An accurate prediction of hydration free energies by combination of molecular integral equations theory with structural descriptors

Ekaterina L. Ratkova,† Gennady N. Chuev,†,‡ Volodymyr P. Sergieievskyi,† and Maxim V. Fedorov*,†

The Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, Leipzig, 04103, Germany, and Institute of Theoretical and Experimental Biophysics, Russian Academy of Science, Pushchino, Moscow Region, 142290, Russia

E-mail: fedorov@mis.mpg.de
Phone: +49 341 9959 756. Fax: +49 341 9959 999

Abstract

In this work we report a novel method for the estimation of the hydration free energy of organic molecules - Structural Descriptors Correction (SDC) model. The method is based on a combination of the Reference Interaction Site Model (RISM) with several empirical corrections. The model requires only a small number of chemical descriptors associated with main features of the chemical structure of solutes: excluded volume, branch, double bond, benzene ring, hydroxyl group, halogen atom, aldehyde group, ketone group, ether group, and phenol fragment. The optimum model was selected after testing of different RISM free energy expressions on a training set of 65 molecules. We show that the correction parameters of the SDC

*To whom correspondence should be addressed
†The Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, Leipzig, 04103, Germany
‡Institute of Theoretical and Experimental Biophysics, Russian Academy of Science, Pushchino, Moscow Region, 142290, Russia
model are transferable between different chemical classes that allows one to cover a wide range of organic solutes. The new model substantially increases the accuracy of calculated HFEs by RISM giving the standard deviation of the error for a test set of 120 organic molecules around 1.2 kcal/mol.

**Introduction**

*Hydration free energy (HFE)* is one of the most important physical/chemical properties characterizing solvation of molecular species in aqueous solutions. This property can be used for description of many processes associated with solvation: from solubility to kinetics of reactions as well as complex formation and association in solution.\(^1\)–\(^6\) The HFEs of bioactive molecules can be used to estimate different pharmacokinetics properties of the compounds such as oral digestion, membrane penetration and absorption in different tissues.\(^5\)–\(^8\)

Despite of the fact that millions of organic compounds are known, experimental data on HFE are reported only for slightly more than a thousand of them.\(^9\)–\(^11\) The main reason is the numerous experimental difficulties associated with accurate measurements of HFE especially for compounds with low solubility and/or low volatility (we note, that many important organic compounds belong to this category).\(^1\)–\(^12\) Thermodynamics of molecular hydration is difficult to obtain through *direct* measurements. Typically, to study the process of molecular hydration one needs to investigate two other processes: dissolution and sublimation.\(^12\)–\(^13\) However, experimental studies of these processes for compounds with low solubility and/or low volatility require precise and laborious measurements with highly sensitive instruments.\(^14\)–\(^15\) Moreover, one has to be cautious about the chemical stability of the solutes under investigation.\(^2\) Although there is a certain progress in the development of experimental methods for obtaining HFEs,\(^1,2,12,16,17\) these studies still remain to be expensive and time-consuming.

Computational methods offer an alternative way to obtain HFEs and much of work has been done in this direction.\(^18\)–\(^26\) These methods are represented by two general approaches which treat solvent effects either (i) *implicitly*\(^30\) or (ii) *explicitly*.\(^27\) The first approach is computationally
inexpensive because it sufficiently reduces the number of degrees of freedom considered in the model.\textsuperscript{28–30} In turn, the explicit models take into account the molecular details of solute and solvent structures.\textsuperscript{3,4,18,20,31–37} However, as it has been recently reported,\textsuperscript{22} even the computationally expensive explicit fully atomistic methods still suffer a sufficient lack of accuracy if applied to an arbitrary selected set of organic molecules. To our opinion, the reported poor performance in Ref.\textsuperscript{22} of various computational methods for HFE prediction is mainly due to the fact that all these methods implicitly or explicitly operate with sets of different parameters which perform well only for selected classes of compounds. However, these parameters are hardly transferable from one class to another. Therefore, the development of an accurate general computational method for HFE prediction is still a challenge.

In general, the implicit models are most popular for HFE calculations due to their low computational costs. In addition, the implicit solvent models can be combined with quantum mechanical description of solutes in a straightforward manner that allows one to model the solvent effects on the electronic structure of the solute.\textsuperscript{28–30} There is a large variety of the implicit models depending on the level of the theoretical treatment of the solvent.\textsuperscript{28–30,38–40} The continuum electrostatics approach\textsuperscript{28,30} is computationally inexpensive even for large biopolymer systems\textsuperscript{41,42} but does not take into account any effects of the solvent molecular structure. However, there exist another class of implicit models based on the integral equation theory of molecular liquids.\textsuperscript{29,43–47} These models operate with solute-solvent density correlation functions, and, therefore, take into account the effects of solvent molecular structure on the properties of solute.\textsuperscript{44,45,48–51}

The reference interaction site model (RISM)\textsuperscript{51,52} has a special place among the statistical mechanics methods, since it combines a reasonable level of molecular description of the solute with low computational costs. The original RISM method pioneered by Chandler and Andersen requires a solution of the site-site Ornstein-Zernike (SSOZ) integral equations combined with a local algebraic relation, so-called hypernetted chain (HNC) closure (see more explanation in the section Main formulae).\textsuperscript{51} Later on this model was extended to the dipolar liquids by the XRISM treatment.\textsuperscript{52}
By now a number of studies have been published on RISM applications to HFE calculations. Many of them used HNC or partially linearized HNC (PLHNC) closure relations (see the section [Main formulae]). The main advantage of the HNC-based methods is that they provide analytical expressions for HFE calculations. Although some of the RISM method applications provide good agreement with experimental data, systematic studies have indicated that the accuracy of HFEs calculated by RISM/HNC or RISM/PLHNC methods is not satisfactory. Moreover, as it has been reported in works, HFEs of organic solutes calculated by RISM combined with HNC or PLHNC closures may differ from the corresponding experimental values by an order of magnitude.

To overcome this shortcoming, various methodologies have been proposed such as the ’three-dimensional’ (3D) extension of the RISM, applications of repulsive bridge corrections, or diagrammatic proper integral equations. However, despite of all these improvements, accurate RISM calculations of HFEs for a wide range of organic compounds still remain a challenge.

Many efforts have been spent to improve the theoretical background of the RISM-based approaches for HFE calculations. Several advanced models have been developed to describe thermodynamics of hydration more accurately than previous methods. One of the earliest models developed by Chandler, Singh and Richardson assumed Gaussian fluctuations (GF) of solvent. Although GF HFE expression provides better agreement with experimental data for some solutes, it is not generally used due to the improper account of molecular effects for polar solutes. Later, yet another approach referred as the partial wave (PW) model has been proposed by Ten-no and Iwata. This approach is based on the distributed partial wave expansion and more accurately accounts angular correlations of solvent molecules around the solute. Nevertheless, the recent analysis of some of us has indicated that the PW model sufficiently overestimates HFE for non-polar solutes. This analysis showed significant correlations between the HFEs calculated with PW method and experimental data. Based on this observation, a semiempirical correction for the PW model - Partial Wave Correction (PWC) model has been proposed in Ref. For twenty organic solutes the PWC model provided better agreement with experimental data than the original PW
Moreover, as it has been reported in Ref.\textsuperscript{74} applications of the empirical corrections for the RISM calculations may significantly improve the models for supramolecular aggregation in water solutions. It was shown that RISM/HNC method with the original PW HFE expression fails to describe the supramolecular structures formed by self-aggregating cyanine dyes in aqueous solution, whereas application of the PWC expression allows to correctly describe the structure and stability of these aggregates.\textsuperscript{74} However, due to the inherent limitations of the model (pure parameterization and small number of corrections) this model cannot be applied 'as is' for a wide range of organic solutes.\textsuperscript{63}

In chemoinformatics applications parameterization of empirical corrections using some scoring functions is a common practice these days to predict different kinds of quantitative structure - activity/property relationships (QSAR/QSPR).\textsuperscript{5,8,32,75–80} With respect to HFE calculations, such parametrization has been used for explicit models\textsuperscript{18} to decrease computational costs and for implicit models\textsuperscript{5,9,26,81–83} to improve the accuracy of calculations within the framework of continuum electrostatics. The choice of functional model for parametrization is rather wide in general case, such, the parametrization can be based on pure statistical analysis,\textsuperscript{84} physical assumptions like the linear response theory,\textsuperscript{85,86} or empirical linear free-energy relationship.\textsuperscript{87–89} The number of required descriptors for empirical corrections may vary from just a few descriptors based on physical/chemical properties of solutes,\textsuperscript{82} up to a hundred descriptors derived from the group/atom contribution approach.\textsuperscript{90} Application of the atomic or group structural descriptors becomes complicated due to the significant amount of the descriptors which is required for predictions of HFEs for a wide range of solutes.\textsuperscript{26,90} In turn, more general physical/chemical descriptors can be successfully applied only for some particular classes of solutes, but it is difficult to transfer the descriptors from one chemical classes of solutes to another.

In the present work we describe a novel hybrid method for computing HFEs of organic solutes which combines the advantages of both RISM and chemoinformatics approaches. The method is based on a combination of the RISM-based HFE expression and several empirical corrections - Structural Descriptors Correction (SDC) model. The model requires only a small number of
empirical chemical descriptors associated with main features of the chemical structure of solutes. We will show that the correction parameters of the SDC model are *transferable* from one class of solutes to another that allows one to cover a wider range of organic solutes. The proposed combination of the chemoinformatical approach with RISM substantially increases the accuracy of calculated HFEs. The paper is organized as follows. In the section [Methods] we describe the basic theoretical and methodological background of our work. In [Results and Discussion] we present the results of our calculations and compare them with experimental HFEs available from the literature. In [Conclusions] we summarize our findings and discuss future avenues of improving our method and expanding this to many other classes of organic molecules. In Appendix 1 we provide the computational details of RISM calculations.

In this pilot paper we would like to demonstrate the efficiency of the SDC model for a set of 120 solutes (non-polar hydrocarbons, polar solutes with hydroxyl, halo, aldehyde, carbonyl and ether groups) for which experimental data of HFE are available in the literature (see Table-S1 of the Supporting information). More solutes with different functional groups will be considered in our forthcoming publications.

**Methods**

**Hydration Free Energy Data Sets**

We used a data set 185 experimental HFEs for neutral organic small solutes which we compiled from different literature sources.\(^9\)-\(^11\),\(^18\),\(^90\)–\(^92\) Chosen solutes can be represented as a combination of several moieties: *alkyl*, *alkenyl*, *phenyl*, *hydroxyl*, *halo*, *aldehyde*, *carbonyl*, and *ether*. In the present work we specified also *phenol fragment* as a separate moiety. We name solutes consisting of either only alkyl moiety or its combination with only *one* another moiety as *simple* solutes. In turn, we name solutes consisting of combination of alkyl moiety with *several* others (of the same or different types) as *polyfragment*. We consider several chemical classes of *simple* and polyfragment solutes (Table 2 and Table S-1 of the Support Information). We used a *training set*
of 65 'simple' solutes for the SDC model calibration. Another 120 solutes formed the test set, which contained 60 'simple' solutes from the same chemical classes as used in the training set as well as 60 polyfragment solutes. Detailed information about the training and test sets is presented in the Support Information together with the corresponding experimental and calculated HFEs.

**Basics of RISM**

**Main formulae**

We use here the so-called one-dimensional RISM (1D RISM) approach,\textsuperscript{45} where the solute and solvent molecules are modeled as sets of sites interacting via pairwise spherically-symmetric potentials. As the main goal of our paper is to develop a computationally inexpensive method for accurate calculations of HFEs we are not using the 3D-RISM method here because of the large computational costs associated with that model.\textsuperscript{124,125} In 1D RISM the solute and solvent density correlations are described by spherically-symmetric site-site correlation functions. In the RISM theory there are three kinds of site-site symmetric site-site correlation functions: intramolecular correlation functions $\omega(r)$, total correlation functions $h(r)$ and direct correlation functions $c(r)$.\textsuperscript{45} In general, these are 3D functions but due to the spherical symmetry we consider only their 1D radial parts $\omega(r)$, $h(r)$ and $c(r)$ which depend only from the radial distance $r$.

**Intramolecular correlation functions** describe the structure of the molecule. We will use notation $\omega_{ss'}(r)$ for the intramolecular correlation function between the sites $s$ and $s'$ of the solute molecule and notation $\omega_{\alpha\xi}(r)$ for the intramolecular correlation function between the sites $\alpha$ and $\xi$ of the solvent molecule. For given two sites $s$ and $s'$ of one molecule the intramolecular correlation function is written as:

$$\omega_{ss'}(r) = \frac{\delta(r - r_{ss'})}{4\pi r_{ss'}^3},$$

where $r_{ss'}$ is the distance between the sites and $\delta(r - r_{ss'})$ is the Dirac delta-function.

The intermolecular solute-solvent correlations in RISM are described by the pairwise total correlation functions $\{h_{s\alpha}(r)\}$, and direct correlation functions $\{c_{s\alpha}(r)\}$ (index $s$ corresponds to the
solute sites and index $\alpha$ corresponds to the solvent site). The total correlation functions $\{h_{s\alpha}(r)\}$ can be expressed via the radial distribution function (RDF)

$$h_{s\alpha}(r) = g_{s\alpha}(r) - 1,$$

(2)

where $g_{s\alpha}(r)$ is RDF of sites $\alpha$ of the solvent molecule around the site $s$ of the solute. The model also uses the solvent-solvent total correlation functions $\{h_{\alpha\xi}^{\text{solv}}(r)\}$ which relates with the distribution of sites $\xi$ of solvent molecules around the site $\alpha$ of a selected single solvent molecule.

Direct correlation functions $c_{s\alpha}(r)$ are connected with the total correlation functions via the set of RISM integral equations:

$$h_{s\alpha}(r) = \exp(-\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r) + B_{s\alpha}(r)) - 1,$$

(4)

where $r = |r_1 - r_2|$ and $\chi_{\alpha\xi}(r) = \omega_{\alpha\xi}^{\text{solv}}(r) + \rho h_{\alpha\xi}^{\text{solv}}(r)$ are the bulk solvent susceptibility functions, $N_{\text{solute}}$ and $N_{\text{solvent}}$ are number of sites in solute and solvent, correspondingly.

To make Eq. (3) complete, $N_{\text{solute}}N_{\text{solvent}}$ closure relations are introduced:

$$h_{s\alpha}(r) = \frac{N_{\text{solute}}N_{\text{solvent}}}{\sum_{s'=1}^{N_{\text{solute}}} \sum_{\xi=1}^{N_{\text{solvent}}} \int \int \omega_{s\xi'}(|r_1 - r'|) c_{s\xi'}(|r' - r''|) \chi_{\alpha\xi'}(|r'' - r_2|) dr'dr''}$$

(3)

where $u_{s\alpha}(r)$ is a pair interaction potential between the sites $s$ and $\alpha$, $B_{s\alpha}(r)$ are site-site bridge functions, $\beta = 1/k_B T$, $k_B$ is the Boltzmann constant, $T$ is the temperature.

We used here the common form of the interaction potential represented by the long range electrostatic term $u_{s\alpha}^{el}(r)$ and short-range Lennard-Jones (LJ) term $u_{s\alpha}^{LJ}(r)$ as:

$$u_{s\alpha}(r) = u_{s\alpha}^{el}(r) + u_{s\alpha}^{LJ}(r),$$

$$u_{s\alpha}^{el}(r) = \frac{q_s q_{\alpha}}{r}, \quad u_{s\alpha}^{LJ}(r) = 4\varepsilon_{s\alpha}^{LJ} \left[ \left( \frac{\sigma_{s\alpha}^{LJ}}{r} \right)^{12} - \left( \frac{\sigma_{s\alpha}^{LJ}}{r} \right)^6 \right],$$

(5)

where $\{q_s, q_{\alpha}\}$ are the partial electrostatic charges of the corresponding solute and solvent sites,
\{\epsilon_{\alpha}^{LJ}, \sigma_{\alpha}^{LJ}\} \text{ are the LJ solute-solvent interaction parameters.}

In general, the exact bridge functions are practically uncomputable and one needs to use some approximation.\textsuperscript{44,45,93} The most straightforward and widely used model is the RISM/HNC approximation, which sets \(B_{\alpha}(r)\) to zero.\textsuperscript{57} The corresponding HNC closure reads then:

\[
h_{\alpha}(r) = \exp\left[\beta u_{\alpha}(r) + h_{\alpha}(r) - c_{\alpha}(r)\right] - 1. \tag{6}
\]

\(s = 1, \ldots, N_{\text{solute}}; \quad \alpha = 1, \ldots, N_{\text{solvent}}\)

However, due to the uncontrolled growth of the argument of the exponent the use of HNC closure can lead to slow convergence rate, and in many cases even divergence of the numerical solution of RISM equations. One way to overcome this problem is to linearize the exponent when its argument is larger than a certain threshold constant \(C\):

\[
h_{\alpha}(r) = \begin{cases} 
\exp(\Xi_{\alpha}(r)) - 1 & \text{when } \Xi_{\alpha}(r) < C \\
\Xi_{\alpha}(r) + \exp(C) - C - 1 & \text{when } \Xi_{\alpha}(r) > C
\end{cases} \tag{7}
\]

where \(\Xi_{\alpha}(r) = -\beta u_{\alpha}(r) + h_{\alpha}(r) - c_{\alpha}(r)\). The linearized HNC closure for the case \(C = 0\) was proposed by Hirata and Kovalenko in Ref.\textsuperscript{69} where it has been called PLHNC closure. When \(C\) goes to infinity, the closure PLHNC (Eq. (7)) becomes the HNC closure (Eq. (6)). In the current work we performed calculations for both limiting cases: \(C = 0\) and \(C \to \infty\). However, we found that the calculations converge much better with the RISM/PLHNC method \((C = 0)\) rather than with the RISM/HNC method (see the Supporting Information). As the difference between the accuracy of SDC HFEs with RISM/HNC and RISM/PLHNC correlation functions is practically negligible (see the Supporting Information), we discuss below only the SDC results based on the correlation functions obtained by the RISM/PLHNC method.

**Hydration Free Energy Expressions in RISM**

Within the framework of the RISM theory there are several expressions which allow one to obtain values of the HFE from the total \(h_{\alpha}(r)\) and direct \(c_{\alpha}(r)\) correlation functions. In our work
we compared the accuracy of the most popular HFE expressions discussed in the Introduction, namely HNC (Eq. (8)); GF (Eq. (9)), KH (Eq. (10)), PW (Eq. (11)), HNCB expression (Eq. (13)), proposed by Kovalenko and Hirata, and PWC method (Eq. (15)), which are given by the equations below.

\[
\Delta \mu_{\text{hyd}}^{\text{HNC}} = 2\pi \rho k_B T \sum_{s=1}^{N_{\text{solute}}} \sum_{\alpha=1}^{N_{\text{solvent}}} \int_0^\infty \left[ -2c_{s\alpha}(r) - h_{s\alpha}(r) (c_{s\alpha}(r) - h_{s\alpha}(r)) \right] r^2 \, dr \tag{8}
\]

\[
\Delta \mu_{\text{hyd}}^{\text{GF}} = 2\pi \rho k_B T \sum_{s=1}^{N_{\text{solute}}} \sum_{\alpha=1}^{N_{\text{solvent}}} \int_0^\infty \left[ -2c_{s\alpha}(r) - c_{s\alpha}(r) h_{s\alpha}(r) \right] r^2 \, dr \tag{9}
\]

\[
\Delta \mu_{\text{hyd}}^{\text{KH}} = \Delta \mu_{\text{hyd}}^{\text{GF}} + 2\pi \rho k_B T \sum_{s=1}^{N_{\text{solute}}} \sum_{\alpha=1}^{N_{\text{solvent}}} \int_0^\infty h_{s\alpha}^2(r) \Theta(-h_{s\alpha}(r)) r^2 \, dr \tag{10}
\]

\[
\Delta \mu_{\text{hyd}}^{\text{PW}} = \Delta \mu_{\text{hyd}}^{\text{GF}} + 2\pi \rho k_B T \sum_{s=1}^{N_{\text{solute}}} \sum_{\alpha=1}^{N_{\text{solvent}}} \int_0^\infty \tilde{h}_{s\alpha}(r) h_{s\alpha}(r) r^2 \, dr \tag{11}
\]

where \( r = |r_1 - r_2| \) and

\[
\tilde{h}_{s\alpha}(|r_2 - r_1|) = \sum_{s' = 1}^{N_{\text{solute}}} \sum_{\alpha' = 1}^{N_{\text{solvent}}} \int \int \tilde{\omega}_{s's}(|r_1 - r'|) \tilde{h}_{s'\alpha'}(|r' - r''|) \tilde{\omega}_{\alpha'\alpha}^{\text{solv}}(|r'' - r_2|) \, dr' \, dr''. \tag{12}
\]

\( \tilde{\omega}_{s's}(r) \) and \( \tilde{\omega}_{\alpha'\alpha}^{\text{solv}}(r) \) are the elements of matrices \( W^{-1}, W_{\text{solv}}^{-1} \) which are inverses to the matrices

\[
W = [\omega_{s's}(r)]_{N_{\text{solute}} \times N_{\text{solute}}} \quad \text{and} \quad W_{\text{solv}} = [\omega_{\alpha'\alpha}^{\text{solv}}(r)]_{N_{\text{solvent}} \times N_{\text{solvent}}}
\]

built from the solute and solvent intramolecular correlation functions \( \omega_{s's}(r) \) and \( \omega_{\alpha'\alpha}^{\text{solv}}(r) \) respectively.

The HFE expression for the HNCB model is

\[
\Delta \mu_{\text{hyd}}^{\text{HNCB}} = \Delta \mu_{\text{hyd}}^{\text{HNC}} + 2\pi \rho k_B T \sum_{s=1}^{N_{\text{solute}}} \int_0^\infty (h_{s\alpha}(r) + 1)(e^{-B_{s\alpha}^R(r)} - 1) r^2 \, dr. \tag{13}
\]

Here \( \{B_{s\alpha}^R(r)\} \) are repulsive bridge correction functions, defined for each pair of solute \( s \) and
solvent $\alpha$ atoms by the expression:

$$\exp(-B_{s\alpha}^R(r)) = \prod_{\xi \neq \alpha} \left\langle \omega_{\alpha\xi} * \exp\left(-\beta \varepsilon_{s\xi} \left(\frac{\sigma_{s\xi}}{r}\right)^{12}\right) \right\rangle$$ (14)

where $\omega_{\alpha\xi}(r)$ are the solvent intramolecular correlation functions, and $\sigma_{s\xi}$ and $\varepsilon_{s\xi}$ are the site-site parameters of the pair-wise Lennard-Jones potential.

The PWC HFE $\Delta\mu_{hyd}^{PW}$ is given by

$$\Delta\mu_{hyd}^{PW} = \Delta\mu_{hyd}^{PW} + a \rho \beta^{-1} V_u + b N_{OH},$$ (15)

where $\Delta\mu_{hyd}^{PW}$ is given by (Eq. (11)), $\rho$ is the number density of solvent (water), $V_u$ is the excluded volume of the solute, and $N_{OH}$ is the number of OH groups in the solute molecule. $a$ and $b$ are the correction coefficients in the PWC model, and are determined by the fit of (Eq. (15)) to the experimental values of the solvation free energies of the training set of solute molecules.

**The SDC Model**

As stated in the Introduction the RISM approach provides a qualitatively correct description of the solvent structure around solute, but it does not provide the quantitatively accurate values of HFE with the standard RISM-based HFE expressions (Eq. (8))-(Eq. (11)). To overcome this drawback of RISM we introduce a set of structural corrections and propose an expression for the SDC model for HFE calculations (see Eq. (17) below).

We define the modeling error $\varepsilon$ of the RISM-based HFEs calculations for a solute as the difference between the calculated and experimental values:

$$\varepsilon = \Delta\mu_{hyd}^{model} - \Delta\mu_{hyd}^{exp},$$ (16)

where $\Delta\mu_{hyd}^{exp}$ is the experimental value of HFE, $\Delta\mu_{hyd}^{model}$ is the HFE calculated by the RISM approach (superscribe model denotes the RISM-based HFE expression, e.g. PW, GF, etc.). We are
mostly interested in the absolute error rather in the relative error because many of the HFE values are close to zero.

The main idea behind the SDC model is that we parametrize the modeling error $\varepsilon$ with a set of descriptors $\{D_i\}$ associated with specific features of the chemical structure of solutes such as partial molar volume, number of branches, number of double bonds, number of hydroxyl groups, number of benzene rings, number of phenol fragments, and others (see justification of the choice of descriptors below). The model is based on an assumption that different structural properties of the solute molecule contribute independently to the error in HFE calculations. We use here the linear regression model, where the impact of the selected chemical properties on the HFE are linearly proportional to the values of the corresponding descriptors $\{D_i\}$ with empirical coefficients $\{a_i^{\text{model}}\}$:

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

where $\Delta \mu_{\text{model hyd}}$ is the HFE calculated with a model expression within the RISM approach (see the section [Hydration Free Energy Expressions in RISM]), the second term is the set of structural corrections, $a_0^{\text{model}}$ is a constant (the meaning of this term will be explained below).

In this paper we discuss the fitting procedure which uses the PW expression for HFE calculations as an initial approximation. The similar analysis for the GF method is performed (see the Support Information for details).

**Choice of descriptors**

*n—Alkanes.* First we considered the training set of n—alkanes that have no specific interactions with water molecules, and, therefore, the excluded volume effect makes a significant contribution to their hydration.\textsuperscript{10,18,94} Several descriptors have been proposed which take into account the excluded volume effect: the solvent accessible surface area (SASA),\textsuperscript{94–96} dimensionless partial molar volume (DPMV),\textsuperscript{63} their combination,\textsuperscript{97} number of carbon atoms (for alkanes),\textsuperscript{98} etc. In the present work we used the DPMV descriptor $\rho V$ (where $V$ is the solute partial molar volume and $\rho$ is the number density of the solvent). The DPMV calculations within the framework of the
RISM approach for the case of infinitely diluted solution are straightforward: \(^{63}\)

\[
\rho V = 1 + \frac{4\pi \rho}{N_{solute}} \sum_s \int_0^\infty \left( h_{oo}^{solv}(r) - h_{so}(r) \right) r^2 dr,
\]

(18)

where \(h_{oo}^{solv}(r)\) is the total oxygen-to-oxygen correlation function of bulk water, \(h_{so}(r)\) is the total correlation function between the solute site \(s\) and the water oxygen.

It is known that \(n\)–alkanes have a linear relationship between their experimental HFEs and excluded volume.\(^{94–96,98}\) For the given training set of \(n\)–alkanes we plotted both \(\Delta \mu_{hyd}^{exp}\) and \(\Delta \mu_{hyd}^{PW}\) versus DPMV (Figure 1, a). The RISM/PW method gives qualitatively correct results (a linear dependence between \(\Delta \mu_{hyd}^{PW}\) and DPMV), but the dependence for the calculated data is considerably shifted with respect to the corresponding experimental data, and these dependencies have different slopes. The major shift of the calculated data can be corrected with the \(a_0^{PW}\) free coefficient (-3.58 kcal/mol) of the SDC expression (Eq. (17), the last term). This correction eliminates a general systematic error of the RISM model which does not depend on the solute excluded volume. For the sake of brevity, in the rest of the paper we will mainly talk only about contributions of the solute structure descriptors, although all calculations were done including the \(a_0^{PW}\) correction as well. We correct the slope of the HFEs calculated with PW method by the DPMV correction \((a_1^{PW}D_1, \text{where } D_1 = \text{DPMV and } a_1^{PW}\) is the slope for the linear dependence between \(\varepsilon\) (see Eq. (16)) and DPMV for \(n\)-alkanes). We note, that this correction was first proposed in the Ref.\(^ {63}\)

**Nonlinear alkanes.** On Figure 1 (b) we plotted the difference between HFEs calculated with the \(a_1^{PW}D_1\) correction and corresponding values of \(\Delta \mu_{hyd}^{exp}\) versus the DPMV for the whole training set of alkanes to check whether this correction is sufficient to provide a reasonable accuracy of HFE calculations for branched alkanes. One can see from this plot that for \(n\)–alkanes the difference is close to zero. The differences for the branched alkanes are shifted up with respect to those for \(n\)–alkanes. We assumed that the values of the shifts are approximately constant for the alkanes with the same number of branches and do not depend on DPMV. Analysis of the Figure 1 (b)
Figure 1: a) Experimental $\Delta \mu_{\text{hyd}}^{\exp}$ and calculated data $\Delta \mu_{\text{hyd}}^{PW}$ versus DPMV ($\rho V$) for the training set of $n$-alkanes. b) Difference between calculated $\Delta \mu_{\text{hyd}}^{(1)}$ and experimental $\Delta \mu_{\text{hyd}}^{\exp}$ data (where $\Delta \mu_{\text{hyd}}^{(1)} = \Delta \mu_{\text{hyd}}^{PW} + a_{PW}^{1} D_{1} + a_{PW}^{0}$) versus DPMV for linear and branched alkanes (the training set).

shows that the shifts are proportional to the number of branches ($N_{br}$). Thus, an introduction of one branch into the carbon skeleton of a solute has a constant effect on the error of calculated HFEs. This effect can be explained by another systematic error of the RISM approach which overestimates the influence of branches on the HFE. Therefore, we introduce another correction for the number of branches in the carbon skeleton ($a_{PW}^{2} D_{2}$, where $D_{2} = N_{br}$). Finally, we found that for alkanes it is sufficient to use the combination of $a_{PW}^{1} D_{1}$ correction with the $a_{PW}^{2} D_{2}$ correction to significantly decrease the error of calculated HFEs.

Other compounds. Next, we analyzed whether the described above empirical corrections ($a_{PW}^{1} D_{1}$ and $a_{PW}^{2} D_{2}$) are sufficient to provide an accurate estimation of HFEs for all other chemical classes from the training set. Figure 2 shows the differences between values of HFEs calculated with these corrections and the corresponding experimental data against $\Delta \mu_{\text{hyd}}^{exp}$ for the whole training set of solutes. The differences for all classes of solutes (except alkanes) are biased with respect to zero. Each class of solutes has its own bias, but the standard deviation inside of the most of the classes is small (Figure 2). Thus, we supposed that the bias for each chemical class can be removed by
Figure 2: The difference between calculated $\Delta \mu_{hyd}^{(2)}$ and experimental $\Delta \mu_{hyd}^{exp}$ HFEs (where $\Delta \mu_{hyd}^{(2)} = \Delta \mu_{hyd}^{PW} + a_1^{PW} D_1 + a_2^{PW} D_2 + a_0^{PW}$) versus $\Delta \mu_{hyd}^{exp}$ for the training set of solutes. Dashed lines indicate mean values of the difference. Arrows indicate the biases of the mean values of corresponding molecular set with respect to zero.

the use of the appropriate fragment correction. From this observation we may conclude that introducing one of the functional groups for each class of solutes introduces a constant error in the HFE calculated with RISM approach. This result reveals the hidden systematic errors in the RISM method due to which the values of HFE are over- or underestimated for solutes with different functional groups.

Previously it was found that the RISM method considerably overestimates the specific interactions of solutes with water (e.g. H-bonds formation) that results in too low values of corresponding HFEs.\textsuperscript{63,65,99} For alcohols this drawback of the RISM method is illustrated by Figure 2 (green triangles). However, the phenol fragment effect on HFE is not so clear. From the Figure 2 (red diamonds) it is obvious that the phenol fragment contribution can not be treated as a sum of OH-group and benzene ring contributions. Thus, we had to introduce a descriptor for the number of phenol fragments ($N_{ph}$). The detailed analysis of the correction for the phenol fragment is shown in the section \textbf{The SDC Model Coefficients}.

Therefore, to obtain a high accuracy predictions for HFE calculations for a given set of solutes we introduced a number of fragment descriptors associated with specific solute structures: number of branches in a carbon skeleton ($N_{br}$), number of double bonds ($N_{db}$), number of benzene rings
Figure 3: Structural descriptors illustrated on the representatives of chemical classes used in this work: a) alkenes; b) alkanes; c) haloalkanes; d) alcohols; e) alkylbenzenes; f) ethers; g) aldehydes; h) phenols; i) ketones. Gray balls are carbons, white balls are hydrogens, red balls are oxygens, green ball is chlorine atom. The colormaps illustrate approximate water density distribution around the molecules, reconstructed from the 1D RDFs (see Support Information for the details).

\( N_{bz} \), number of hydroxyl groups \( (N_{OH}) \), number of phenol fragments \( (N_{ph}) \), number of halogen atoms \( (N_{Hal}) \), number of ether groups \( (N_{eth}) \), number of aldehyde groups \( (N_{ald}) \), and number of ketone groups \( (N_{ket}) \). Thus, for the given set of solutes (see [Hydration Free Energy Data Sets]) the second term of the SDC model (Eq. (17)) consists of ten structural descriptors: one is a whole-molecule type descriptor\(^{79}\) (the DPMV descriptor) and nine fragments descriptors. Representatives for each of given class of solutes and their structural features requiring corrections are presented on Figure 3. As the RISM approach takes into account the molecular details of the solvent structure, one can see that each of specified structural features changes the solvent distribution around the solute (Figure 3).
Computational Details

The HFEs were calculated with the RISM method using the home-made collection of numerical routines developed by our group.\textsuperscript{56,100} Calculations were performed for the case of infinitely diluted aqueous solutions at T=300K. We used the Lue and Blankshtein version of the SPC/E model of water (MSPC/E).\textsuperscript{101} It differs from the original SPC/E water model\textsuperscript{102} by the additional LJ potential parameters for the water hydrogen, which were modified to prevent possible divergence of the algorithm.\textsuperscript{52,60,63,69} We took the MSPC/E bulk solvent correlation functions from the work\textsuperscript{103} where they are calculated by RISM equations for solvent-solvent correlations\textsuperscript{45} using wavelet-based algorithms.\textsuperscript{104,105}

RISM equations were numerically solved by the Fourier iterative algorithm on a finite mesh.\textsuperscript{100} Solutions were discretized on the uniform grid with 4096 grid points and grid step 0.05 Bohr (0.0265Å). The multi-grid technique was used to speedup the calculations.\textsuperscript{106,107} Calculations were performed with $L_2$-norm accuracy $\varepsilon = 10^{-4}$, which corresponds to the numerical error in the HFE of about 0.01 kcal/mol for non-polar solutes, and to the numerical error $\approx 0.025$ kcal/mol for the polar solutes (see Appendix 1).

To perform the calculations one needs three sets of input data: 1) atoms coordinates, 2) partial charges of atoms, and 3) atom-atom potential parameters representing the Van der Waals interactions. Coordinates for linear alkanes, several alkylbenzenes and phenols were taken from the Cambridge Structural Database.\textsuperscript{108} Due to the fact that hydrogen positions determined by standard X-ray methods differ systematically from those determined by neutron methods\textsuperscript{109} we optimized the length of the carbon-hydrogen bonds (C-H) using QM (quantum mechanical) energy minimization at the MP2/6-311G(d,p) level of theory with constrained bonds between heavy atoms (e.g. C-C). The geometrical parameters of all other solutes (not presented in the Cambridge Structural Database) were found by the structural optimization at the same level of theory but without geometrical constrains for the bond lengths between heavy atoms. For QM calculations we used Gaussian03 quantum chemistry software.\textsuperscript{110} We modeled all compounds with OPLS-AA (Optimized Potential for Liquid Simulations - All Atom) partial charges and LJ potential parame-
These parameters were assigned to each atom automatically by the Maestro software (the Schroedinger Inc.). We note, that the convergence of the RISM calculations with original geometric rules for OPLS-AA is very poor (see Supporting Information). To avoid this problem with convergence we did our calculations with the Lorentz-Berthelot mixing rule for the solute-water LJ potential parameters, similar to work. In the present work for calculations of the LJ potential parameters for interactions between solute sites and water hydrogens we used the following LJ parameters for water hydrogen: $\sigma_{Hw} = 0.8\text{Å}$ and $\epsilon_{Hw} = 0.046$ kcal/mol.

The set of structural descriptors was assigned to each molecules automatically using the computer program "checkmol".

**Results and Discussion**

**The SDC Model Coefficients**

Values of coefficients $a_i^{PW}$ of the SDC model (Eq. (17)) with the considered set of descriptors (see the section choice of descriptors) were obtained by the multiple linear regression analysis with the training set of solutes (see the section Hydration Free Energy Data Sets). Regression analysis was performed with the function regress from Matlab Statistics Toolbox (MATLAB version 7.8.0.347(R2009a). The MathWorks Inc., 2009). Errors of the regression coefficients were estimated as a square root of their variation. Results are shown in Table 1. All determined coefficients have the same order of magnitude showing that each descriptor from the considered set is significant.

In Figure 4 the SDC model corrections together with experimental and calculated HFEs are shown for two solutes: non-polar alkane (2,3-dimethylpentane) and polar alcohol (2-methylpentan-3-ol). For the simplicity of comparison these two solutes were chosen in such a way that they have almost the same structure but different side groups at the third carbon atom. As one can see, the HFEs calculated with PW method (Figure 4, upper grey bars) are overestimated for both solutes: alkane with positive $\Delta \mu_{hyd}^{exp}$ and alcohol with negative $\Delta \mu_{hyd}^{exp}$ (Figure 4, red bars). As it was
Figure 4: Hydration free energies and structural corrections of the SDC model for alkane (2,3-dimethylpentane) and alcohol (2-methylpentan-3-ol). Red bars are experimental data, blue bars are HFEs calculated with the SDC model, grey bars are contributions of the SDC model. Depicted structural corrections can be presented as products of coefficients and corresponding descriptors (e.g. DPMV correction equals \( a_1^{PW} D_1 \)) (see Eq. (17) and Table 1). Values of dimensionless descriptors are given in the inset table. Values of required coefficients \( a_1^{PW} \), \( a_2^{PW} \), and \( a_5^{PW} \) are presented in Table 1.

Table 1: Descriptors of the SDC model and the corresponding multiple regression coefficients for PW method.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Coefficient (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensionless partial molar volume</td>
<td>( a_1^{PW} = -1.51 \pm 0.01 )</td>
</tr>
<tr>
<td>Number of branches</td>
<td>( a_2^{PW} = 1.07 \pm 0.05 )</td>
</tr>
<tr>
<td>Number of double bonds</td>
<td>( a_3^{PW} = -0.92 \pm 0.20 )</td>
</tr>
<tr>
<td>Number of benzene rings</td>
<td>( a_4^{PW} = -1.70 \pm 0.15 )</td>
</tr>
<tr>
<td>Number of OH-groups</td>
<td>( a_5^{PW} = 0.73 \pm 0.14 )</td>
</tr>
<tr>
<td>Number of phenol fragments</td>
<td>( a_6^{PW} = -1.52 \pm 0.22 )</td>
</tr>
<tr>
<td>Number of halogen atoms</td>
<td>( a_7^{PW} = -2.10 \pm 0.16 )</td>
</tr>
<tr>
<td>Number of ether groups</td>
<td>( a_8^{PW} = -1.69 \pm 0.19 )</td>
</tr>
<tr>
<td>Number of aldehyde groups</td>
<td>( a_9^{PW} = -0.91 \pm 0.20 )</td>
</tr>
<tr>
<td>Number of ketone groups</td>
<td>( a_{10}^{PW} = -2.44 \pm 0.20 )</td>
</tr>
<tr>
<td>Correlation coefficient of multiple regression</td>
<td>( R^2 = 0.9472 )</td>
</tr>
</tbody>
</table>
shown above (see the section Choice of descriptors) the major part of the difference between $\Delta \mu_{\text{hyd}}^{PW}$ and $\Delta \mu_{\text{hyd}}^{\text{exp}}$ can be eliminated with the $a_0^{PW}$ and $a_1^{PW}D_1$ corrections. In turn, structural corrections are required to increase the accuracy of HFE calculations by removing other hidden systematic errors. Thus, the OH-group correction ($a_5^{PW}D_5$) has positive value and compensates the overestimation of the strengths of specific interactions between the OH-groups of the polar solute and water molecules. For the branched solutes it is also necessary to take into account the branches correction ($a_2^{PW}D_2$).

Results of HFE calculations with the SDC model for the whole training set of solutes are shown on Figure 5. Correlation coefficient between $\Delta \mu_{\text{hyd}}^{\text{SDC}}$ and $\Delta \mu_{\text{hyd}}^{\text{exp}}$ equals 0.9870 showing that the SDC model with small set of structural descriptors can accurately describe HFEs of 65 solutes with different chemical nature (see Table 2).

Figure 5: (a) Correlation between calculated and experimental HFEs for the training set of solutes, (b) Difference between $\Delta \mu_{\text{hyd}}^{\text{SDC}}$ and $\Delta \mu_{\text{hyd}}^{\text{exp}}$ versus experimental HFEs for the training set. Dashed lines indicate the corresponding standard deviation (see Table 2).

**The SDC model predictive ability**

The predictive ability of the SDC model was analyzed using the test set of 120 solutes and the same set of coefficients from Table 1 as for the training set. Comparison of predicted and experimental HFEs is shown on Figure 6. As we previously noted the test set contains 60 polyfragment
solute\(s\) (see the section Hydration Free Energy Data Sets). Among them there are dienes, dihy- 
dric alcohols, unsaturated aliphatic alcohols, styrenes, phenyl alcohols, di- and polychloroalkanes, 
chlorobenzenes, chlorophenols, hydroxybenzaldehydes, alkenals, alkoxyphenones, oxylcohols, 
phenylethers, and alkoxyphenols. As one can see (Figure 6) the proposed SDC model allows to 
predict HFEs of "simple" solutes with high accuracy. Details of the SDC model statistical profile 
are presented in Table 2.

In the case of polyfragment solutes predictability of the SDC model is sensitive to the chemical 
nature of solutes. One can see on Figure 6 several polyfragment solutes for which the difference 
\(\Delta \mu_{\text{SDC hyd}} - \Delta \mu_{\text{exp hyd}}\) exceeds the standard deviation. Some of them are small chloroalkanes \((N_{Hald} \text{ equals } 2-5)\). Others are chlorobenzenes with 3 or 4 atoms. We suppose that the main reason of this devi- 
ation is the fact that OPLS-AA partial charges do not take into account redistribution of electron 
density around electronegative groups. Indeed, the ability to describe properties of polyfragment 
solutes with OPLS-AA charges can be analyzed on phenols as an example. As it was shown above 
(see the section Choice of descriptors) the phenol fragment contribution can not be treated as a 
plain sum of the OH-group and benzene ring contributions. Results of the regression analysis 
confirm that even taking into account all required corrections (DPMV, branches, benzene ring, and 
OH-group corrections) the difference \(\Delta \mu_{\text{calc hyd}} - \Delta \mu_{\text{exp hyd}}\) for phenols is sufficiently biased with respect 
to zero (it is about 2 kcal/mol). We attribute this effect to the oversimplified character of OPLS-
AA partial charges for phenol. It is well known that substitution of a benzene hydrogen to an 
OH-group influences the electron density distribution and, correspondingly, the partial charges of 
carbon atoms in phenols.\(^{115}\) The OPLS-AA partial charges (which were used in the present work) 
take into account only the change of the partial charge of the carbon atom closest to OH-group 
(it has \(q = 0.150\) instead of \(q = -0.115\) for benzene’s carbons).\(^{111}\) All other atoms in the benzene 
ring have the same parameters as for the "neat" benzene (see Fig. S-2 of the Support Informa-
tion). However, partial charges of other carbon atoms also change because of the electron density 
redistribution, and, as a result, meta—, orto—, and para—positions in the phenol ring become dis-
tinguishable.\(^{115}\) OPLS-AA partial charges do not reflect these details. Thus, phenols properties are
not described in a proper way. Consequently, we included the subset of phenols into the training set and introduced additional correction for the number of phenol fragments \( (a_6^{PW} D_6) \), where \( D_6 \) is the total number of phenol fragments). However, the OPLS-AA partial charges perform satisfactory for phenyl alcohols containing a carbon spacer between the benzene ring and the OH-group (see Support Information). In this case the HFE correction can be approximated as a sum of the benzene ring and the OH-group contributions with the additional DPMV correction for the spacer.

One can see that for the rest of polyfragment solutes the deviation of predicted HFEs are comparable for those of "simple" from the test set. That means that coefficients of the model determined with the training set of "simple" solutes are transferable to polyfragment solutes. This indicates a great potential of the SDC model for HFE predictions of various organic molecules. We think that the quality of the model can be farther improved by reparametrization this model for QM-derived partial charges which is the subject of our future work.

![Figure 6](image_url)

**Figure 6:** (a) Correlation between experimental and predicted HFEs for the test set of solutes, (b) Difference between \( \Delta \mu_{SDC}^{\text{hyd}} \) and \( \Delta \mu_{\text{exp}}^{\text{hyd}} \) versus experimental HFEs for the test set. Dashed lines indicate the corresponding standard deviation.

To analyze the accuracy of the data obtained with the SDC model we calculated the mean values, standard deviation (std) of the difference \( \Delta \mu_{\text{hyd}}^{SDC} - \Delta \mu_{\text{hyd}}^{\text{exp}} \) for both training and test sets (Table 2). As one can see, even with polychloroalkanes and chlorobenzenes the std does not exceed 1.24 kcal/mol for the test set of solutes. We analyzed whether this difference is biased with respect to zero. For this purpose we calculated mean value of the difference \( \Delta \mu_{\text{hyd}}^{SDC} - \Delta \mu_{\text{hyd}}^{\text{exp}} \).
Table 2: Composition of the training and test sets by chemical classes. Statistical profile of the SDC model: mean value and standard deviation (std) of the difference $\Delta \mu_{SDC}^{hyd} - \Delta \mu_{exp}^{hyd}$ for the training and test sets of solutes (kcal/mol).

<table>
<thead>
<tr>
<th>No.</th>
<th>Chemical class</th>
<th>Fragment$^a$</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>mean</td>
</tr>
<tr>
<td>1</td>
<td>alkanes</td>
<td>R-R(R$_n$)</td>
<td>11</td>
<td>1.1e-17</td>
</tr>
<tr>
<td>2</td>
<td>alkenes</td>
<td>R=R</td>
<td>6</td>
<td>-7.4e-17</td>
</tr>
<tr>
<td>3</td>
<td>alkylbenzenes</td>
<td>Ph(R$_m$)</td>
<td>6</td>
<td>-1.2e-15</td>
</tr>
<tr>
<td>4</td>
<td>monohydric alcohols</td>
<td>R(OH)</td>
<td>8</td>
<td>3.3e-16</td>
</tr>
<tr>
<td>5</td>
<td>phenols</td>
<td>Ph(R$_l$)(OH)</td>
<td>5</td>
<td>-1.8e-16</td>
</tr>
<tr>
<td>6</td>
<td>chloroalkanes</td>
<td>R-Hal</td>
<td>10</td>
<td>1.1e-16</td>
</tr>
<tr>
<td>7</td>
<td>aldehydes</td>
<td>R-CHO</td>
<td>6</td>
<td>-1.5e-16</td>
</tr>
<tr>
<td>8</td>
<td>ketones</td>
<td>(R$_2$)C=O</td>
<td>6</td>
<td>-6.7e-16</td>
</tr>
<tr>
<td>9</td>
<td>ethers</td>
<td>R-O-R</td>
<td>7</td>
<td>3.5e-16</td>
</tr>
<tr>
<td>6</td>
<td>Polyfragment Solutes$^b$</td>
<td>c</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>TOTAL:</td>
<td></td>
<td>65</td>
<td>8.2e-17</td>
</tr>
</tbody>
</table>

$^a$ [R= alkyl; n = 2,3; m = 1...6; l = 1...5];

$^b$ dienes, dihydric alcohols, unsaturated aliphatic alcohols, styrenes, phenyl alcohols, di- and polychloroalkanes, chlorobenzenes, chlorophenols, hydroxybenzaldehydes, alkenals, alkoxyphenones, oxyalcohols, phenylethers, alkoxyphenols;

$^c$ combinations of the previous fragments

the test set of solutes it equals -0.55 ± 0.11 kcal/mol (in this case and in the rest of the paper the error of mean is std divided by square root of number of solutes). As one can see, the difference between experimental and predicted data is slightly biased, and the accuracy of the predicted HFEs depends mostly on the deviation of calculated data.

We compared the accuracy of HFEs calculated with different RISM expressions (see the section [Hydration Free Energy Expressions in RISM]). The mean values and std of the difference $\Delta \mu_{hyd}^{method} - \Delta \mu_{hyd}^{exp}$ for each method are presented in Figure 7. As one can see the HNC and KH expressions for HFE calculations (Eq. (8)) provide significantly overestimated values. The bias of the difference $\Delta \mu_{hyd}^{HNC} - \Delta \mu_{hyd}^{exp}$ with respect to zero is the major contribution to the error (mean value of the difference is about 38.8 kcal/mol). However, the bias removing is not sufficient to get a reasonable accuracy of HFE calculations with HNC method because standard deviation of the data is considerable. HFEs calculated with PW method (Eq. (11)) are also biased with respect to experimental data (mean value of $\Delta \mu_{hyd}^{PW} - \Delta \mu_{hyd}^{exp}$ is 10.91 ± 0.20 kcal/mol) but they are char-
characterized with noticeably less deviation. Thus, removing the bias of $\Delta \mu_{\text{hyd}}^{\text{PW}} - \Delta \mu_{\text{hyd}}^{\text{exp}}$ provides an ability to predict HFEs with high accuracy. Application of the PWC model for the correction of PW data leads to the decrease of the difference between experimental and calculated values (mean value of $\Delta \mu_{\text{hyd}}^{\text{PWC}} - \Delta \mu_{\text{hyd}}^{\text{exp}}$ is $0.63 \pm 0.18$ kcal/mol) but only weakly affects std. In turn, the PW data corrected with the SDC model are less biased with respect to experimental values (see above), and the standard deviation of these results is 1.5 times less than that for the PW method. Thus, the SDC model improves the quality of the initial PW model and considerably reduces the error of HFE calculation.

The results in Figure 7 show that the HFEs calculated with the GF expression (Eq. (9)) are less biased with respect to the experimental values than the PW HFEs but have the standard deviation which is about 1.5 times higher as that for the PW data. We applied the SDC model for the GF data as well (see details in the Supporting Information). It was shown that the combination of the SDC method with the PW HFE expression performs better for prediction of HFEs for "simple" organic

<table>
<thead>
<tr>
<th>Method</th>
<th>mean</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>HNC</td>
<td>38.78</td>
<td>13.03</td>
</tr>
<tr>
<td>HNCB</td>
<td>-6.56</td>
<td>3.79</td>
</tr>
<tr>
<td>KH</td>
<td>38.83</td>
<td>12.97</td>
</tr>
<tr>
<td>GF</td>
<td>2.30</td>
<td>3.08</td>
</tr>
<tr>
<td>PW</td>
<td>10.91</td>
<td>2.19</td>
</tr>
<tr>
<td>PWC</td>
<td>0.63</td>
<td>1.95</td>
</tr>
<tr>
<td>SDC</td>
<td>-0.55</td>
<td>1.24</td>
</tr>
</tbody>
</table>
solutes. The standard deviation of $\Delta \mu_{\text{hyd}}^{GF+SDC} - \Delta \mu_{\text{hyd}}^{exp}$ for them is about 60% larger than that for the combination of PW and SDC methods. However, performance of the GF/SDC for polyfragment solutes is practically the same as for the PW/SDC model (see Supporting Information).

Table 3: Statistical profile of the SDC model and chemoinformatics for "simple" and polyfragment solutes from the test set

<table>
<thead>
<tr>
<th></th>
<th>SDC model</th>
<th>Chemoinformatics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>“simple” solutes (N = 60)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.06</td>
<td>-0.03</td>
</tr>
<tr>
<td>std</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>max</td>
<td></td>
</tr>
<tr>
<td><strong>polyfragment solutes (N = 60)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>-1.15</td>
<td>-3.10</td>
</tr>
<tr>
<td>std</td>
<td>1.44</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>max</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL test set (N = 120)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>-0.55</td>
<td>-1.57</td>
</tr>
<tr>
<td>std</td>
<td>1.24</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td>max</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: Predictability of the SDC model and chemoinformatics for the polyfragment solutes from the test set. Solid lines correspond to the mean values of the difference between calculated and experimental data. Dashed lines indicate the corresponding standard deviation of the difference.

We show that the SDC model yields more accurate HFE predictions with respect to other
RISM-based HFE expressions: GF, PW, HNCB, KH, and HNC. However, the biased data obtained by the PW calculations (Figure 7) lead to some doubts that the RISM approach can be a good starting point for the HFE calculations. To clarify this question, we compare the data obtained by the SDC model with those derived from pure chemoinformatics calculations based on the same descriptors (see Table 3 and Figure 8). Indeed, in the case of "simple" solutes the RISM approach does not provide a significant improvement with respect to that obtained by the chemoinformatics (Table 3). However, the situation changes drastically for polyfragment solutes, for which the pure chemoinformatics calculations are significantly less efficient. Figure 8 indicates that the structure of errors for the SDC model and the pure chemoinformatics calculations is similar, the errors increase as the absolute value of HFE decreases. However, the absolute values of errors obtained within the chemoinformatics framework are twice larger than that calculated with the use of the RISM. Therefore, we conclude that the RISM-based SDC model is efficient for accurate HFE predications for a wide set of solutes.

Conclusions

Results of the present study reveal the hidden systematic errors in the 1D RISM method with standard HFE expressions (Eq. (8))–(Eq. (11)) and HNC/PLHNC closure relations due to which the values of HFE are over- or underestimated for solutes with different functional groups. There are two main sources of these errors: (i) oversimplified form of the bridge approximation in HNC/PLHNC closure relations; (ii) replacement of multidimensional molecular correlations functions by a combination of spherically symmetric functions. Although there has been done certain progress towards the improvement of the theoretical background of the integral equations by introducing more sophisticated forms of bridge functionals and 3D extension of RISM as well as combining RISM with QM methods, we are not aware of any integral equations theory method capable to provide accurate estimations of HFEs for a wide set of organic compounds.

We showed that one can eliminate the systematic errors of HFEs calculated with RISM ap-
proach by a set of empirical corrections. The major part of the difference \( \Delta \mu_{\text{hyd}}^{\text{calc}} - \Delta \mu_{\text{hyd}}^{\exp} \) can be eliminated with the \( a_{\text{PW}0} \) and \( a_{\text{PW}1} \) corrections. The \( a_{\text{PW}0} \) correction eliminates a general systematic error of the RISM model which does not depend on the solute properties. In turn, structural corrections associated with structural fragments of the molecular solutes are required to increase the accuracy of HFE calculations by removing other hidden systematic errors. The resulted SDC model (Eq. (17)) containing the small set of structural descriptors predicts reasonably well the HFEs of small organic molecules with different chemical nature. We note, that the obtained standard deviation of the error around 1.2 kcal/mol is comparable with the accuracy of HFE predictions by more expensive molecular mechanics methods which also give accuracy of HFE predictions around 1 kcal/mol.^{32,33}

The theoretical foundation of the SDC model is quite general and, in principal, this model can be used for correction of results of any HFE computational model. However, we have found that the efficiency of this hybrid method depends on the theoretical origin of the initial model. The combination of the SDC model with RISM/PLHNC calculations of correlation functions and PW expression for HFE provides the best results within the models considered in our study (see details in the Supporting Information). However, the search for the best combination of methods is far from being complete and we plan to make more investigations of this subject in our future works.

Our results show that once calculated the SDC model correlation coefficients can be transferred between different classes of solutes. This indicates a great potential of the method for HFE predictions of various organic molecules. We also note, that the method can be combined in a relatively straightforward manner with quantum mechanical description of the solute molecules via one of the RISM/QM schemes described in the literature.\(^{45,118–121,126}\) Our future research will be focused on the extension of this method to HFE calculations of bioactive compounds.

**Acknowledgement**

The authors would like to acknowledge the use of the Chemical Database Service at Daresbury. Maxim V. Fedorov and Gennady G. Chuev acknowledge the financial support from the FP7-
PEOPLE-IIF-2008 program (Grant No 235064, Programme: People). We are thankful to Andrey Frolov and Nikolaj Georgi for critical reading of the manuscript and useful discussions.

Appendix 1. Details of the RISM calculations

The set of the RISM equations (Eq. (3), Eq. (4)) was solved by the standard numerical iterative scheme using Bessel-Fourier transforms for the calculation of the convolution integrals. For speeding-up the iterations the multi-grid technique was used. Six levels of numerical grids were employed for the calculations. The coarsest grid, where the most of the iterations were done, had 128 grid points and grid-step of 0.4 Bohr (0.212 Å). The solution was obtained on the finest grid, which had 4096 grid points, grid step was 0.05 Bohr (0.0265 Å) and cutoff distance was 204.8 Bohr (108.4 Å). The accuracy of the iterations was controlled by the norm of difference between the solutions on the sequential iterations (Eq. (19)). Iteration process was stopped when the accuracy of n-th iteration had reached the threshold $\varepsilon_{\text{thres}}$: 

$$\Delta_n < \varepsilon_{\text{thres}}.$$  

where $h^{(n)}_{s\alpha}(r), c^{(n)}_{s\alpha}(r), h^{(n+1)}_{s\alpha}(r), c^{(n+1)}_{s\alpha}(r)$ are the total and direct correlation functions approximations on the n-th and $(n+1)$-th iteration steps respectively.

In the current work the RISM equations were solved up to the accuracy $\varepsilon_{\text{thres}} = 10^{-4}$. To check, whether this accuracy is sufficient for the accurate HFE calculations additional numerical experiments were performed. It was shown, that for 10 randomly chosen non-polar compounds the numerical error of the RISM HFE calculations with PW method is about 0.008 kcal/mol. For polar compounds the calculation error is approximately 0.024 kcal/mol. These errors are essentially lower than typical error of experimental HFE measurements (approximately 0.24 kcal/mol). Therefore, we assume that the numerical accuracy $\varepsilon_{\text{thres}} = 10^{-4}$ is sufficient.
Supporting Information Available

This material is available free of charge via the Internet at http://pubs.acs.org/.

References


(100) Fedorov, M. V.; Flad, H. J.; Chuev, G. N.; Grasedyck, L.; Khoromskij, B. N. Computing 2007, 80, 47–73.


