Journal of Hazardous Materials 170 (2009) 1014–1022

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Predicting water solubility of congeners: Chloronaphthalenes—A case study

Tomasz Puzyn[a,*], Aleksandra Mostrąg[a], Jerzy Falandysz[a], Yana Kholod[b], Jerzy Leszczynski[b]

[a] Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland
[b] NSF CREST Nanotoxicity Center, Department of Chemistry, Jackson State University, 1325 Lynch St, Jackson, MS 39217-0510, USA

ARTICLE INFO

Article history:
Received 6 November 2008
Received in revised form 15 April 2009
Accepted 18 May 2009
Available online 22 May 2009

Keywords:
Chloronaphthalenes
QSPR
DFT
WSKOWIN
COSMO-RS

ABSTRACT

Since the important physicochemical data for chloronaphthalenes (PCNs) are still scarce, we have predicted water solubility (log S) of all 75 congeners with the Quantitative Structure–Property Relationship (QSPR) scheme. The values of log S, predicted by the most efficient model, varied from 0.01 to 1660 μg dm^{-3} (2.85×10^{-11}–1.02×10^{-5} mol dm^{-3}), depending on the number of chlorine atoms present in the molecule and the substitution pattern. We found that the main factor determining relative differences in solubility between the congeners is the solvent accessible volume related to the cavitation process occurring in the solvent. The results are presented as a case study of QSPR modeling for those Persistent Organic Pollutants (POPs) that exist as families of congeners. By investigating the impact of (i) the way of the molecular descriptors' calculation, (ii) the size of applied database and (iii) chemometric method of modeling (Multiple Linear Regression, MLR, and/or Partial Least Squares regression, PLS) on the quality of the models we proposed general recommendations for dealing with congeners. We found that the combination of the B3LYP functional with 6-311++G(d,p) basis set was the most optimal technique of the molecular descriptors' calculation for congeners when comparing with semi-empirical PM3, *ab initio* Hartee–Fock (HF), and Møller–Pleset 2 (MP2) method carried out with different-size basis sets. Moreover, the model developed with a larger and more general database that includes chloronaphthalenes, polychlorinated dibezno-*p*-dioxins, furans and biphenyls predicted the values of log S for PCNs noticeable worse than the model calibrated only on PCNs. In the later case it was possible to obtain satisfactory results by employing even the simplest MLR method and only one molecular descriptor. The values of log S were also calculated with the WSKOWIN and COSMO-RS models as the reference techniques and then compared to our results.

1. Introduction

Solubility in water plays one of the most important roles among many physicochemical parameters that characterize a chemical pollutant. It influences behavior of the chemical compound in many physical and biological processes, involving information on the ability of the compound to take part in metabolic processes as well as assessing its environmental persistence, transport and fate [1].

Polychlorinated naphthalenes (chloronaphthalenes, PCNs, CNs) form a set of 75 two-ringed aromatic compounds, containing from one to eight chlorine atoms per molecule in different positions (congeners). CNs were commercially synthesized between 1910s and 1970s and used in many technical applications. All CN congeners are planar and some of them (CNs nos. 48, 54, 66, 67, 68, 69, 70, 71, 73, and 75) belong to the group of so-called 'dioxin-like' chemicals due to their contribution to the aryl hydrocarbon receptor-mediated mechanism of toxicity. Their confirmed persistence in the natural environment, tendency to be accumulated in biota and extensive toxicity are the reasons that chloronaphthalenes are concerned as one of the major groups of hazardous environmental pollutants [2–4].

According to the results of the global atmospheric passive sampling study [5], the total concentration of CN congeners in the atmosphere measured between December 2004 and March 2005 ranged from the levels below detection limit to 32 pg/m^3 with a geometric mean of 1.6 pg/m^3. Significant levels of PCNs were noticed primarily in the northern hemisphere with the highest concentrations in the urban, industrial areas of Eastern Europe (Czech Republic, Poland) and China. Elevated levels were observed also at the urban sites of Turkey, Kuwait and Philippines. At Arctic sites the total concentration of PCNs ranged between 1 and 8 pg/m^3, confirming the long-range transport potential of these compounds. The most frequently detected congeners were CN nos. 24, 33/34/37, 47, 27/30/39, 52/60, 50, 51, 54, 66/67, and 75. Interestingly, PCN air concentrations are declining more slowly than expected, taking into account fact that the technical mixtures of chloronaphthalenes are no longer widely used. This could be attributed to continued emission of PCNs from various combustion sources (waste incineration, domestic heating, etc.). Thus, the problem of PCNs is still up-to-date and should be further studied.

* Corresponding author. Tel.: +48 58 523 5451; fax: +48 58 523 5472.
E-mail addresses: puzi@qsar.eu.org, puzi@pcb.chem.univ.gda.pl (T. Puzyn).

0304-3894/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2009.05.079

Table 1
Molecular descriptors calculated in the study.

No.	Symbol	Name	Unit	Method of calculation
1.	*nCl*	The total number of chlorine atoms	–	Manually
2.	*nClp1*	The number of chlorine atoms in the first aromatic ring	–	Manually
3.	*nClp2*	The number of chlorine atoms in the second aromatic ring	–	Manually
4.	*D*	The dipole moment	Debye	Gaussian 03, DFT
5.	*A*	Mean polarizability[a]	Å^3	Gaussian 03, DFT
6.	*MaxQ+*	The maximal positive partial Mulliken's charge	–	Gaussian 03, DFT
7.	*MaxQ-*	The maximal negative partial Mulliken's charge	–	Gaussian 03, DFT
8.	*HOMO*	The energy of the highest occupied molecular orbital	Hartree	Gaussian 03, DFT
9.	*LUMO*	The energy of the lowest unoccupied molecular orbital	Hartree	Gaussian 03, DFT
10.	*Hard*	The molecular hardness[b]	Hartree	Gaussian 03, DFT
11.	*IP*	The adiabatic ionization potential	eV	Gaussian 03, DFT
12.	*EA*	The adiabatic electron affinity	eV	Gaussian 03, DFT
13.	*Et*	The total energy of the molecule	Hartree	Gaussian 03, DFT
14.	*Cv*	The heat capacity (for v = const)	kJ mol^{-1}	Gaussian 03, DFT
15.	*S*	Entropy	J mol^{-1} K^{-1}	Gaussian 03, DFT
16.	*SASw*	The solvent accessible molecular surface area in the water[c]	Å^2	Gaussian 03, DFT
17.	*SAVw*	The solvent accessible molecular volume in the water[c]	Å^3	Gaussian 03, DFT
18.	*TEESolw*	The total electrostatic energy of solvation in the water[c]	Hartree	Gaussian 03, DFT
19.	*PolSSw*	The polarized solute–solvent interaction energy in the water[c]	kJ mol^{-1}	Gaussian 03, DFT
20.	*CEw*	The cavitation energy in the water[c]	kJ mol^{-1}	Gaussian 03, DFT
21.	*DEw*	The dispersion energy in the water[c]	kJ mol^{-1}	Gaussian 03, DFT
22.	*TNEw*	The total non-electrostatic energy of solvation[c]	kJ mol^{-1}	Gaussian 03, DFT

[a] Mean polarizability (*A*) was calculated as a mean eigenvalue taken from diagonalization of the polarizability tensor.
[b] Hardness (*Hard*) was computed as a half of the HOMO–LUMO energy difference.
[c] Calculations in solute (water) simulated by the Conductor-like Screening MOdel (COSMO).

Although data on water solubility of CN congeners are of vital importance in environmental modeling, such information is scarce. The experimental data are available only for selected 15 CN congeners [6]. These, however, incomplete data have been cited many times in the scientific reports devoted to risk assessment of chloronaphthalenes [7,8]. Analyzing the reports we pointed out that the prediction of water solubility for the remaining 60 compounds from this group would be imperative, and it would fill significant gaps in the existing data. If the complete data exist, it would be possible to model environmental transport and fate not only for 15, but also for all 75 congeners.

Problems with lacking data are also very common for similar congeneric compounds, such as polychlorinated and polybrominated dibenzo-*p*-dioxins, dibezofurans, diphenyl ethers, tiophenes, etc. Specificity of this groups results from a relatively small variability in both physicochemical and molecular parameters characterizing individual members of the congeneric families. As a consequence, computational predictions for such compounds require very accurate techniques that are able to express these small differences among the congeners, especially containing the same number of chloro- or bromo-substituents (i.e., 1,2-dichloronaphthalene and 1,4-dichloronaphthalene).

The efficient way to obtain a complete set of the data, without necessity of performing expensive laboratory experiments is application of the Quantitative Structure–Property Relationship (QSPR) techniques, including advanced quantum chemical, combinatorial and chemometric methods [9–12].

The main purpose of this study was to predict reliable values of water solubility for all chloronaphthalene congeners. However, we also wanted to present the results as a case study and discuss some methodological aspects interesting for QSPR modelers dealing with the congeneric compounds. By investigating the impact of (i) the way of the molecular descriptors' calculation, (ii) the size of applied database and (iii) the chemometric methods of modeling (Multiple Linear Regression, MLR, and/or Partial Least Squares regression, PLS) on the quality of the models we developed some more general recommendations for dealing with congeners. We have confirmed the hypothesis that in the case of such congeneric compounds as chloronaphthalenes, the application of relatively simple local QSPRs, even constructed on the very limited, but homogenous experimental database, could provide more precise results than comprehensive and 'universal' models.

2. Materials and methods

In the first step, the structures of all possible CN congeners were combinatorially generated with the ConGENER package [13]. Then, the molecular coordinates of the structures were optimized at the Density Functional Theory (DFT) level, using B3LYP functional and 6-311++G(d,p) basis set [14–17]. Molecular geometry of each congener was optimized three times *in vaccuo*: (i) as a neutral molecule, (ii) a corresponding cation and (iii) a corresponding anion. In addition, each molecule (as a neutral molecule) was optimized in water simulated by the Conductor-like Screening MOdel (COSMO) [18]. All quantum-mechanical calculations were carried out using the Gaussian 03 package [16].

In the second step, 22 molecular descriptors were either directly extracted from the Gaussian output files or calculated manually (Table 1). After obtaining the descriptors, additional calculations by Hartree–Fock (HF), Møller–Pleset with the 2nd order corrections (MP2), and semi-empirical PM3 methods (for details see [17]) were additionally performed to evaluate the influence of the calculation method and the basis set on the descriptor values. Because of the limited number of experimental data, this influence was examined based on the dipole moments for mono- and dichloronaphthalenes taken from a database by Eucken and Hellwege [19]. The dipole moments seem to be a very good measure for such comparison, because they arise from electron distribution and they are more sensitive to molecular geometry than most the other properties. Moreover, the dipole moments from *ab initio* (HF) and DFT (B3LYP) calculations can be compared to the semi-empirical results more reliable than the energies, because in the case of the semi-empirical methods, the energies are related to the heat of formation [17].

A compilation of all available experimental data on water solubility of PCNs and critical evaluation of those data preceded the next step of modeling. As already have been mentioned, experimentally derived values of water solubility were available only for 15 CN congeners (20% of all congeners) [6]. They ranged from 0.08 to 2870 μg dm^{-3} (2.0×10^{-10}–1.8×10^{-5} mol dm^{-3}). Those congeners, for which the experimental data had been available (Set 1), were

Table 2
Comparison of the experimental and calculated values of the dipole moment D (in Debyes) for mono- and dichloronaphthalenes.

CN congener	Exp.[a]	Calculated									
		HF/6-31G(d)		MP2/6-31G(d)		B3LYP/6-31G(d)		B3LYP/6-311++G(d,p)		PM3	
	D	*D*	Res.	*D*	Res.	*D*	Res	*D*	Res.	*D*	Res.
1-Chloronaphthalene	1.59	2.07	0.48	1.79	0.20	1.85	0.26	1.79	0.2	0.90	−0.69
2-Chloronaphthalene	1.72	2.36	0.64	2.07	0.35	2.13	0.41	2.04	0.32	1.07	−0.65
1,2-Dichloronaphthalene	2.47	3.38	0.91	2.94	0.47	2.99	0.52	2.85	0.38	1.45	−1.02
1,3-Dichloronaphthalene	1.78	2.55	0.77	2.25	0.47	2.30	0.52	2.2	0.42	1.15	−0.63
1,4-Dichloronaphthalene	0.48	0.75	0.27	0.75	0.27	0.73	0.25	0.69	0.21	0.37	−0.11
1,5-Dichloronaphthalene	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	0.00	0.00
1,6-Dichloronaphthalene	1.44	1.90	0.46	1.59	0.15	1.66	0.22	1.59	0.15	0.83	−0.61
1,7-Dichloronaphthalene	2.55	3.41	0.86	2.93	0.38	3.01	0.46	2.91	0.36	1.48	−1.07
1,8-Dichloronaphthalene	2.82	3.82	1.00	3.24	0.42	3.33	0.51	3.21	0.39	1.59	−1.23
2,3-Dichloronaphthalene	2.55	3.65	1.10	3.12	0.57	3.20	0.65	3.01	0.46	1.60	−0.95
2,6-Dichloronaphthalene	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2,7-Dichloronaphthalene	1.53	2.06	0.53	1.84	0.31	1.86	0.33	1.8	0.27	0.91	−0.62
Mean error	–	–	0.59	–	0.30	–	0.34	–	0.26	–	−0.63

[a] Experimental data taken from [19].

divided into two subsets: the training subset containing 10 congeners and the validation subset with 5 compounds. The splitting procedure was designed to assure that there are the high, low and medium values of water solubility represented in each subset. The validation compounds were randomly selected from those characterized by the values of log S below 1.0 (2 congeners), log S between 1.0 and 3.0 (2 congeners), and log S over 3.0 (1 congener). The remaining congeners were used as the training set.

In the same way, we have collected the solubility data for a wider set of structurally similar, chloroaromatic, congeneric compounds, known as environmental pollutants (Set 2). They were: previously characterized chloronaphthalenes (15 congeners), polychlorinated biphenyls, PCBs (15 congeners), polychlorinated dibenzo-*p*-dioxins, PCDDs (15 congeners), and polychlorinated dibenzofurans, PCDFs (7 congeners) [20]. The values of water solubility in this extended database varied between 7.4×10^{-5} and 7079 $\mu g\,dm^{-3}$ (1.61×10^{-13}–4.59×10^{-5} mol dm^{-3}). The compounds from Set 2 were also split into the training subset and the validation subset. The molecular structures of these compounds were optimized with the same quantum-mechanical methods as applied to PCNs and, finally, the same set of 22 descriptors as derived before was calculated.

Then, based on the descriptors and the experimental data, we developed QSPR models for predicting water solubility of PCNs. To be able to investigate the influence of the model's domain on the prediction error, we simultaneously constructed individual models with use of the compounds from Set 1 and Set 2. Moreover, in each case we used two different chemometric approaches: Multiple Linear Regression (MLR) method, and Partial Least Squares (PLS) regression to determine the influence of the modeling technique on the output. The same training and validation subsets were used in each model in order to make the comparison between the models more reliable. MLR and PLS are both the standard techniques and their detailed characteristics are described elsewhere [21]. In the case of MLR, we arbitrary selected the best model, after considering possible solutions and verifying orthogonality of the descriptors to avoid the known problem of yielding a very good MLR model just by chance when a large number of 'screened' descriptors are calculated [22]. The most optimal set of the descriptors for PLS were selected by employing the standard Genetic Algorithm (GA) implemented in the PLS Toolbox 4.1 [23]. Before modeling, all descriptors were autoscaled (transformed to mean equal 0 and variance equal 1) for equalizing the impact of each variable in the models.

The obtained QSPR models were validated according to the best practice and the five OECD recommendations [24,25]. These 'golden rules' state that the 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; and (v) a mechanistic interpretation, if possible.

The applicability domains of the models were explored by use of the Williams plots. The plots played a double role. Firstly, they described the impacts of the objects on models by the values of the objects' leverages (diagonal elements of the Hat or Influence Matrix $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$). Secondly, they presented the Euclidean distances of the compounds to the models measured by the jackknifed (standardized and cross-validated) residuals. The leverage (h) greater than the warning $h^* = 3p'/n$ (p': the number of variables plus one; n: the number of compounds in the training set) suggested that the compound was very influential on the model, while the residual standard deviations R.S.D. > 2.5 classified the compound as an outlier. The robustness of each model was expressed by the cross-validated (leave-one-out technique, LOO) validation coefficient (Q^2_{LOO}) and the root mean square errors of LOO cross-validation (RMSECV). The predictive abilities of the models were compared to each other according to the values of the external validation coefficients (Q^2_{ext}) and the root means square error of prediction in the validation sets (RMSEP) [24,26]. Successfully validated QSPR models with confirmed predictive abilities were used to predict water solubility for all 75 CN congeners. All QSPR calculations were performed in the MATLAB 7.6 environment [27].

In addition, the quality of the QSPR results was verified by comparing the predictions to the results obtaining with use of other common computational techniques. We performed reference calculations with the WSKOWIN [28] and the COSMO-RS [29,30] models.

WSKOWIN is a software package recommended by U.S. Environmental Protection Agency. It calculates water solubilities from n-octanol/water partition coefficients, molecular weights and melting points using two QSPR models. These models were developed based on 1450 training compounds. The models were also externally validated on 817 compounds. The authors reported the values of standard deviation and absolute mean error of prediction as 0.615 and 0.480 log units, respectively [31].

Calculations in COSMO*therm* package are based on the COSMO-RS theory by Klamt and co-workers [30]. This theory was used to develop quantum-mechanical COnductor-like Screening Model, by putting solute and solvent in a perfect conductor and calculating the polarization charge densities for both. This way of calculation allows determining the chemical potentials for the solute and the solvent and – finally – solubility of the solute. It is worth noting, that the initial step in the COSMO*therm* algorithm

Table 3
Description of the QSPR models developed in this study.

	MLR	PLS
Set 1		
	$\log S = 11.6(\pm 0.8) - 0.015(\pm 0.001)$ *SAVw*	PLS model with 1 latent vector (8 descriptors: *nClp1*, *HOMO*, *Hard*, *Et*, *SASw*, *SAVw*, *DEw*, *TNEw*)
$n = 10$	$s = 0.324$ $F_{1,8} = 153$	
$n_{val} = 5$	*Intercept*: $p < 0.001$, $t = 12.35$	
	SAVw: $p < 0.001$, $t = 14.26$	
	Goodness-of-fit: $R^2 = 0.950$, RMSEC[a] = 0.290	Goodness-of-fit: $R^2 = 0.947$, RMSEC = 0.298
	Robustness: $Q^2_{LOO} = 0.897$, RMSECV = 0.418	Robustness: $Q^2_{LOO} = 0.947$, RMSECV = 0.383
	Predictivity: $Q^2_{ext} = 0.933$, RMSEP = 0.260	Predictivity: $Q^2_{ext} = 0.966$, RMSEP = 0.185
Set 2		
$n = 34$	–	PLS model with 4 latent vectors (8 descriptors: *nClp1*, *HOMO*, *Hard*, *Et*, *SASw*, *SAVw*, *DEw*, *TNEw*)
$n_{val} = 18$		
		Goodness-of-fit: $R^2 = 0.941$, RMSEC = 0.492
		Robustness: $Q^2_{LOO} = 0.941$, RMSECV = 0.597
		Predictivity: $Q^2_{ext} = 0.894$, RMSEP = 0.607

[a] The values of RMSEP, RMSECV, and RMSEP were calculated according to the formula:

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}})^2}{n}}$$

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{\mathrm{obs}} - y_i^{\mathrm{cvpred}})^2}{n}}$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{j=1}^{n_{val}}(y_j^{\mathrm{obs}} - y_j^{\mathrm{pred}})^2}{n_{val}}}$$

where y^{obs}: observed (experimental) value of log S, y^{pred}: predicted value of log S, y^{cvpred}: cross-validated value of log S.

is quantum-mechanical COSMO calculation—the same procedure that we performed generating the matrix of molecular descriptors for QSPR.

3. Results and discussion

3.1. Molecular descriptors

The calculated values of the molecular descriptors were characterized by the normal distribution, without the presence of outliers. Because the triple standard deviations were always greater than the errors of computation, we assumed that all descriptors would be able to reliably characterize molecular differences amongst CNs.

As mentioned, the dipole moments were utilized for comparison of quality of quantum-mechanical geometry optimization. Both HF and PM3 methods provided the poorest description of geometry and – in consequence – structural differences between the congeners (Table 2). We do not recommend application of these approaches. We observed that for the considered compounds the very fast B3LYP method in conjunction with the triple zeta basis set gave even slight better results than calculations at the costly (long time of calculations) MP2 level with the double zeta basis set. Therefore, DFT (B3LYP) methods might be recommended for similar QSPR studies.

3.2. QSPR modeling

By applying the QSPR methodology it was possible to develop only three statistically satisfying models, namely: MLR model for Set 1, PLS model for Set 1, and PLS model for Set 2. The forth possible one (MLR model for Set 2) was characterized by insufficient goodness-of-fit, thus, it was eliminated from further considerations.

When analyzing goodness-of-fit, robustness and predictivity of both models for the narrow set (Set 1) one can conclude that the differences are not very large (Table 3). However, the PLS model predicts slightly better than the MLR-based one. This might be mechanistically explained when regarding the theory of the dissolving process. The MLR utilizes only one descriptor (*SAVw*) which is directly related to the cavitation. Formation of 'caves' in the solvent (cavitation) plays the critical role in dissolving of such highly hydrophobic compounds as PCNs [1]. That is why we were able to obtain a satisfied QSPR model with the only one descriptor. It is worth noting that the solvent accessible volume, in the case of chloronaphthalenes, depends mainly on the chlorination degree—it increases from mono- to octachloronaphthalenes. The influence of the substitution pattern on the *SAVw* values is less pronounced. Of less importance for the dissolving process are the electrostatic and dispersive interactions occurring between the solvent and solute after formation of the caves. For chloronaphthalenes, those factors become especially important when comparing each other congeners with the same number of chlorine substituents. The PLS model utilizes not only descriptors related to the cavitation (*SASw*, *Et*), but also those linked to the dispersive (*DEw*, *TNEw*) and electrostatic interactions (*nClp1*, *HOMO*, *Hard*). One can be surprised, why *nClp1*, *HOMO* and *Hard* in this context are interpreted as descriptors of the electrostatic interactions. The explanation is relatively simple. Differences in electron density on particular atoms are responsible for variations of the electrostatic interactions between the solvent and various congeners. On the other hand, it is obvious that the differences in the electron density for the congeners having the same number of chlorine atoms result from various substitution patterns. Therefore, when investigating congeners with the same number of chlorine substituents, the substitution pattern seems to be the most important parameter governing the observed differences in their solubility. High concentration of the chlorine substituents on one aromatic ring (denoted by *nClp1*), because of their electron withdrawing properties, favors formation of a small dipole moment in the molecule. Also the energy of the highest occupied molecular orbital (*HOMO*) and the molecular hardness (*Hard*) in this particular case were used for simple numerical description of the chlorine substitution pattern rather than the electron transfer processes (direct link to *HOMO*). As it was proved in our previous studies [2], *beta*-substituted chloronaphthalene congeners are

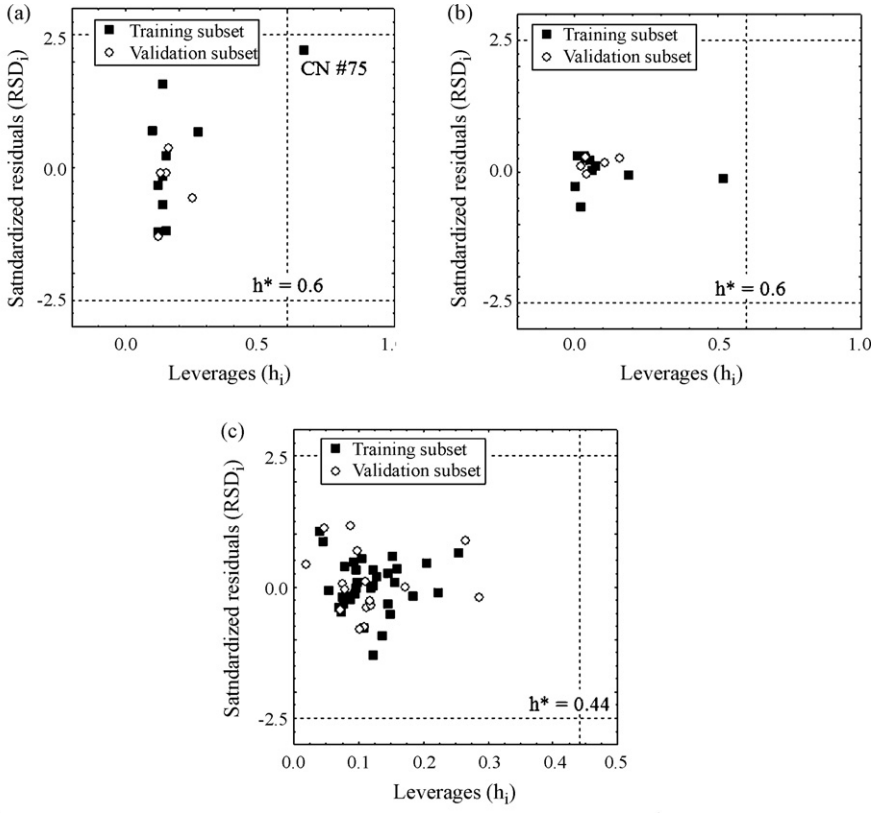

Fig. 1. The Williams plots describing the applicability domains of the models developed in this study: (a) MLR model for Set 1; (b) PLS model for Set 1; and (c) PLS model for Set 2.

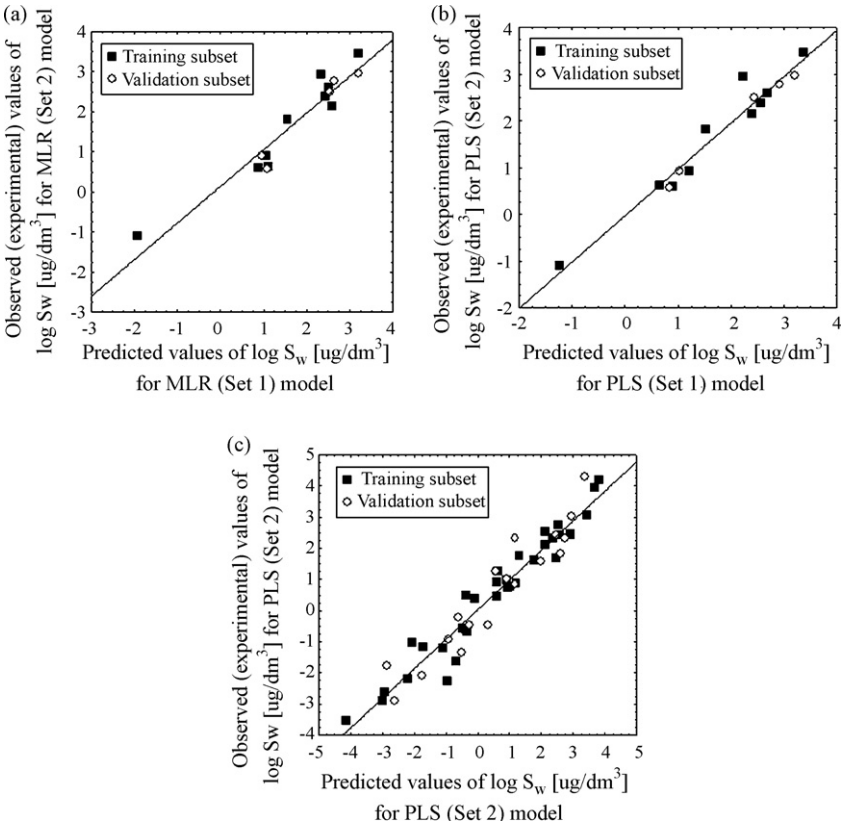

Fig. 2. The correlations between experimental (observed) and predicted values of water solubility for the models developed in this study: (a) MLR model for Set 1; (b) PLS model for Set 1; and (c) PLS model for Set 2.

Table 4
Experimental and predicted values of water solubility (S) for chloronaphthalene congeners (μg/dm^3).

#CN	CN congener	Exp. log S[a]	Predicted log S				
			WSKOWIN	COSMO-RS	QSPR[b]		
					MLR (Set 1)	PLS (Set 1)	PLS (Set 2)
1	1-Chloronaphthalene	3.46[T]	4.49	4.10	3.22	3.38	3.07
2	2-Chloronaphthalene	2.97[V]	4.42	4.19	3.20	3.21	3.02
3	1,2-Dichloronaphthalene	2.14[T]	3.88	3.58	2.61	2.40	2.53
4	1,3-Dichloronaphthalene				2.45	2.28	2.41
5	1,4-Dichloronaphthalene	2.50[V]	3.68	3.39	2.55	2.45	2.44
6	1,5-Dichloronaphthalene	2.60[T]	3.67	3.40	2.53	2.71	2.42
7	1,6-Dichloronaphthalene				2.45	2.54	2.36
8	1,7-Dichloronaphthalene	2.37[T]	3.76	3.48	2.44	2.58	2.33
9	1,8-Dichloronaphthalene	2.77[V]	3.76	3.79	2.64	2.93	2.32
10	2,3-Dichloronaphthalene	2.94[T]	3.80	3.69	2.35	2.25	2.45
11	2,6-Dichloronaphthalene				2.35	2.43	2.23
12	2,7-Dichloronaphthalene				2.34	2.35	2.33
13	1,2,3-Trichloronaphthalene				1.82	1.47	1.94
14	1,2,4-Trichloronaphthalene				1.81	1.49	1.89
15	1,2,5-Trichloronaphthalene				1.81	1.74	1.86
16	1,2,6-Trichloronaphthalene				1.71	1.61	1.75
17	1,2,7-Trichloronaphthalene				1.72	1.59	1.80
18	1,2,8-Trichloronaphthalene				1.91	1.99	1.77
19	1,3,5-Trichloronaphthalene				1.71	1.64	1.73
20	1,3,6-Trichloronaphthalene				1.61	1.44	1.71
21	1,3,7-Trichloronaphthalene	1.81[T]	2.91	2.76	1.56	1.52	1.61
22	1,3,8-Trichloronaphthalene				1.79	1.83	1.67
23	1,4,5-Trichloronaphthalene				1.90	2.00	1.73
24	1,4,6-Trichloronaphthalene				1.71	1.66	1.69
25	1,6,7-Trichloronaphthalene				1.70	1.90	1.67
26	2,3,6-Trichloronaphthalene				1.60	1.45	1.69
27	1,2,3,4-Tetrachloronaphthalene	0.62[T]	2.37	2.18	1.10	0.64	1.26
28	1,2,3,5-Tetrachloronaphthalene	0.57[V]	2.36	2.12	1.08	0.84	1.25
29	1,2,3,6-Tetrachloronaphthalene				0.97	0.66	1.19
30	1,2,3,7-Tetrachloronaphthalene				0.98	0.70	1.15
31	1,2,3,8-Tetrachloronaphthalene				1.18	1.06	1.18
32	1,2,4,5-Tetrachloronaphthalene				1.16	1.04	1.17
33	1,2,4,6-Tetrachloronaphthalene				0.97	0.73	1.10
34	1,2,4,7-Tetrachloronaphthalene				0.97	0.69	1.15
35	1,2,4,8-Tetrachloronaphthalene				1.18	1.06	1.16
36	1,2,5,6-Tetrachloronaphthalene				1.07	1.08	1.20
37	1,2,5,7-Tetrachloronaphthalene				0.97	0.95	1.12
38	1,2,5,8-Tetrachloronaphthalene				1.18	1.33	1.09
39	1,2,6,7-Tetrachloronaphthalene				0.96	0.95	1.10
40	1,2,6,8-Tetrachloronaphthalene				1.06	1.19	1.01
41	1,2,7,8-Tetrachloronaphthalene				1.19	1.31	1.16
42	1,3,5,7-Tetrachloronaphthalene	0.60[T]	2.00	1.73	0.88	0.90	0.92
43	1,3,5,8-Tetrachloronaphthalene	0.91[T]	2.37	2.18	1.05	1.21	0.98
44	1,3,6,7-Tetrachloronaphthalene				0.86	0.84	0.97
45	1,3,6,8-Tetrachloronaphthalene				0.94	1.01	0.95
46	1,4,5,8-Tetrachloronaphthalene				1.26	1.54	1.00
47	1,4,6,7-Tetrachloronaphthalene	0.91[V]	2.32	2.02	0.96	1.02	1.01
48	2,3,6,7-Tetrachloronaphthalene				0.86	0.79	1.01
49	1,2,3,4,5-Pentachloronaphthalene				0.57	0.27	0.70
50	1,2,3,4,6-Pentachloronaphthalene				0.35	−0.07	0.64
51	1,2,3,5,6-Pentachloronaphthalene				0.34	0.17	0.60
52	1,2,3,5,7-Pentachloronaphthalene				0.24	0.09	0.45
53	1,2,3,5,8-Pentachloronaphthalene				0.44	0.43	0.48
54	1,2,3,6,7-Pentachloronaphthalene				0.23	0.04	0.48
55	1,2,3,6,8-Pentachloronaphthalene				0.33	0.26	0.43
56	1,2,3,7,8-Pentachloronaphthalene				0.46	0.41	0.53
57	1,2,4,5,6-Pentachloronaphthalene				0.44	0.39	0.51
58	1,2,4,5,7-Pentachloronaphthalene				0.32	0.25	0.42
59	1,2,4,5,8-Pentachloronaphthalene				0.53	0.61	0.44
60	1,2,4,6,7-Pentachloronaphthalene				0.22	0.08	0.42
61	1,2,4,6,8-Pentachloronaphthalene				0.33	0.29	0.38
62	1,2,4,7,8-Pentachloronaphthalene				0.46	0.41	0.54
63	1,2,3,4,5,6-Heksachloronaphthalene				−0.15	−0.37	0.05
64	1,2,3,4,5,7-Heksachloronaphthalene				−0.28	−0.50	−0.08
65	1,2,3,4,5,8-Heksachloronaphthalene				−0.06	−0.14	−0.03
66	1,2,3,4,6,7-Heksachloronaphthalene				−0.40	−0.70	−0.06
67	1,2,3,5,6,7-Heksachloronaphthalene				−0.40	−0.43	−0.11
68	1,2,3,5,6,8-Heksachloronaphthalene				−0.30	−0.23	−0.16
69	1,2,3,5,7,8-Heksachloronaphthalene				−0.28	−0.20	−0.17
70	1,2,3,6,7,8-Heksachloronaphthalene				−0.28	−0.23	−0.15
71	1,2,4,5,6,8-Heksachloronaphthalene				−0.19	−0.03	−0.20

Table 4 (*Continued*)

#CN	CN congener	Exp. log S[a]	Predicted log S				
			WSKOWIN	COSMO-RS	QSPR[b]		
					MLR (Set 1)	PLS (Set 1)	PLS (Set 2)
72	1,2,4,5,7,8-Heksachloronaphthalene				−0.18	−0.03	−0.18
73	1,2,3,4,5,6,7-Heptachloronaphthalene				−0.89	−0.98	−0.65
74	1,2,3,4,5,6,8-Heptachloronaphthalene				−0.77	−0.78	−0.67
75	1,2,3,4,5,6,7,8-Oktachloronaphthalene	-1.10^{T}	0.98	0.05	−1.94	−1.24	−1.21

[T]Training subset, [V]Validation subset.

[a] Experimental data taken from [6].

[b] This study.

characterized by higher values of ionization potential (negative value of *HOMO*) that the other compounds containing the same number of chlorine atoms. Summarizing, the predictive ability (of the PLS model) increased when we included descriptors not only related to the cavitation, but also those involving the other types of the solute–solvent interactions. However, the influence of the cavitation is more vital than the influence of the other phenomena. Therefore, taking into account the rule of making the algorithm as simple as possible, we could recommend the simplest model (MLR), which uses only one descriptor for further predictions.

It is interesting that, in this particular case, the values of RMSEP are lower than RMSECV. This unusual situation probably resulted from very limited size of the validation set. However, we decided to use even a small number of validation compounds in order to externally confirm predictive ability of the models.

In this study we initially assumed that development of the QSPR models with use of the small training set of similar compounds should be more accurate for congeners that application of the models with wider applicability domains. This assumption has been confirmed by the results of comparison between both PLS models: the model for Set 1 and Set 2. As we expected, the predictions by the model developed with larger and more diversified structural domain (Set 2) are noticeable worse that the predictions by the model calibrated only on PCNs (Set 1). When designing the QSPRs for congeneric sets, there is often a dilemma, which strategy of modeling to choose. Indeed, is it better to have a model based on only 10 congeners, even if the ratio of descriptors-to-compounds is relatively small or is it better to have a model calibrated on greater number, but not of such structurally similar compounds? As usually, there is no one unambiguous answer. Everything should depend on the purpose of modeling. When the precision of the results is very important, we recommend to reduce the model's domain and to make predictions based on the small set of congeners. However, when the model should be predictive not only for one class of compounds (i.e., chloronaphthalenes), but also for other species (i.e., dibenzofurans, dibenzo-*p*-dioxins) the second strategy must be applied. Obviously, the second strategy is the only one possibility, when the number of congeners from one class with available experimental data is insufficient to construct even a simple predictive model.

An association of particular congeners from the training and validation subsets with the models' applicability domains was confirmed by use of the Williams plots (Fig. 1). We have not observed the presence of outliers. But, when comparing both models for Set 1, the high influence of CN#75 on the MLR model can be noticed. For all the models also a strong linear correlation between the experimentally obtained and the predicted values of water solubility was observed (Fig. 2). In this way the OECD recommendations of the QSAR's quality have been fulfilled.

3.3. Water solubility of the individual CN congeners

Before the prediction of water solubility for the remaining 60 congeners, we verified if they were situated inside of the applicability domains of the three models by calculating the leverage values (Fig. 3). Even the most influencing congeners (CN nos. 73 and 74) had the leverage values significantly lower than the critical thresholds.

Based on the results obtained for PCNs (Table 4) some environmentally related conclusions can be made. Persistent Organic Pollutants (POPs) contaminating the natural environment may exist in aquatic ecosystems in different forms—dissolved in water, absorbed on various organic particles or accumulated in tissues of numerous aquatic species. Lower chlorinated CN congeners show the highest water solubility and this fact influences their fate, mobility and physical form in the aqueous environment. When comparing the results to the other POPs we hypothesized that mono- and dichloronaphthalenes, due to their relatively high water solubility, should be generally present in the pelagial water, while tri- and tetra-chloronaphthalenes could exist in the pelagial water as long as they can be absorbed on suspended organic particles. They might also be bioaccumulated due to their high lipophilicity [3]. Penta- and hexa-chloronaphthalenes are much less soluble in water and much more lipophilic [3]. Thus, they should be mainly accumulated in biota. Chloronaphthalenes containing seven or eight chlorine atoms are almost insoluble in water and hardly bioaccumulated; they would be deposited in sediments or absorbed on organic particles. Such hypothesis should be further verified by measuring concentration of the congeners in the compartments (water, organisms, and sediments) of water ecosystems (lakes, rivers, sea). It is also worth noting that the congeneric compounds, even containing the same number of chlorine atoms in the molecule, can differ by ability to be transported in the aqueous systems and bioaccumulated by aquatic organisms [4]. Because of that estimating environmental risk of a congeneric set of pollutants it is usually very important to have data on solubility determined for all possible congeners.

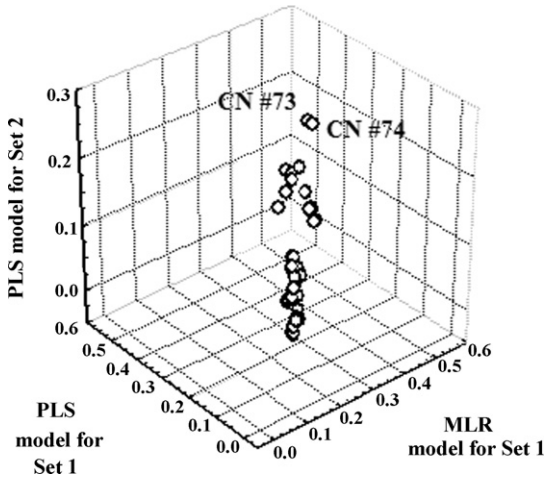

Fig. 3. The plot of the leverage values for chloronaphthalenes neither included in the training nor in the validation subsets.

3.4. Reference calculations and comparisons

Many previously published QSPR models of water solubility were trained on the large sets of structurally differentiated compounds. Those strategies usually lead to the models, characterized by wide applicability domains. A comprehensive comparison of the models could be found in a review by Delaney [32]. Ten QSPR models discussed by the author were obtained using different chemometric methods, including MLR and artificial neural networks techniques. The number of the model parameters varied between 3 and 118 including a wide spectrum of the two-dimensional and three-dimensional molecular descriptors. The values of the standard error of prediction for these models calculated for 21 common chemicals (validation set) ranged between 0.55 and 0.91 log units.

Similar or even better predictive ability was reported for the WSKOWIN program. We used this model as a 'golden standard' for the reference calculations. We calculated the values of water solubility for those 15 congeners, for which experimental data exist (Table 4). To be as much fair as possible in the comparison, we split these compounds into two groups: a group corresponding to the training subset used in the QSPR models for Set 1 ($n = 10$) and a group corresponding to the external validation subset ($n_{val} = 5$). Then, we calculated the root mean square errors of prediction separately for these two groups. They were $RMSEP_{10}{}^{WSKOWIN} = 1.44$ and $RMSEP_5{}^{WSKOWIN} = 1.40$ (log units), respectively.

In the same way, we performed additional reference calculations with use of the COSMO-RS model. The values of the root mean square errors of prediction by COSMO-RS (for $n = 5$ and $n = 10$) were $RMSEP_5{}^{COSMO\text{-}RS} = 1.78$ and $RMSEP_{10}{}^{COSMO\text{-}RS} = 1.12$ (log units).

When we put together the values of $RMSEP_5{}^{WSKOWIN}$, $RMSEP_5{}^{COSMO\text{-}RS}$ and the RMSEPs for the models presented in this study, we observed up to two times better predictive ability of the QSPRs in comparison to the WSKOWIN and COSMO-RS models. Moreover, because only data for 5 compounds were used, we decided to compare also the root mean square errors of cross-validated data (RMSECV) to the $RMSEP_{10}{}^{WSKOWIN}$. In that case also the QSPR models were characterized by significantly lower errors. However, the second comparison should be treated with care, since, in fact, we put together data from external ($RMSEP_{10}{}^{WSKOWIN}$) and internal (RMSECV) validation. Although this did not provide a direct comparison between the values predicted from WSKOWIN and COSMO-RS with the fitted values from QSPR (but the cross-validated residuals), the external validation is always a stronger criterion than the internal predictive ability measured by the cross-validated residuals.

In our study we tried to restrict the applicability domain of the models as much as possible. By decreasing 'universalism' of the models it was possible to improve their predictive characteristics. However, it is worth noting that even the PLS model for the extended domain (Set 2) predicts water solubility of PCNs more precisely than any of the standard models mentioned above. This observation does not mean that the 'golden standard' models are wrong. We have no enough data to make such a conclusion. On contrary, this excellent models are very useful nowadays, when we have roughly 100 000 different chemicals that are commercially produced and for which we are in the need of a sound risk assessment procedure. However, we have demonstrated that whenever even a very limited data set is available for a congeneric family (such as for PCNs) and whenever very reliable data are needed, development of the local QSPRs instead of using the 'golden standards' is justified.

4. Conclusions

This study provides the values of water solubility predicted for 75 chloronaphthalene congeners by the local QSPR models. The values calculated from the most recommended, simplest MLR model (for Set 1) varied from 0.01 to 1660 $\mu g\,dm^{-3}$ (2.85×10^{-11}–1.02×10^{-5} $mol\,dm^{-3}$), depending on the number of chlorine atoms present in the molecule and the substitution pattern. In this way, significant gaps in the environmentally relevant physicochemical data on PCNs are now bridged. In further works it will be possible to use this data for modeling of environmental transport and fate not only of selected congeners, but also for all chloronaphthalenes.

The molecular parameters calculated at the Density Functional Theory level with the 6-311++G(d,p) basis set were able to successfully describe differences in the estimated property, even among very similar compounds (congeners). This was impossible, when only semi-empirical descriptors were used. In addition, due to low cost of DFT computations in comparison to such 'classic' *ab inito* methods with comparable accuracy, as MP2, the presented approach is strongly recommended.

The errors of prediction by the presented local QSPRs (for Set 1) were significantly lower than those of more general QSPRs (for Set 2) as well as of WSKOWIN and COSMO-RS 'golden standard' models. Although there were not enough experimental data for more sophisticated comparisons and general conclusions, it was demonstrated that whenever even a very limited data set is available and more reliable values of properties are required the development of such local QSPRs is justified.

Acknowledgements

The authors thank to Prof. Paola Gramatica, Prof. Janusz Rak and the three anonymous reviewers for their stimulating comments. The support by the Basic Research Project BT25-08-41 grant from the U.S. Army Engineer Research and Development Center (ERDC) is acknowledged. Computations were carried out using supercomputers in TASK—Academic Computer Center in Gdańsk. T.P. thanks to U.S. Environmental Protection Agency for the training devoted to EPISuite software. T.P. is the recipient of the "HOMING" fellowship granted by the Foundation for the Polish Science and financed by the EOG Financial Mechanism in Poland.

References

[1] S.N. Bhattachar, L.A. Deschenes, J.A. Wesley, Solubility: it's not just for physical chemists, Drug Discov. Today 11 (2006) 1012–1018.
[2] T. Puzyn, J. Falandysz, P.D. Jones, J.P. Giesy, Quantitative structure–activity relationships for prediction of relative in vitro potencies (RePs) for chloronaphthalenes, J. Environ. Sci. Technol. A 42 (2007) 1–18.
[3] T. Puzyn, J. Falandysz, QSPR modeling of partition coefficients and Henry's Law constants for 75 chloronaphthalene congeners by means of six chemometric approaches—a comparative study, J. Phys. Chem. Ref. Data 36 (2007) 203–214.
[4] J. Falandysz, Chloronaphthalenes as food-chain contaminants: a review, Food Addit. Contam. 20 (2003) 995–1014.
[5] S.C. Lee, T. Harner, K. Pozo, M. Shoeib, F. Wania, D.C. Muir, L.A. Barrie, K.C. Jones, Polychlorinated naphthalenes in the Global Atmospheric Passive Sampling (GAPS) study, Environ. Sci. Technol. 41 (2007) 2680–2687.
[6] A. Opperhuizen, E.W. van der Velde, F.A.P.C. Gobas, D.A.K. Liem, J.M.D. van der Steen, Relationships between bioconcentration in fish and steric factors of hydrophobic chemicals, Chemosphere 14 (1985) 1871–1896.
[7] Chlorinated naphtalenes, World Health Organisation, International Program on Chemical Safety, Geneva, 2001.
[8] Polychlorinated naphtalenes, National Industrial Chemicals Notification and Assessment Scheme, Sydney, 2002.
[9] A. Lewis, N. Kazantzis, I. Fishtik, J. Wilcox, Integrating process safety with molecular modeling-based risk assessment of chemicals within the REACH regulatory framework: benefits and future challenges, J. Hazard. Mater. 142 (2007) 592–602.
[10] Y. Pan, J. Jiang, R. Wang, H. Cao, J. Zhao, Prediction of auto-ignition temperatures of hydrocarbons by neural network based on atom-type electrotopological-state indices, J. Hazard. Mater. 157 (2008) 510–517.
[11] M.H. Keshavarz, Prediction of heats of sublimation of nitroaromatic compounds via their molecular structure, J. Hazard. Mater. 151 (2008) 499–506.
[12] P. Gramatica, E. Papa, Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure, Environ. Sci. Technol. 41 (2007) 2833–2839.

[13] M. Haranczyk, T. Puzyn, P. Sadowski, ConGENER—a tool for modeling of the congeneric sets of environmental pollutants, QSAR Comb. Sci. 27 (2008) 826–833.
[14] C.W. Lee, W. Yang, R.G. Parr, Development of the Colle–Salvetti correlation energy formula into a functional of the electron density, Phys. Rev. B 37 (1988) 785–789.
[15] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange, J. Chem. Phys. 98 (1993) 5648–5652.
[16] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery, T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, Y. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez and J.A. Pople, GAUSSIAN 03, Gaussian Inc., 2003.
[17] F. Jensen, Introduction to Computational Chemistry, John Wiley & Sons, Chichester, 1999.
[18] J. Tomasi, M. Persico, Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent, Chem. Rev. 94 (1994) 2027–2094.
[19] A. Eucken, K.H. Hellwege, Landolt-Börnstein Zahlenwerte und Funktionen aus Physic, Chemie, Astronomie, Geophysik, Technik, Springer-Verlag, Berlin, 1951.
[20] D. Mackay, W.Y. Shiu, K.C. Ma, S.C. Lee, Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals, 2nd ed., Taylor & Francis, Boca Raton, London, New York, 2007.
[21] M.A. Sharaf, D.H. Illman, B.R. Kowalski, Chemometrics, John Wiley & Sons Inc., 1986.
[22] D.J. Livingstone, D.W. Salt, Judging the significance of multiple linear regression models, J. Med. Chem. 48 (2005) 661–663.
[23] PLS Toolbox Version 4.1, 2007, Eigenvector Research, http://www.eigenvector.com (accessed January 2008).
[24] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007), 694-670.
[25] OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure–Activity Relationships models, 2004, Available at: http://www.oecd.org/dataoecd/33/37/37849783.pdf (accessed January 2008).
[26] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for K-oc prediction, J. Mol. Graph. Modell. 25 (2007) 755–766.
[27] MATLAB Version 7.6, Mathworks, 2008, http://www.mathworks.com (accessed January 2008).
[28] WSKOWIN, WSKOWIN, v. 1.41, U.S. Environmental Protection Agency, 2000.
[29] F. Eckert, A. Klamt, COSMO*therm*, Version C2.1, Release 01.05, COSMO*logic* GmbH & Co. KG, 2005.
[30] F. Eckert, A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, AIChE J. 48 (2002) 369–385.
[31] W.M. Meylan, P.H. Howard, R.S. Boethling, Improved method for estimating water solubility from octanol/water partition coefficient, Environ. Toxicol. Chem. 15 (1996) 100–106.
[32] J.S. Delaney, Predicting aqueous solubility from structure, Drug Discov. Today 10 (2005) 289–295.