all 10 training and 5 validation compounds (Fig. 4b) confirmed that the local model can be applied to make satisfying predictions of water solubility within the group of chloronaphthalenes.

The last aspect that should be taken into account when comparing both models is the selection of molecular descriptors employed in each case. One can be surprised that we compared two models utilizing the descriptors

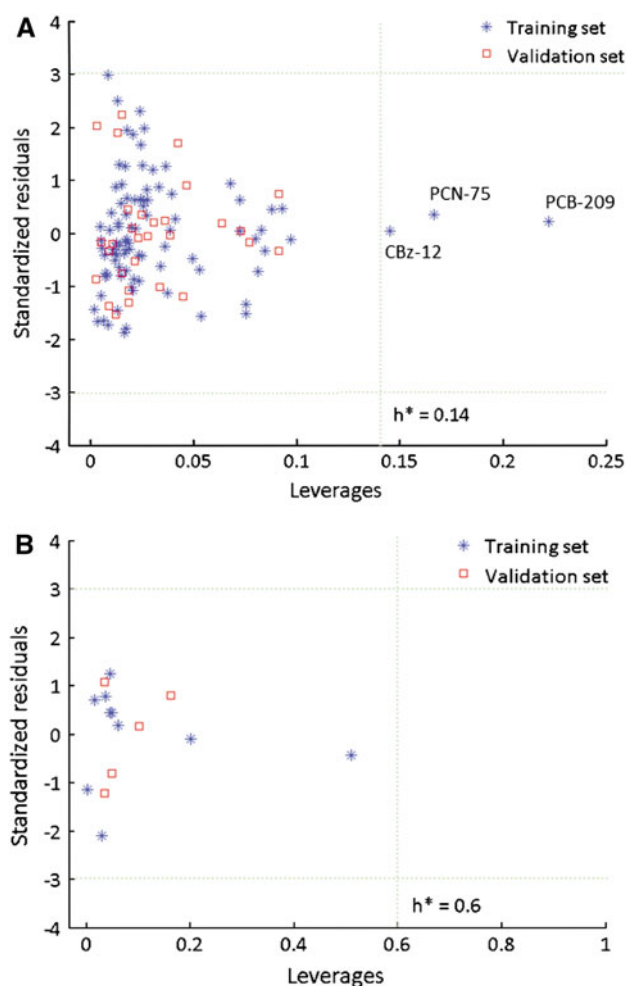


Fig. 4 Williams plot describing applicability domains of global (a) and local (b) QSPR models. Dotted lines represent the residual threshold (0 ± 3 standard deviation units), and the critical leverage value (h^*), respectively

calculated at different levels of theory (i.e., the global model has been developed based on molecular descriptors from semiempirical PM6 calculations, whereas the local model used DFT descriptors). However, we previously demonstrated [39] that eventual differences in the numerical values of molecular descriptors for POPs calculated with both methods could be neglected. We proved that QSPR models employing the descriptors calculated at the level of novel semiempirical methods (PM6 and RM1) were of similar accuracy that the models utilizing descriptors from DFT (B3LYP functional with 6-311++G(*d*, *p*) basis set). This level of accuracy was out of reach for the models employing earlier semiempirical methods (e.g., PM3 and AM1).

Moreover, it may be unclear why, when putting together both model equations (Eqs. 10 and 11) for the same property ($\log S$), the selected descriptors are different (e.g.,

LUMO in Eq. 10 and HOMO in Eq. 11)? To clarify these apparent contradictions, one needs to refer to the theory of dissolution (described in section “Global QSPR model of water solubility”) and to keep in mind three following important issues.

First, the quantum–mechanical descriptors that we used are internally correlated. Thus, they form groups of descriptors related to the same “global” property (latent vectors) and, because of that, having very similar meaning. In consequence, one descriptor from the particular group (latent vector) can be replaced with another one from the same group without changing of the global interpretation of the model. For instance, in a group of chlorinated congeners, both total energy and the solvent accessible surface area mainly depend on the number of chlorine substituents present in molecules with the same carbon skeleton. Therefore, in this context, both descriptors have very similar meaning. For that reason, we decided to use PLS method of modeling instead of much simpler and more intuitively interpretative multiple linear regression (MLR).

Second, molecular descriptors for both local and global models were selected with use of the genetic algorithm. The algorithm is, in fact, an automatic probability-based procedure, blind on the mechanistic interpretation. In effect, when the algorithm has a choice between two strongly correlated descriptors related to the same “global” property (see above), it might select the first or the second descriptor only by chance.

Third, when considering a local model, developed for only one congeneric group (i.e., polychlorinated naphthalenes), the model is much more sensitive on the number of substituents (chlorine atoms) and the substitution pattern than the global model calibrated for more groups, in which the main differences between particular compounds are related to their carbon skeletons (i.e., the number of aromatic rings, presence of heteroatoms, etc.).

Hence, no one should expect exactly the same model equations for the global and local models being compared in our study. In the context of the dissolution mechanism, three structural features (“global” properties) of POPs’ congeners seem to be very important. They are: (i) the size of the parent molecule (carbon skeleton), (ii) the type and the number of substituents present in the molecule, and (iii) the substitution pattern. The first “global” property is obviously related to the cavitation process. We observed that the solubility decreases with the increasing size of the molecule. The type and the number of substituents are, of course, also strongly related to the size, and consequently, to the cavitation stage. Generally, molecules based on the same skeleton and substituted with the same number of bromine atoms are less soluble than their chlorinated analogues, due to larger radius of bromine substituents in comparison to chlorine atoms. Similarly, the solubility

increases with the increasing number of halogen substituents (e.g., monochloronaphthalenes are more soluble than dichloronaphthalenes). The descriptors related to the factors influencing the cavitation process, namely: the size of a molecule, the type and number of substituents are nAT , nX , and SAS in Eq. 10, as well as E_t , SAS_w , SAV_w , DE_w , and TNE_w in Eq. 11.

The substitution pattern is the main factor deciding on differences in solubility between congeners containing the same number of substituents of the same type. Differences in the distribution of the substituents (over the same carbon skeleton) decide on differences in polarity of particular congeners. For example, 1,2,3,4-tetrachloronaphthalene is more polar than 2,3,6,7-tetrachloronaphthalene. Subsequently, electrostatic interactions with water as a solvent are stronger in the second case. Thus, the second congener in this pair is more soluble. Interestingly, as we demonstrated in many previous contributions [35, 40–42] such descriptors as HOMO and LUMO are strongly dependent on the substitution pattern. Thus, in study they should not be interpreted as those describing redox properties of the molecules (according to the well-known Koopman’s theorem), but rather their substitution patterns. Another descriptors related to the substitution pattern are Q_+ and Shift in Eq. 10, as well as $nClpl$ and Hard in Eq. 11. Therefore, the mechanistic interpretation of both global and local QSPR models would be very similar.

In summary, from the economical point of view, both models are acceptable, since they require a relatively small number of experimental data. In fact, both are based on the data taken from the literature, thus performing of any extensive empirical work was unnecessary. However, the use of the global QSPR would be more profitable, because it enables to make predictions for those groups of POPs, for which the number of experimental data is insufficient to develop appropriate local models. For example, the experimentally determined data on water solubility are available only for eight congeners of PCDFs, which is evidently too small for calibrating and validating a local model. Moreover, time and, in consequence, costs of obtaining the predicted values of $\log S$ can be significantly reduced by employing the global modeling scheme.

Comparing other global and local QSPR models

In addition, to extend the investigations on the other phys/chem properties, we performed similar pairwise comparisons for the other, previously published QSPR models. We used two of our previous global models developed for predicting n-octanol/water partition coefficient ($\log K_{OW}$) [40] and subcooled liquid vapor pressure ($\log P_L$) [42],

Table 2 Statistical parameters of local and global models of $\log P_L$ and $\log K_{OW}$

Model	Feature	Measure	Local QSPR model	Global QSPR model
$\log P_L$	Goodness-of fit	R^2	0.99 ^a	0.97
		RMSE _C	n/a	0.21
	Robustness	Q_{CV}^2	0.96 ^a	0.97
		RMSE _{CV}	n/a	0.22
	Predictivity	Q_{Ext}^2	n/a	0.97
		RMSE _P	n/a	0.22
$\log K_{OW}$	Goodness-of fit	R^2	0.91 ^b	0.92
		RMSE _C	0.23 ^b	0.32
	Robustness	Q_{CV}^2	0.91 ^b	0.92
		RMSE _{CV}	0.23 ^b	0.32
	Predictivity	Q_{Ext}^2	n/a	0.92
		RMSE _P	n/a	0.30

^a Taken from [43], ^b taken from [44]

n/a No data provided in the original paper

respectively, for a group of 1436 POPs, including chloro- and bromo-analogues of dibenzo-*p*-dioxins, dibenzofurans, biphenyls, naphthalenes, diphenyl ethers, and benzenes. The models were compared with two corresponding local models published by other groups. The first one was designed to predict $\log K_{OW}$ for 209 PCBs [43], whereas the second one the values of $\log P_L$ for 210 congeners of PCDDs and PCDFs [44].

Interestingly, the conclusions from both comparisons (based on predicting $\log K_{OW}$ and $\log P_L$) are even more optimistic than that for $\log S$. The statistical measures of goodness-of-fit and robustness were very similar in pairs for corresponding global and local models (Table 2). Moreover, the observed differences between the experimentally measured and predicted values of by both methods of modeling (i.e., local and global) were not statistically significant ($p > 0.05$) (Table 3), which was consistent with our assumption. Regarding that (i) both global models have been developed for a much wider applicability domain (covering of about 85% more compounds) and (ii) they practically did not differ from their local counterparts in quality, we concluded that the employment of global QSPRs would be much more efficient than the development of particular local ones.

Table 3 Comparison between the residuals derived from the predictions of $\log P_L$ and $\log K_{OW}$ with local and global GA-PLS models by the Student *t* test

Model	Statistics	
$\log P_L$	Test statistic	2.01
	<i>p</i> Value	0.051
$\log K_{OW}$	Test statistic	1.81
	<i>p</i> Value	0.072

Conclusions

We have verified the efficiency of two modeling strategies. The first one assumes the reduction of the model's domain and the development of QSPR based on a small number of structurally similar compounds (local QSPR). According to the second one, the model is calibrated with use of the wider and more structurally diversified training set (global QSPR), even if this leads to a small decrease of the model's predictivity.

Based on the obtained results, we recommend that whenever global models fulfill all quality criteria proposed by the Organization for Economic Cooperation and Development (OECD), they should be applied in practice without necessity of developing a series of local QSPRs. Such a recommendation is reasonable, because of three reasons. First, the global models allow for simultaneous predictions of physicochemical properties for even many hundreds of compounds. This feature is very important from the economic point of view, regarding that the number of new chemicals synthesized and/or identified in the environmental compartments is growing exponentially. Second, the global modeling approach may be the only possibility of modeling, when the number of chemicals from one specific class of the chemically related compounds is insufficient to calibrate and appropriately validate a local QSPR model. Third, as demonstrated, the performance (predictive ability) of global models is not always worse than these of local ones.

Acknowledgments The authors thank the editors for rapidly considering our submission and the anonymous reviewer for valuable comments, which helped to improve scientific quality of this contribution. T.P. thanks the Foundation for Polish Science for granting him with a fellowship and a research grant in frame of the HOMING Program supported by Norwegian Financial Mechanism and EEA Financial Mechanism in Poland. This work was supported by the

Polish Ministry of Science and Higher Education, Grant No. DS/8430-4-0171-11. This research was supported in part (to M.H.) by the U.S. Department of Energy under contract DE-AC02-05CH11231.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Yang G, Zhang X, Wang Z, Liu H, Ju X (2006) Estimation of the aqueous solubility ($-lgS_w$) of all polychlorinated dibenzo-furans (PCDF) and polychlorinated dibenzo-p-dioxins (PCDD) congeners by density functional theory. *J Mol Struct Theochem* 766:25–33
2. Rotkin-Ellman M, Navarro KM, Solomon GM (2010) Gulf oil spill air quality monitoring: lessons learned to improve emergency response. *Environ Sci Technol* 44:8365–8366
3. UNEP (2001) Stockholm convention on persistent organic pollutants. United Nations Environment Programme, Geneva
4. Harańczyk M, Puzyn T, Ng EG (2010) On enumeration of congeners of common persistent organic pollutants. *Environ Pollut* 9:2786–2789
5. Schultz TW, Cronin MTD, Walker JD, Aptula AO (2003) Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *J Mol Struct Theochem* 622:1–22
6. Castro EA, Toropova AP, Toropov AA, Mukhamedjanova DV (2005) QSPR modeling of Gibbs free energy of organic compounds by weighting of nearest neighboring codes. *Struct Chem* 16:305–324
7. Golmohammadi H, Dashtbozorgi Z (2010) Quantitative structure–property relationship studies of gas-to-wet butyl acetate partition coefficient of some organic compounds using genetic algorithm and artificial neural network. *Struct Chem* 21: 1241–1252
8. Öberg T, Liu T (2008) Global and local PLS regression models to predict vapor pressure. *QSAR Comb Sci* 27:273–279
9. Lei B, Ma Y, Li J, Liu X, Yo X, Gramatica P (2010) Prediction of the adsorption capability onto activated carbon of a large data set of chemicals by local lazy regression method. *Atmos Environ* 44:2954–2960
10. Hayward D (1998) Identification of bioaccumulating polychlorinated naphthalenes and their toxicological significance. *Environ Res* 76:1–18
11. Dearden JC, Cronin MTD, Kaiser KLE (2009) How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ Res* 20: 241–266
12. Dunnivant FM, Elzerman AW (1988) Aqueous solubility and Henry's law constant for PCB congeners for evaluation of quantitative structure-property relationships (QSPRs). *Chemosphere* 17:525–541
13. Miller MM, Ghodbane S, Wasik SP, Tewari YB, Martire DE (1984) Aqueous solubilities, octanol/water partition coefficients, and entropies of melting of chlorinated benzenes and biphenyls. *J Chem Eng Data* 29:184–190
14. Govers HAJ, Krop HB (1998) Partition constants of some chlorinated dibenzofurans, and dibenzo-p-dioxins. *Chemosphere* 37:2139–2152
15. Ruelle P, Kesselring UW (1997) Aqueous solubility prediction of environmentally important chemicals from the mobile order thermodynamics. *Chemosphere* 34:275–298
16. Tittlemier SA, Halldorson T, Stern GA, Tomy GT (2002) Vapor pressures, aqueous solubilities, and Henry's law constants of some brominated flame retardants. *Environ Toxicol Chem* 21:1804–1810
17. Opperhuizen A, Velde EW, Gobas FAP, Liem DAK, Steen JMD (1985) Relationship between bioconcentration in fish and steric factors of hydrophobic chemicals. *Chemosphere* 14:1871–1896
18. Doucette WJ, Andren AW (1988) Aqueous solubility of selected biphenyl, furan, and dioxin congeners. *Chemosphere* 17:243–252
19. Hewitt M, Cronin MTD, 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
20. Harańczyk M, Puzyn T, Sadowski P (2008) ConGENER—a tool for modeling of the congeneric sets of environmental pollutants. *QSAR Comb Sci* 27:826–833
21. Harańczyk M, Gutowski M (2007) Quantum mechanical energy-based screening of combinatorially generated library of tautomers. TauTGen: a tautomer generator program. *J Chem Inf Model* 47:686–694
22. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J Mol Model* 13:1173–1213
23. Stewart JJP (2009) In: Chemistry SC (ed) MOPAC2009 <http://openmopac.net/MOPAC2009.html>. Accessed 14 April 2010
24. OECD (2004) OECD principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models. In: 37th joint meeting of the chemicals committee and working party on chemicals, pesticides and biotechnology. Organisation for Economic Co-Operation and Development, Paris
25. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab Syst* 58:109–130
26. Holland J (1992) Adaptation in natural and artificial systems. MIT Press, Michigan, MI
27. MATLAB (2008) MATLAB 7.6.0.324. Mathworks
28. PLS_Toolbox (2009) PLS_Toolbox 5.2. Eigenvector Research Inc., Wenatchee, WA
29. 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
30. Gramatica P (2007) Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 26:694–701
31. Atkinson AC (1985) Plots, transformations, and regression. An introduction to graphical methods of diagnostic regression analysis. Oxford Statistical Science Series, Oxford
32. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska J, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C (2005) Current status of methods for defining the applicability domain of (quantitative structure–activity relationships). The report and recommendations of ECVAM workshop 52. *Altern Lab Anim* 33:155–173
33. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *Altern Lab Anim* 33:445–459
34. OECD (2007) Guidance document on the validation of (quantitative) structure–activity relationships (QSAR) models. Organisation for Economic Co-Operation and Development, Paris
35. Puzyn T, Mostrąg A, Falandysz J, Kholod Y, Leszczynski J (2009) Predicting water solubility of congeners: chloronaphthalenes—a case study. *J Hazard Mater* 170:1014–1022

36. Delaney JS (2005) Predicting aqueous solubility from structure. *Drug Discov Today* 10:289–295
37. Topliss JG, Costello RJ (1972) Chance correlations in structure–activity studies using multiple regression analysis. *J Med Chem* 15:1066–1068
38. Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, Skrzyński M, Worth AP (2011) Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Struct Chem*
39. Puzyn T, Suzuki N, Harańczyk 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
40. Puzyn T, Suzuki N, Harańczyk M (2008) How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine? *Environ Sci Technol* 42:5189–5195
41. Puzyn T, Mostrąg A, Suzuki N, Falandysz J (2008) QSPR-based estimation of the atmospheric persistence for chloronaphthalene congeners. *Atmos Environ* 42:6627–6636
42. Gajewicz A, Harańczyk M, Puzyn T (2010) Predicting logarithmic values of the subcooled liquid vapor pressure of halogenated persistent organic pollutants with QSPR: how different are chlorinated and brominated congeners? *Atmos Environ* 44: 1428–1436
43. Padmanabhan J, Parthasarathi R, Subramanian V, Chattaraj PK (2006) QSPR models for polychlorinated biphenyls: n-octanol/water partition coefficient. *Bioorg Med Chem* 14:1021–1028
44. Yang P, Chen J, Chen S, Yuan X, Scharmm KW, Kettrup A (2003) QSPR models for physicochemical properties of polychlorinated diphenyl ethers. *Sci Total Environ* 305:65–76