JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Combination of RISM and Cheminformatics for Efficient Predictions of Hydration Free Energy of Polyfragment Molecules: Application to a Set of Organic Pollutants

Ekaterina L. Ratkova and Maxim V. Fedorov*

The Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, Leipzig, 04103, Germany

Supporting Information

ABSTRACT: Here, we discuss a new method for predicting the hydration free energy (HFE) of organic pollutants and illustrate the efficiency of the method on a set of 220 chlorinated aromatic hydrocarbons. The new model is computationally inexpensive, with one HFE calculation taking less than a minute on a PC. The method is based on a combination of a molecular integral equations theory, one-dimensional reference interaction site model (1D RISM), with the cheminformatics approach. We correct HFEs obtained by the 1D RISM with a set of empirical corrections. The corrections are associated with the partial molar volume and structural descriptors of the molecules. We show that the introduced corrections can significantly improve the quality of the 1D RISM HFE predictions obtained by the partial wave free energy expression [Ten-no, S. J. *Chem. Phys.* **2001**, *115*, 3724] and the Kovalenko−Hirata closure [Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095]. We also show that the quality of the model can be further improved by the reparametrization using QM-derived partial charges instead of the originally used OPLS-AA partial charges. The final model gives good results for polychlorinated benzenes (the mean and standard deviation of the error are 0.02 and 0.36 kcal/mol, correspondingly). At the same time, the model gives somewhat worse results for polychlorobiphenyls (PCBs) with a systematic bias of −0.72 kcal/mol but a small standard deviation equal to 0.55 kcal/mol. We note that the error remains the same for the whole set of PCBs, whereas errors of HFEs predicted with continuum solvation models (data were taken from Phillips, K. L. et al. *Environ. Sci. Technol.* **2008**, *42*, 8412) increase significantly for higher chlorinated PCB congeners. In conclusion, we discuss potential future applications of the model and several avenues for its further improvement.

INTRODUCTION

Chlorinated aromatic hydrocarbons (CAHs) are a group of compounds that belong to the category of "persistent organic pollutants" (POPs). This class of pollutants is characterized by (i) long-term persistence, (ii) long-range atmospheric transport and deposition, (iii) bioaccumulation, and (iv) adverse effects on biota.[1,2] For a long time, in many countries, CAHs (such as polychlorobiphenyls, hexachlorobenzene, etc.) were used in agriculture as pesticides, fungicides, and agents controlling arthropods.[1] Although CAHs have been banned from further use and production,[3] their persistence in biological compartments (e.g., soil, water, plants, and sediment) means that they still pose a significant environmental hazard. Understanding and clarifying the global fate of CAHs is one of the most important environmental and ecological problems.[1,2,4,5] The semivolatile nature of CAHs allows them to evaporate from soil and water into the atmosphere, where they can exist both in gaseous and particle-absorbed forms (these can be atmospheric aerosol particles, e.g., cloud droplets, as well as dust particles). Several dominant mechanisms that determine the distribution of CAHs between atmosphere and water are shown in Figure 1.

There are several physical/chemical properties of CAHs that determine their global fate: vapor pressure; aqueous solubility; partition coefficients between different media; and half-lives in the air, solids, and water. These parameters are intensively used in mathematical models describing the global fate and long-range transport of CAHs.[6−10] One of the most important parameters in these models is the flux of a compound across surfaces, which characterizes the exchange of the compound between the corresponding compartments.[2,11] As an example, the flux of molecules i between two compartments 1 and 2 can be modeled as

$$F_{1\rightarrow 2} = K_{1/2(i)}\left(C_{1(i)} - \frac{C_{2(i)}}{P_{i,\mathrm{eq}}}\right) \quad (1)$$

where $F_{1\rightarrow 2}$ is the flux from compartment 1 to compartment 2, $K_{1/2(i)}$ is the kinetic parameter represented by the mass transfer coefficient on the molecules i, $C_{1(i)}$ and $C_{2(i)}$ are equilibrium molecular concentrations of the molecules i in compartments 1 and 2, respectively, and $P_{i,\mathrm{eq}}$ is the equilibrium partition coefficient of the molecules i between the two compartments.

Thus, accurate data for the partition coefficients are of a high importance for modeling CAH exchange between compartments. In the case of the air−water flux, the widely used partition coefficient is the Henry's law constant (K_{H}), which shows the distribution of a compound between gaseous and aqueous phases:

$$K_{\mathrm{H}} = \frac{C_{\mathrm{aq}(i)}}{C_{\mathrm{g}}(i)} \quad (2)$$

Received: November 12, 2010
Published: April 22, 2011

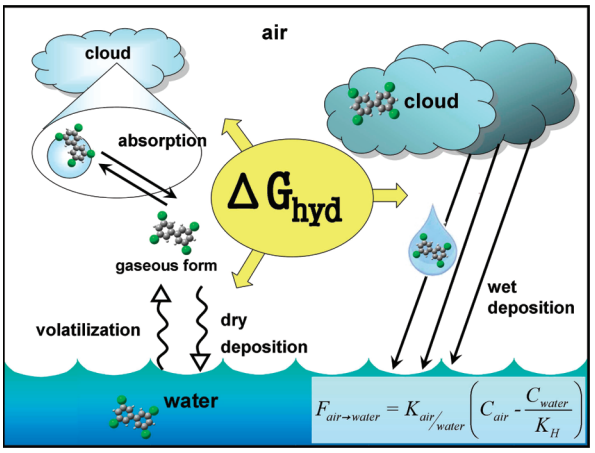

Figure 1. Dominant mechanisms that determine the distribution of CAHs between atmosphere and water. Hydration free energy (ΔG_{hyd}) is an important thermodynamic parameter used to describe the main processes of CAH distribution between atmosphere and water. It is closely related to the Henry's law constant (K_H) as $\Delta G_{hyd} = -RT \ln(K_H)$. In turn, K_H is widely used to model the flux of a molecule from air to water, $F_{air\to water}$ (see the inset equation and eq 1 for the notation).

where $C_{aq(i)}$ and $C_{g(i)}$ are equilibrium molecular concentrations of the molecules i in aqueous and gaseous phases, respectively.

We note that the Henry's law constant is closely related with the HFE as[12]

$$\Delta G_{hyd} = -RT \ln(K_H) \quad (3)$$

where ΔG_{hyd} is the hydration free energy, K_H is the Henry's law constant, R is the ideal gas constant, and T is the temperature.

Recently, we reported a novel computational method for accurate estimations of the HFEs of organic molecules—the structural descriptors correction (SDC) model.[13] The method is based on a combination of the computationally inexpensive one-dimensional RISM (1D RISM) with several corrections that can be obtained in a straightforward manner from the molecular structure. The main advantage of the model is a small number of chemical descriptors associated with main structural features of solutes: partial molar volume (PMV), aromatic rings, electron-donating/withdrawing substituents, etc. We have shown that the 1D RISM-SDC model with the OPLS-AA partial charges[14,15]—1D RISM-SDC(OPLSq) model—allows one to obtain HFEs for monofragment solutes with high accuracy.[13] In the case of polyfragment solutes, the 1D RISM-SDC(OPLSq) model is more sensitive to the chemical nature of solutes. Thus, the model allows one to predict HFEs with an accuracy of about 1 kcal/mol for chlorinated benzenes with fewer than three chlorine atoms, but it provides worse results for chlorinated benzenes with a larger number of chlorine atoms.[13] The main reason for this deviation is the fact that OPLS-AA partial charges are not sensitive to the mesomeric effect of aromatic polyfragment solutes.[13]

Here, we show that the quality of the 1D RISM-SDC model can be further improved by the model reparametrization using QM-derived partial charges (1D RISM-SDC(QMq) model) instead of the originally used OPLS-AA partial charges. In this paper, we would like to demonstrate the efficiency of the 1D RISM-SDC(QMq) model for two classes of CAHs: (i) polychlorinated benzenes and (ii) polychlorobiphenyls. Other classes of CAHs will be considered in our forthcoming publications.

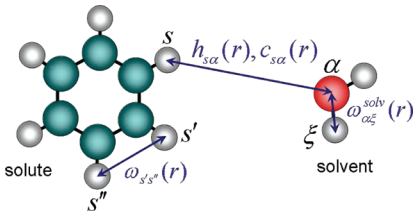

Figure 2. Representations of solute and solvent molecules and correlation functions in the 1D RISM approach. Both molecules are modeled as sets of sites (atoms). The molecules structures are described with site−site intramolecular correlation functions: $\omega_{ss'}(r)$ and $\omega^{solv}_{\alpha\xi}(r)$. Solvent density distributions around the solute molecule are described with intermolecular total $h_{s\alpha}(r)$ and direct $c_{s\alpha}(r)$ correlation functions.

METHODS

1D RISM Approach. We use here the 1D RISM approach,[16] where the solute and solvent molecules are modeled as sets of sites (atoms) interacting via pairwise spherically symmetric potentials (Figure 2).[16] We use the common form of the interaction potential represented by the long-range electrostatic term and short-range Lennard-Jones (LJ) term.[17] The 1D RISM operates with site−site correlation functions: intramolecular correlation functions $\omega_{ss'}(\mathbf{r})$, $\omega^{solv}_{\alpha\xi}(\mathbf{r})$, total correlation functions $h_{s\alpha}(\mathbf{r})$, and direct correlation functions $c_{s\alpha}(\mathbf{r})$ (where s and s' are solute atoms, and α and ξ are solvent atoms; Figure 2).[16] In general, these are 3D-functions, but due to the spherical symmetry, we consider only their 1D-radial parts $\omega_{ss'}(r)$, $\omega^{solv}_{\alpha\xi}(r)$, $h_{s\alpha}(r)$, and $c_{s\alpha}(r)$, which depend only on the radial distance r. Direct correlation functions are connected with the total correlation functions via the set of 1D RISM integral equations:[16]

$$h_{s\alpha}(|\mathbf{r}_1 - \mathbf{r}_2|) = \sum_{s'=1}^{N_{solute}} \sum_{\xi=1}^{N_{solvent}} \int_{R^3}\int_{R^3} \omega_{ss'}(|\mathbf{r}_1 - \mathbf{r}'|)\, c_{s'\xi}(|\mathbf{r}' - \mathbf{r}''|)\, \chi_{\alpha\xi}(|\mathbf{r}'' - \mathbf{r}_2|)\, d\mathbf{r}'\, d\mathbf{r}'' \quad (4)$$

where $\chi_{\alpha\xi}(r) = \omega^{solv}_{\alpha\xi}(r) + \rho h^{solv}_{\alpha\xi}(r)$ are the bulk solvent susceptibility functions and N_{solute} and $N_{solvent}$ are the numbers of sites in the solute and solvent, correspondingly. We note that in the current work we expressed the intramolecular correlation functions in terms of Dirac δ functions considering molecules as rigid objects. However, molecules under investigation are almost rigid, and this simplification does not lead to considerable changes in the description of the hydration process. In general, for more flexible compounds, the changes in molecular conformations upon hydration have to be taken into account (e.g., with a coupled RISM/MD or RISM/MC simulation methodology[18−20]).

To make eq 4 complete, $N_{solute} \times N_{solvent}$ site−site *closure* relations are introduced:

$$h_{s\alpha}(r) = \exp(-\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r) + B_{s\alpha}(r)) - 1$$
$$s = 1, \ldots, N_{solute}; \alpha = 1, \ldots, N_{solvent} \quad (5)$$

where $u_{s\alpha}(r)$ is a pair interaction potential between the sites s and α, $B_{s\alpha}(r)$ are site−site bridge functions, and $\beta = 1/k_B T$, where k_B is the Boltzmann constant and T is the temperature. In general, the exact bridge functions are practically uncomputable, and one needs to use some approximation.[16,21,22] The most straightforward and widely used model is the HNC approximation, which

sets the bridge functional $B_{s\alpha}(r)$ to zero.[23] However, due to the uncontrolled growth of the argument of the exponent (eq 5), use of the HNC closure can lead to a slow convergence rate, and in many cases even divergence of the numerical solution of 1D RISM equations. One way to overcome this problem is to linearize the exponent when its argument is larger than a certain threshold constant C:

$$h_{s\alpha}(r) = \begin{cases} \exp(\Xi_{s\alpha}(r)) - 1 & \text{when} \quad \Xi_{s\alpha}(r) < C \\ \Xi_{s\alpha}(r) + \exp(C) - C - 1 & \text{when} \quad \Xi_{s\alpha}(r) > C \end{cases} \tag{6}$$

where $\Xi_{s\alpha}(r) = -\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r)$.

The linearized HNC closure for the case $C = 0$ was proposed by Hirata and Kovalenko in ref 24, where it has been called KH closure. More details of the theoretical and computational background behind the 1D RISM equations can be found in refs 16, 25, and 26. In the current work, we performed 1D RISM calculations with the KH closure. Previously, we showed[27,28] that the efficiency of HFE calculations with a set of semiempirical corrections is almost not sensitive to the choice of closure relation (KH, HNC). However, the KH closure allows one to perform the quickest and the most stable RISM calculations. That is why we used the KH closure in this work rather than the HNC.

Within the framework of the 1D RISM theory, there are several free energy expressions which allow one to obtain values of the HFE from the total and direct correlation functions: HNC,[16,23] GF,[29] KH,[30] PW,[31] HNCB,[32] and PWC.[33] Comparisons of these expressions[13,31,33−38] show that the PW free energy expression has better performance than the KH, HNC, and HNCB free energy expressions. Therefore, as in our previous work,[13] we use here the PW free energy expression to calculate HFE values:

$$\Delta G_{\text{hyd}}^{\text{PW}} = 2\pi\rho k_{\text{B}}T \sum_{s=1}^{N_{\text{solute}}} \sum_{\alpha=1}^{N_{\text{solvent}}} \int_0^\infty [-2c_{s\alpha}(r) - c_{s\alpha}(r)\, h_{s\alpha}(r) + \tilde{h}_{s\alpha}(r)\, h_{s\alpha}(r)] r^2\, \mathrm{d}r \tag{7}$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$ and

$$\tilde{h}_{s\alpha}(|\mathbf{r}_1 - \mathbf{r}_2|) = \sum_{s'=1}^{N_{\text{solute}}} \sum_{\xi=1}^{N_{\text{solvent}}} \int_{R^3}\int_{R^3} \tilde{\omega}_{ss'}(|\mathbf{r}_1 - \mathbf{r}'|)\, h_{s'\xi}(|\mathbf{r}' - \mathbf{r}''|)\, \tilde{\omega}_{\alpha\xi}^{\text{solv}}(|\mathbf{r}'' - \mathbf{r}_2|)\, \mathrm{d}\mathbf{r}'\, \mathrm{d}\mathbf{r}''$$

$\tilde{\omega}_{ss'}$ and $\tilde{\omega}_{\alpha\xi}^{\text{solv}}(r)$ are the elements of matrices $\mathbf{W}^{-1}$ and $\mathbf{W}_{\text{solv}}^{-1}$, which are inverses to the matrices $\mathbf{W} = [\omega_{ss'}(r)]_{N_{\text{solute}}\times N_{\text{solute}}}$ and $\mathbf{W}_{\text{solv}} = [\omega_{\alpha\xi}^{\text{solv}}(r)]_{N_{\text{solvent}}\times N_{\text{solvent}}}$ built from the solute and solvent intramolecular correlation functions, respectively.

SDC Model. We define the *modeling error* (ε) of the RISM-based HFEs calculations for a solute as the difference between the calculated and experimental values:

$$\varepsilon = \Delta G_{\text{hyd}}^{\text{model}} - \Delta G_{\text{hyd}}^{\text{exp}} \tag{8}$$

where $\Delta G_{\text{hyd}}^{\text{exp}}$ is the experimental value of HFE and $\Delta G_{\text{hyd}}^{\text{model}}$ is the HFE calculated by the RISM approach (superscript model denotes the RISM-based HFE expression, e.g., PW).

The main idea behind the SDC model is that we parametrize the modeling error ε with a set of descriptors $\{D_i\}$ associated with specific features of the chemical structure of solutes such as partial molar volume (PMV), aromatic rings, electron-donating/withdrawing substituents, etc. (Figure 3). The model contains

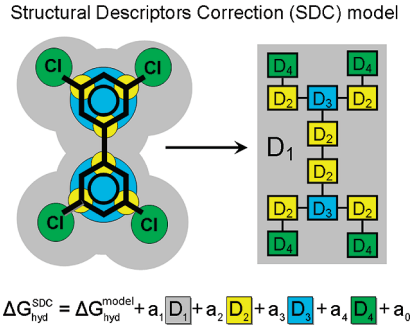

Figure 3. Schematic representation of a molecule (3,3′,5,5′-tetrachlorobiphenyl) as a combination of fragment counts. The SDC model equation as a linear combination of the corresponding structural corrections: a_1D_1 is the correction on dimensionless partial molar volume, a_2D_2 is the correction on branches, a_3D_3 is the correction on the benzene ring, a_4D_4 is the correction on the halogen atom, and a_0 is a constant (solute-independent systematic error; see eq 9).

the assumption that different structural properties of the solute molecule contribute independently to the error in HFE calculations. We use here the multilinear regression model where the impact of the selected chemical properties on the HFE is linearly proportional to the values of the corresponding descriptors $\{D_i\}$ with empirical coefficients $\{a_i^{\text{model}}\}$:

$$\Delta G_{\text{hyd}}^{\text{SDC}} = \Delta G_{\text{hyd}}^{\text{model}} + \sum_i a_i^{\text{model}} D_i + a_0^{\text{model}} \tag{9}$$

where the second term is the set of structural corrections and a_0^{model} is a systematic error.[13]

Training and Test Sets. Another basic idea behind the SDC model is to calibrate the empirical coefficients $\{a_i^{\text{model}}\}$ on a set of "simple" solutes. They can be represented as an alkyl chain (linear or branched) which can contain only *one* substituent (e.g., benzene ring or chlorine atom). In the present work for the 1D RISM-SDC(QMq) model, we used a training set of 46 small neutral organic solutes: 22 alkanes, 17 alkylbenzenes, and 7 monochloroalkanes (see the Supporting Information). The modeling error for alkanes can be parametrized with corrections on PMV and branches. The set of alkylbenzenes requires an additional correction on the benzene ring; in turn, the modeling error for chloroalkanes can be parametrized with a linear combination of corrections on PMV, branches, and chlorine atoms. Experimental HFEs for all solutes from the training set were taken from ref 13, where HFEs were collected from several literature sources and then averaged.

We tested the calibrated 1D RISM-SDC model on a set of 220 chlorinated aromatic hydrocarbons (CAHs): 11 polychlorinated benzenes (from chlorobenzene to hexachlorobenzene, Table 2) and 209 polychlorinated biphenyls, PCBs (see the Supporting Information). The set of experimental HFEs for CAHs was compiled from different literature sources: (i) HFEs were taken from ref 13; (ii) log P(water/gas) values were collected from ref 39 and recalculated to HFEs with eq 10 ; (iii) K_{H} constants were taken from refs 40−44 and recalculated to HFEs with eq 3.

$$\Delta G_{\text{hyd}} = -(\ln 10)RT \log P(\text{water/gas}) \tag{10}$$

where ΔG_{hyd} is the hydration free energy, log P(water/gas) is the logarithm of the partition coefficient between the gaseous phase and water, R is the ideal gas constant, and T is the temperature.

Table 1. Descriptors and Corresponding Multilinear Regression Coefficients of the 1D RISM-SDC(QMq) Model for the Training Set of Solutes

descriptor		coefficient (kcal/mol)
		$a_0^{PW} = -4.19$
dimensionless partial molar volume	$(D_1 = \rho\overline{V})^a$	$a_1^{PW} = -1.48$
number of branches	$(D_2 = N_{br})$	$a_2^{PW} = 0.98$
number of benzene rings	$(D_3 = N_{benz})$	$a_3^{PW} = -3.11$
number of halogen atoms	$(D_4 = N_{hal})$	$a_4^{PW} = -1.30$

[a] $\overline{V}$ is the partial molar volume of the solute; $\rho = 0.0337$ Å^{-3}.

Computational Details. The HFEs were calculated with the 1D RISM method using the collection of numerical routines developed by our group.[45−47] Calculations were performed for the case of infinitely diluted aqueous solutions at $T = 300$ K. We used the Lue and Blankschtein version of the modified SPC/E model of water (MSPC/E),[48] proposed earlier by Pettitt and Rossky.[49] It differs from the original SPC/E water model[50] by the addition of Lennard-Jones (LJ) potential parameters for the water hydrogen ($\sigma_{Hw}^{LJ} = 0.8$ Å and $\varepsilon_{Hw}^{LJ} = 0.046$ kcal/mol). We took the MSPC/E bulk solvent correlation functions $h_{\alpha\beta}^{solv}(r)$ from ref 37.

To perform the calculations, one needs three sets of input solute data: (1) coordinates of atoms, (2) partial charges on atoms, and (3) atom LJ potential parameters. Coordinates of atoms for each molecule were optimized using the Gaussian 03 quantum chemistry software[51] at the B3LYP/6-31G(d,p) level of theory. The initial configurations for the solutes from the training set were taken from ref 13. In the case of the test set, atomic coordinates for several PCBzs and PCBs were taken from the Cambridge Structural Database.[52] Due to the fact that the hydrogen positions determined by standard X-ray methods can be inadequate,[53] we optimized the length of the C−H bonds with constrained C−C and C−Cl bonds. The geometrical parameters of all other CAHs (not presented in the Cambridge Structural Database) were found by structural optimization at the same level of theory without constrained bonds. Partial charges for all molecules were obtained by with the CHELPG procedure[54] at the B3LYP/6-31G(d,p) level of theory using the Gaussian 03 software.[51] We modeled all compounds with OPLS-AA (optimized potential for liquid simulations−all atom) LJ potential parameters[14,15,55] which were assigned to each atom automatically by the Maestro software (Schroedinger Inc.). The set of structural descriptors (eq 9) was assigned to each molecule automatically by the computer program "checkmol"[56] with the use of Python scripts.

RESULTS AND DISCUSSION

The 1D RISM-SDC(QMq) Model Calibration. Values of coefficients $\{a_i^{PW}\}$ of 1D RISM-SDC(QMq) model eq 9 with the considered set of descriptors were obtained using multilinear regression[57] against a training set of 46 solutes. The regression analysis was performed with the function *regress* from the Matlab Statistics Toolbox (MATLAB, version 7.8.0.347(R2009a), The MathWorks Inc., 2009). As one can see (Table 1), coefficients a_2^{PW}, a_3^{PW}, and a_4^{PW} have the same order of magnitude, indicating that each structural descriptor from the considered set is significant.

HFEs calculated by the 1D RISM-SDC(QMq) model for the training set of solutes are shown in Figure 4. Correlation coefficient

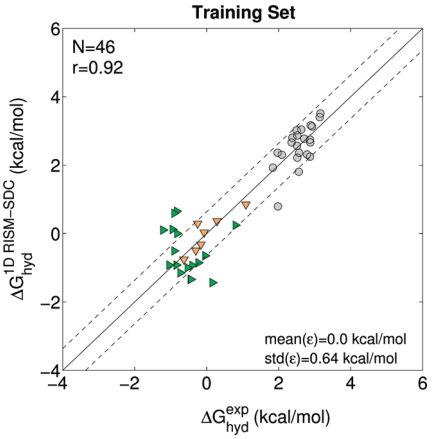

Figure 4. Correlation between the calculated and experimental HFEs for the training sets of solutes (gray circles are alkanes, orange triangles are alkylbenzenes, green triangles are chlorinated alkanes). The inset data show the statistical profile of the error $\varepsilon = \Delta G_{hyd}^{1DRISM\text{-}SDC} - \Delta G_{hyd}^{exp}$. Solid line illustrates the ideal correlation. Dashed lines indicate the standard deviation of the error.

r between the calculated and experimental HFEs was obtained with the function *corrcoef* from the same Matlab Statistics Toolbox and equals 0.92. It shows that the 1D RISM-SDC(QMq) model with four structural descriptors describes HFEs of 46 solutes from different chemical classes with high accuracy (the standard deviation of the error is 0.64 kcal/mol).

Predictive Ability of the 1D RISM-SDC(QMq) Model for CAHs. The predictive ability of the 1D RISM-SDC(QMq) model for HFE calculations was analyzed on the test set of 220 CAHs (see the section Training and Test Sets) and *the same* set of coefficients from Table 1 as for the training set. A comparison of the predicted and experimental HFEs is discussed below. We note that the reliable experimental data are very important for estimations of the accuracy of predicted results. Due to that, before the analysis of calculated data, we performed an estimation of reliability of experimentally obtained HFE values for each class of compounds from the test set.

Polychlorinated Benzenes (PCBzs). First of all, we analyzed the difference between the experimental HFEs for PCBzs obtained by different sources (see Table 2). Despite the fact that for several solutes (1,2,3-trichlorobenzene, 1,3,5-trichlorobenzene, and hexachlorobenzene) the HFE values differ by 0.5−0.6 kcal/mol (see Table 2), on average, HFE values obtained with different techniques deviate from the mean value by 0.2−0.3 kcal/mol. Thus, we concluded that experimental data for polychlorinated benzenes are sufficiently accurate and can be used for the estimation of the accuracy of the predicted data.

The comparison of the predicted and experimental HFE values is shown in Figure 5. To quantify the accuracy, we calculated statistical parameters of the error $\varepsilon = \Delta G_{hyd}^{1DRISM-SDC} - \Delta G_{hyd}^{exp}$ for the test set of polychlorinated benzenes (Figure 5, inset data). As one can see, results obtained with the 1D RISM-SDC(QMq) model are nonbiased (mean of the difference equals 0.02 ± 0.11 kcal/mol), and the standard deviation of the error is in the range of the deviation between different sources of the corresponding experimental data (~0.4 kcal/mol).

Polychlorobiphenyls (PCBs). For PCBs, experimental values of neither hydration free energy nor log P(water/gas) are available in the literature. However, since the 1980s, there have been

Table 2. Descriptors of the 1D RISM-SDC Model (eq 9) and Hydration Free Energies (ΔG_{hyd}) for Polychlorinated Benzenes[a]

					ΔG_{hyd} (kcal mol^{-1})			
name	D_1 ($\rho\overline{V}$)	D_2 (N_{br})	D_3 (N_{benz})	D_4 (N_{hal})	1D RISM-PW	1D RISM-SDC	$exp_{average}$	$exp_{\lvert max\rvert\_\lvert min\rvert}$
1,2,3,4-tetrachlorobenzene	5.24	4	1	4	14.75	−1.83	−1.32[39,40,43]	0.07
1,2,3-trichlorobenzene	4.85	3	1	3	13.77	−1.79	−1.49[39,43]	0.50
1,2,4,5-tetrachlorobenzene	5.30	4	1	4	15.40	−1.27	−1.34[39,43]	0.00
1,2,4-trichlorobenzene	4.89	3	1	3	14.23	−1.40	−1.22[39,40,43]	0.29
1,2-dichlorobenzene	4.45	2	1	2	12.85	−1.69	−1.47[13,39−41,43]	0.27
1,3,5-trichlorobenzene	4.93	3	1	3	14.71	−1.97	−1.09[39,43]	0.63
1,3-dichlorobenzene	4.48	2	1	2	13.24	−1.34	−1.13[39,41,43]	0.29
1,4-dichlorobenzene	4.49	2	1	2	13.15	−1.44	−1.15[39,41,43]	0.21
2-chlorotoluene	4.52	2	1	1	12.76	−0.55	−1.14[39]	
chlorobenzene	4.04	1	1	1	11.98	−1.51	−1.07[13,39−41,43]	0.22
hexachlorobenzene	5.95	6	1	6	16.33	−2.17	−2.26[43,42]	0.50

[a] Hydration free energies (ΔG_{hyd}) predicted by the uncorrected PW free energy expression and the 1D RISM-SDC model with QM-derived partial charges. Experimental values were averaged over different sources ($exp_{average}$); $exp_{|max|-|min|}$ shows the difference between the maximum and minimum values from different literature sources.

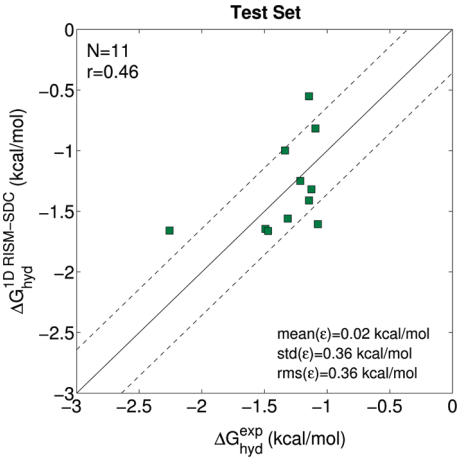

Figure 5. Correlation between the experimental HFE and values predicted by the 1D RISM-SDC(QMq) model for the test set of polychlorinated benzenes. The inset data show the statistical profile of the error $\varepsilon = \Delta G_{hyd}^{calc} - \Delta G_{hyd}^{exp}$. Solid line illustrates the ideal correlation. Dashed lines indicate the std(ε).

several experimental investigations of K_H values of PCBs reported, where the experiments were carried out with two dynamic techniques: (i) the gas stripping method (GSM)[58−61] and (ii) the "wetted-wall column" (WWC) or the concurrent flow technique.[44,62,63] All values are presented in Figure 6a; corresponding HFEs recalculated with eq 3 are presented in Figure 6b. One can see that the experimental K_H values are presented mainly by two sets of data obtained by the GSM (Bamford[59]) and the WWC technique (Brunner et al.[44]). Other sets of K_H values are not very large and contain about 20−30 values from 209 possible. Figure 6 shows that, for the same solutes, experimental K_H values from the GSM and WWC sets can differ considerably. The difference increases with the increase in the number of chlorine atoms in a solute. In terms of HFE, the difference varies from 1 kcal/mol for lighter PCB congeners (PCB with 4−5 chlorine atoms) to up to 3 kcal/mol for heavier congeners (higher chlorinated PCBs) (Figure 6b).

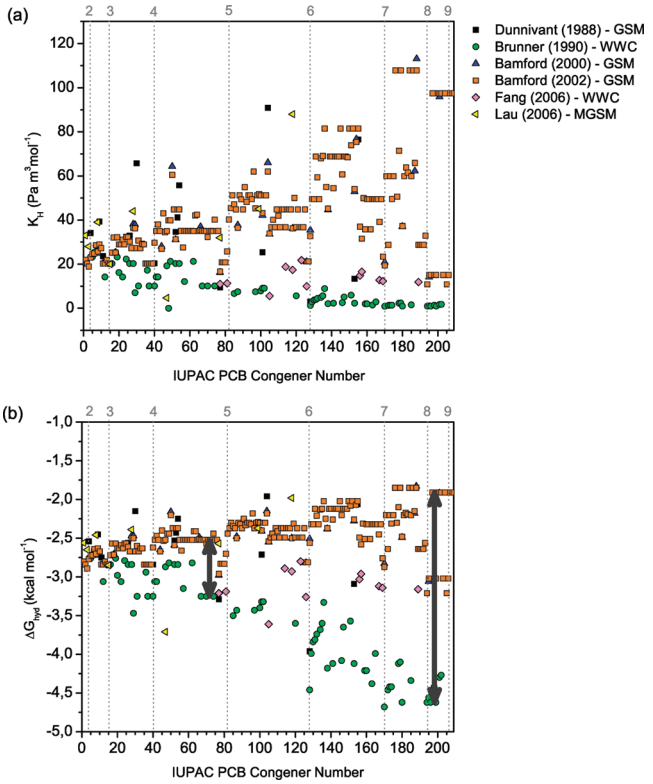

Figure 6. Experimental data for PCB congeners: (a) Henry's law constants, K_H, obtained with wetted-wall column (WWC), gas stripping method (GSM), or modified GSM (MGSM). (b) Hydration free energies (ΔG_{hyd}) recalculated from K_H. Black arrows show the deviation of experimental data obtained by the different techniques. Dashed lines show the separation of the whole set of PCBs with respect to the number of chlorine atoms (shown on the top).

Recently, it was found that the GSM overestimates K_H values for highly chlorinated biphenyls .[64,65] The problem is hidden in the technical implementation of the GSM. Within the method, the K_H of a compound is determined as a ratio of the equilibrium

concentrations of the compound in aqueous solution and vapor, accordingly. The compound is stripped from the aqueous phase into a gaseous phase using a bubble column apparatus (see the Supporting Information).[66] It was found that the sorption of the solute molecules to the surface of gas bubbles leads to a higher compound concentration in the gaseous phase, which, in turn, results in the overestimated K_H value. With the WWC technique, one can avoid this drawback. The technical implementation of the method consists of the equilibration of a compound between a thin layer of water and a concurrent flow of gas within the contact region.[66] Due to that, we accepted the experimental data obtained by the WWC method[44] as the most reliable set. Unfortunately, the total number of experimental values published in ref 44 is only 57 from 209 possible.

A comparison of HFEs, predicted by the 1D RISM-SDC-(QMq) model, with the experimental data (Table 3) shows that the calculated values are biased with respect to experimental ones, mean(ε) = −0.72 ± 0.07 kcal/mol, but have a small standard deviation of error. Figure 7 shows that the error remains the same for the whole set of PCBs and does not increase for the heavier PCB congeners.

Table 3. Statistical Profiles of Errors for Results Obtained by the Implicit Solvent Models for the Test Set of Polychlorobiphenyls (N = 57): Mean Value, Standard Deviation (std), and Root Mean Square (rms) of the Error $\varepsilon = \Delta G_{hyd}^{calc} - \Delta G_{hyd}^{exp}$ (kcal/mol)[a]

	model			
	1D RISM-PW	1D RISM-SDC	SM_6[67]	COSMO-SAC[67]
mean(ε)	20.35	−0.72	1.28	1.15
std(ε)	1.62	0.55	0.78	0.94
rms(ε)	20.42	0.91	1.50	1.49
r	−0.80	0.65	−0.35	−0.70

[a] r is the correlation coefficient. Results obtained by the SM_6 and COSMO-SAC methods were collected from ref 67.

Also, we performed a comparison of our results with HFEs obtained by other implicit models, SM_6 and COSMO-SAC (the data were taken from ref 67). Both of them treat the solvent as a homogeneous medium characterized by its dielectric constant (continuum solvent methods). Statistical analysis of the literature results is shown in Table 3. As one can see (Figure 7), HFEs obtained by these models are in good agreement with each other. However, the models allow predictions of HFE with high accuracy only for light congeners, whereas for the heavier PCBs, the error of HFE increases with the increase in the number of chlorine atoms. In the case of the highly chlorinated biphenyls (N_{Cl} = 8−9), the error is ∼3 kcal/mol. We explain these results as follows. In the case of lighter PCB congeners, the chlorine atoms are well-separated from each other. Thus, the total effect of chlorine atoms interactions with the solvent molecules can be presented as a sum of single chlorine atoms' contributions. Increasing the number of chlorine atoms in biphenyl leads to the interference of the chlorine atoms' interactions with the solvent molecules and, as a result, to a nonlinearity of the solvent response in the process of hydration. We underline that the 1D RISM approach considers these effects in a proper way, even in the case of highly chlorinated compounds. In turn, the continuum solvent models (SM_6 and COSMO-SAC) are not sensitive to the nonlinear solvent response. We note that using the RISM model for solvent is essential for the efficiency of the SDC model. As such, we tested our correction scheme with the use of the PBSA solvent model instead of the 1D RISM (see the Supporting Information). The results show that the RISM-based SDC model is superior to the PBSA-based model. Thus, for the test set of polychlorinated benzenes, the mean of error and the standard deviation of error of the PBSA-SDC model are ∼11.3 kcal/mol and ∼6.7 kcal/mol, accordingly; that is much worse than the 1D RISM-SDC results.

The results of this work show the potential of the 1D RISM-SDC(QMq) approach for the description of a hydration/solvation process for a wide range of chemical solutes. It makes the model a good candidate for use in large-scale environmental modeling of hydration pathways of organic pollutants.

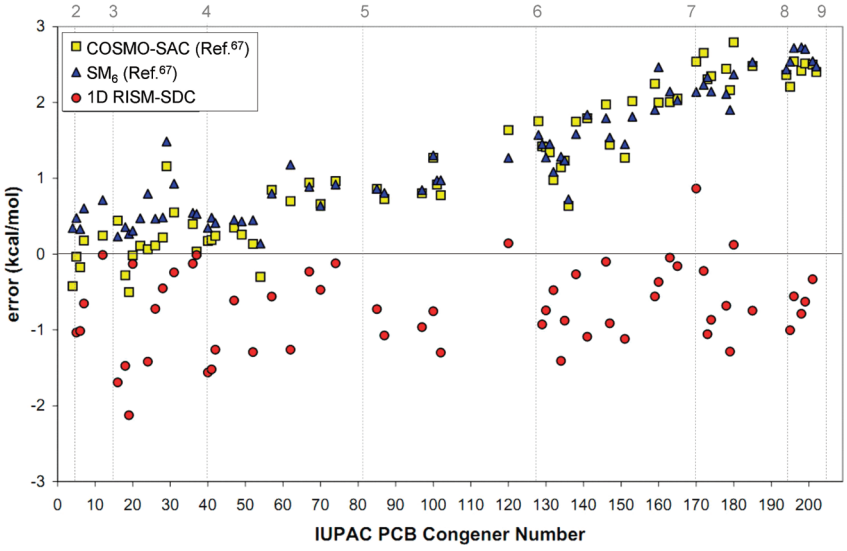

Figure 7. Errors for HFE predictions by the 1D RISM-SDC(QMq) model proposed in this work for the test set of polychlorobiphenyls (PCBs). The errors are compared with the corresponding literature data for SM_6 and COSMO-SAC (taken from ref 67). The HFE prediction error increases for SM_6 and COSMO-SAC with the increase in IUPAC number. At the same time, the 1D RISM-SDC(QMq) error remains the same for all PCBs. Dashed lines show the separation of the whole set of PCBs with respect to the number of chlorine atoms (shown on the top).

CONCLUSIONS

Here, we discussed a new method for predicting the hydration free energy (HFE) of organic pollutants and illustrated the efficiency of the method on a set of 220 chlorinated aromatic hydrocarbons that are in the list of persistent organic pollutants. The model is computationally inexpensive, and one HFE calculation takes only a minute on a standard PC (3.3 GHz). The method provides good accuracy for the test set of organic pollutant molecules. However, analysis of the results shows that the model performs better for polychlorinated benzenes than for polychlorobiphenyls. On one hand, that means that the SDC model might still require some improvement. That can be done in two directions: (i) one can use more sophisticated molecular theories, such as 3D RISM[16,68,69] [we note, however, that the 3D approach is significantly more computationally expensive than the 1D RISM approach used here (roughly by 2 orders of magnitude) and that might limit its application for large-scale screening of pollutants]; (ii) one can also work on the improvement of the theoretical part of the model by developing new, more efficient forms of the HFE functional. This is the subject of our future research.

On the other hand, the observed ~1 kcal/mol bias of the model results from experimental data for PCBs may be attributed to the differences in the quality of experimental data for polychlorinated benzenes and polychlorobiphenyls. We note that the sources of experimental data for these two classes of pollutants are different. However, as shown in Figure 6, the PCB congeners HFEs obtained from different sources can vary by several kilocalories per mole. We note that the problem of the lack of reliable experimental data for pollutants was highlighted in refs 2 and 65. Computational and theoretical scientists can do very little to improve the situation in that respect, but we hope that our results and analysis of the available experimental data will provoke experimentalists to revisit the question and, hopefully, to make additional independent measurements of HFE for CAHs. Such new experimental data would be very valuable in creating and testing new models for environmental modeling with high predictive ability.

ASSOCIATED CONTENT

Supporting Information. The composition of the training set and test subset of polychlorobiphenyls together with the corresponding experimental and calculated hydration free energies. Performance of the SDC model with the continuum PBSA model as the initial approximation. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49 341 9959 804. Fax: +49 341 9959 999. E-mail: fedorov@mis.mpg.de.

ACKNOWLEDGMENT

This work was partially funded by Deutsche Forschungsgemeinschaft (DFG)—German Research Foundation, Research Grant FE 1156/2-1 and REA—Research Executive Agency, Grant No. 247500 "BioSol", Programme: FP7-PEOPLE-2009-IRSES. The authors would like to acknowledge the use of the Chemical Database Service at Daresbury. We are thankful to Andriy Kovalenko for his insightful comments on the history of RISM and to Sergey Gusarov for useful discussions on hybrid RISM/QM models. We thank David S. Palmer and Andrey I. Frolov for critical reading of the manuscript and useful discussions. We acknowledge Katarzyna Baier and Evelyn Weiser for their help with finding references for this paper.

REFERENCES

(1) Vallack, H. W.; *Environ. Toxicol. Pharmacol.* **1998**, *6*, 143–175.
(2) Valsaraj, K. T.; Thibodeaux, L. J. *J. Phys. Chem. Lett.* **2010**, *1*, 1694–1700.
(3) Aarhus Protocol on Persistent Organic Pollutants (POPs). http://www.unece.org/env/lrtap/pops_h1.htm (accessed December 18, 2009).
(4) Jones, K. C.; de Voogt, P. *Environ. Pollut.* **1999**, *100*, 209–221.
(5) Wine, P. H. *J. Phys. Chem. Lett.* **2010**, *1*, 1749–1751.
(6) Mackay, D. In *Multimedia Environmental Models: The Fugacity Approach*, 2nd ed.; CRC Press: Boca Raton, FL, 2001; Chapter Environmental Chemicals and Their Properties, pp 29−54.
(7) Beyer, A.; Biziuk, M. *Rev. Environ. Contam. Toxicol.* **2009**, *201*, 137–158.
(8) Scheringer, M.; Wegmann, F.; Fenner, K.; Hungerbuhler, K. *Environ. Sci. Technol.* **2000**, *34*, 1842–1850.
(9) Wania, F.; Mackay, D. *The Global Distribution Model. A Non-Steady State Multicompartment Mass Balance Model of the Fate of Persistent Organic Pollutants in the Global Environment*; University of Toronto: Scarborough, Canada, 2000
(10) Strand, A.; Hov, O. *Water, Air, Soil Pollut.* **1996**, *86*, 283–316.
(11) Liss, P. S.; Slater, P. G. *Nature* **1974**, *247*, 181–184.
(12) Modarresi, H.; Modarress, H.; Dearden, J. C. *SAR QSAR Environ. Res.* **2005**, *16*, 461–482.
(13) Ratkova, E. L.; Chuev, G. N.; Sergiievskyi, V. P.; Fedorov, M. V. *J. Phys. Chem. B* **2010**, *114*, 12068–12079.
(14) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
(15) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
(16) *Molecular Theory of Solvation*; Hirata, F., Ed.; Kluwer Academic Publishers: Dordrecht, Netherlands, 2003.
(17) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation*; Academic Press: New York, 2002; Chapter Basics, pp 7−23.
(18) Kinoshita, M.; Okamoto, Y.; Hirata, F. *J. Am. Chem. Soc.* **1998**, *120*, 1855–1863.
(19) Kinoshita, M.; Okamoto, Y.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 4090–4100.
(20) Freedman, H.; Truong, T. N. *Chem. Phys. Lett.* **2003**, *381*, 362–367.
(21) Monson, P. A.; Morriss, G. P. *Adv. Chem. Phys.* **1990**, *77*, 451–550.
(22) Duh, D. M.; Haymet, A. D. J. *J. Chem. Phys.* **1995**, *103*, 2625–2633.
(23) Singer, S. J.; Chandler, D. *Mol. Phys.* **1985**, *55*, 621–625.
(24) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095–10112.
(25) Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1972**, *57*, 1930–1937.
(26) Hirata, F.; Rossky, P. J. *Chem. Phys. Lett.* **1981**, *83*, 329–334.
(27) Palmer, D. S.; Sergiievskyi, V. P.; Jensen, F.; Fedorov, M. V. *J. Chem. Phys.* **2010**, *133*, 044104.
(28) Karino, Y.; Fedorov, M. V.; Matubayasi, N. *Chem. Phys. Lett.* **2010**, *496*, 351–355.
(29) Chandler, D.; Singh, Y.; Richardson, D. M. *J. Chem. Phys.* **1984**, *81*, 1975–1982.
(30) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10403–10417.
(31) Ten-no, S. *J. Chem. Phys.* **2001**, *115*, 3724–3731.
(32) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *113*, 2793–2805.
(33) Chuev, G. N.; Fedorov, M. V.; Crain, J. *Chem. Phys. Lett.* **2007**, *448*, 198–202.
(34) Lee, P. H.; Maggiora, G. M. *J. Phys. Chem.* **1993**, *97*, 10175–10185.
(35) Ten-no, S.; Jung, J.; Chuman, H.; Kawashima, Y. *Mol. Phys.* **2010**, *108*, 327–332.
(36) Sato, K.; Chuman, H.; Ten-no, S. *J. Phys. Chem. B* **2005**, *109*, 17290–17295.

(37) Chuev, G. N.; Fedorov, M. V. *J. Comput. Chem.* **2004**, *25*, 1369–1377.
(38) Chuev, G. N.; Fedorov, M. V.; Chiodo, S.; Russo, N.; Sicilia, E. *J. Comput. Chem.* **2008**, *29*, 2406–2415.
(39) Abraham, M. H.; Andonianhaftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777–1791.
(40) Ryu, S. A.; Park, S. J. *Fluid Phase Equilib.* **1999**, *161*, 295–304.
(41) Ashworth, R. A.; Howe, G. B.; Mullins, M. E.; Rogers, T. N. *J. Hazard. Mater.* **1988**, *18*, 25–36.
(42) Jantunen, L. M.; Bidleman, T. F. *Chemosphere* **2006**, *62*, 1689–1696.
(43) Rounds, S. A.; Pankow, J. F. *J. Chromatogr.* **1993**, *629*, 321–327.
(44) Brunner, S.; Hornung, E.; Santl, H.; Wolff, E.; Piringer, O. G.; Altschuh, J.; Brueggemann, R. *Environ. Sci. Technol.* **1990**, *24*, 1751–1754.
(45) Fedorov, M. V.; Kornyshev, A. A. *Mol. Phys.* **2007**, *105*, 1–16.
(46) Fedorov, M. V.; Flad, H. J.; Chuev, G. N.; Grasedyck, L.; Khoromskij, B. N. *Computing* **2007**, *80*, 47–73.
(47) Sergiievskyi, V. P.; Hackbusch, W.; Fedorov, M. V. *J. Comput. Chem.* **2011**, in press, DOI: 10.1002/jcc.21783.
(48) Lue, L.; Blankschtein, D. *J. Phys. Chem.* **1992**, *96*, 8582–8594.
(49) Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1982**, *77*, 1451–1457.
(50) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
(51) Frisch, M. J. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.
(52) Fletcher, D. A.; McMeeking, R. F.; Parkin, D. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 746–749.
(53) Hope, H.; Ottersen, T. *Acta Crystallogr., Sect. B: Struct. Sci.* **1978**, *34*, 3623–3626.
(54) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.
(55) Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. *J. Phys. Chem. B* **2002**, *106*, 11673–11680.
(56) Feldman, H.; Dumontier, M.; Ling, S.; Haider, N.; Hogue, C. *FEBS Lett.* **2005**, *579*, 4685–4691.
(57) *Handbook of Chemoinformatics. 4 Bde. From Data to Knowledge*, 1st ed.; Gasteiger, J., Ed.; Wiley-VCH: New York, 2003.
(58) Dunnivant, F. M.; Coates, J. T.; Elzerman, A. W. *Environ. Sci. Technol.* **1988**, *22*, 448–453.
(59) Bamford, H. A.; Poster, D. L.; Baker, J. E. *J. Chem. Eng. Data* **2000**, *45*, 1069–1074.
(60) Bamford, H. A.; Poster, D. L.; Huie, R. E.; Baker, J. E. *Environ. Sci. Technol.* **2002**, *36*, 4395–4402.
(61) Lau, F. K.; Charles, M. J.; Cahill, T. M. *J. Chem. Eng. Data* **2006**, *51*, 871–878.
(62) Fendinger, N.; Glotfelty, D. *Environ. Toxicol. Chem.* **1990**, *9*, 731–735.
(63) Fang, F.; Chu, S. G.; Hong, C. S. *Anal. Chem.* **2006**, *78*, 5412–5418.
(64) Goss, K.-U.; Wania, F.; McLachlan, M. S.; Mackay, D.; Schwarzenbach, R. P. *Environ. Sci. Technol.* **2004**, *38*, 1626–1628.
(65) Shunthirasingham, C.; Lei, Y. D.; Wania, F. *Environ. Sci. Technol.* **2007**, *41*, 3807–3814.
(66) Bamford, H.; Baker, J. In *Review of Methods and Measurements of Selected Hydrophobic Organic Contaminant Aqueous Solubilities, Vapor Pressures, and Air-Water Partition Coefficients*; National Institute of Standards and Technology: Gaithersburg, MD, 1998; Chapter Physical and Chemical properties.
(67) Phillips, K. L.; Sandler, S. I.; Greene, R. W.; Di Toro, D. M. *Environ. Sci. Technol.* **2008**, *42*, 8412–8418.
(68) Palmer, S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. *J. Phys.: Condens. Matter* **2010**, *22*, 492101.
(69) Frolov, A. I.; Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. *J. Phys. Chem. B* **2011**, doi: 10.1021/jp111271c.