Fast and General Method to Predict the Physico-Chemical Properties of Druglike Molecules using the Integral Equation Theory of Molecular Liquids

David S. Palmer · Maksim Mišin · Maxim V. Fedorov · Antonio Llinas

Received: date / Accepted: date

Abstract We report a method to predict physico-chemical properties of druglike molecules using a classical statistical mechanics based solvent model combined with machine learning. The RISM-MOL-INF method introduced here provides an accurate technique to characterize solvation and desolvation processes based on solute-solvent correlation functions computed by the 1D Reference Interaction Site Model of the Integral Equation Theory of Molecular Liquids. These functions can be obtained in a matter of minutes for most small organic and druglike molecules using existing software (RISM-MOL) (Sergiievskyi, V. P.; Hackbusch, W.; Fedorov, M. V. *J. Comput. Chem.* **2011**, *32*, 1982-1992.). Predictions of caco-2 cell permeability and hydration free energy obtained using the RISM-MOL-INF method are shown to be more accurate than the state-of-the-art tools for benchmark datasets. Due to the importance of solvation and desolvation effects in biological systems, it is anticipated that the RISM-MOL-INF approach will find many applications in biophysical and biomedical property prediction.

David S. Palmer
Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland G1 1XL, UK
Tel.: +44-141-5484178
E-mail: david.palmer@strath.ac.uk

Maksim Misin
Scottish Universities Physics Alliance (SUPA), Department of Physics, University of Strathclyde, John Anderson Building, 107 Rottenrow, Glasgow, Scotland G4 0NG, UK

Maxim V. Fedorov
Scottish Universities Physics Alliance (SUPA), Department of Physics, University of Strathclyde, John Anderson Building, 107 Rottenrow, Glasgow, Scotland G4 0NG, UK

Antonio Llinas
AstraZeneca R&D, Respiratory, Inflammation and Autoimmune iMed, Pepparedsleden 1, SE-431 83, Mölndal, Sweden

1 Introduction

The Integral Equation Theory (IET) of Molecular Liquids is a promising theoretical framework for modeling solvent in biomolecular simulations.[1] IET is based on the molecular Ornstein-Zernike (MOZ) equation, which allows the density distribution of solvent molecules around a solute to be calculated from a set of integral equations and a closure relationship, without the need for long molecular dynamics or Monte Carlo simulations. Since IET methods operate with averaged quantities (total and direct solute-solvent correlation functions) and they do not spend computer time on the averaging procedure, they can treat an infinite number of solvent molecules and the results are almost independent of the size of the simulation cell and free from the statistical noise which causes problems in numerical simulation. IET has found an ever-increasing number of successful applications in the last few years, including computing solubility of druglike molecules[2], fragment-based drug design[3], modelling the binding of water[4] and ions[5] by proteins, predicting tautomer ratios[6], interpreting solvent densities around biomacromolecules[7], and sampling molecular conformations[8]. Here we propose a general method to compute solution-phase properties of druglike molecules using IET combined with statistical or machine learning algorithms. Proof-of-concept results are provided for the prediction of caco-2 cell permeability and hydration free energy for which the new method is shown to perform better than many state-of-the-art tools.

Caco-2 cells originate from an intestinal cell line derived from human colorectal carcinoma that under the correct conditions exhibits many of the same properties as the eneterocytes lining the small intestine.[9] For druglike molecules, permeability measured across a monolayer of caco-2 cells has been shown to be well correlated with human oral absorption,[10,11] which means that caco-2 cell permeability assays are useful in vitro models of in vivo absorption behaviour.[9] Experimental measurements and computational predictions of caco-2 cell permeability are now widely used in early stage drug discovery as a means to assess oral absorption and to guide medicinal chemists, with computational methods having the advantage that they can be applied to virtual libraries of compounds prior to their synthesis. There are two important routes for cell permeation: passive diffusion and carrier-mediated influx via active transport mechanisms. Passive diffusion can be further classified as occurring by either paracellular or transcellular routes.

Hydration free energy (ΔG_{hyd} or HFE) - the change in free energy associated with transferring a molecule from gas phase to aqueous solution under standardised conditions - is one of the most important properties in solution chemistry.[12,13] Accurate computation of HFEs is of great interest because it is difficult and time-consuming to measure experimentally,[14,15] and because many physico-chemical and biomedical properties of molecules are defined by their solvation and acid-base behavior, which can be estimated from their HFEs.[1,16–18] For example, HFEs have been used in the calculation of aqueous solubility,[2,19] protein-ligand binding affinity,[20] acid-base dissociation constant (pKa)[21], and octanol-water partition coefficient[22–24], amongst others properties. Computed HFEs are also important for assessing the environmental impact of new organic molecules in consumer products (i.e pharmaceuticals, agrochemicals, etc).[25,26] Improving the accuracy of computational methods to calculate HFEs would have widespread benefits.

2 Theory

The method proposed here is based on properties that can be computed using the 1D Reference Interaction Site Model (1D RISM) proposed by Chandler et al. We begin with a description of the standard 1D RISM method before outlining our approach to physico-chemical property prediction.

2.1 1DRISM

The 1D RISM allows the thermodynamics of molecular solutions to be modelled by a set of integral equations and closure relationships.[27] In the following, we provide only a brief overview of the theory, a more thorough discussion is provided elsewhere.[1,28] In the RISM approach both the solute and the solvent molecules are treated as sets of sites with spherically-symmetric properties. In the simplest case, the sites are just the atoms of the molecules. Three types of site-site correlation functions are considered in the RISM: intramolecular correlation functions, total correlation functions and direct correlation functions. Intramolecular correlation functions describe the structure of the molecule. For the two sites, s and s' of one molecule, the intramolecular correlation function is:

$$\omega_{ss'}(r) = \frac{\delta(r - r_{ss'})}{4\pi r_{ss'}^2} \tag{1}$$

where $r_{ss'}$ is the distance between the sites and $\delta(r - r_{ss'})$ is the Dirac delta-function. Total correlation functions $h_{s\alpha}(r)$ and direct correlation functions $c_{s\alpha}(r)$ are defined for each pair of solute and solvent sites (s and α, respectively). The total correlation functions can be expressed as $h_{s\alpha}(r) = g_{s\alpha}(r) - 1$, where $g_{s\alpha}(r)$ is the radial distribution function of solvent sites around the solute sites. Bulk solvent total correlation functions $h_{\alpha\xi}^{\mathrm{bulk}}(r)$ are also considered, and represent the distribution of sites ξ of solvent molecules around the site α of the selected solvent molecule. Direct correlation functions $c_{s\alpha}(r)$ are calculated using the set of RISM equations for the case of infinitely diluted solution[28]:

$$\begin{array}{l} h_{s\alpha}(r) = \sum_{s'\xi} \left\langle \omega_{ss'} * c_{s'\xi} * [\omega_{\alpha\xi}^{\mathrm{bulk}} + \rho h_{\alpha\xi}^{\mathrm{bulk}}] \right\rangle (r) \\ s = 1, \ldots, N_{solute},\ \alpha = 1, \ldots, N_{solvent},\ r \in [0; \infty) \end{array} \tag{2}$$

Here $\langle x * y \rangle (r)$ is the radial part of the spherically symmetric three-dimensional convolution $\langle x * y \rangle (r) = \int_{R^3} x(\mathbf{r} - \mathbf{r}')y(\mathbf{r}')d\mathbf{r}'$, and ρ is a number density of the bulk solvent. To complete the set of RISM equations, one needs to use a closure relationship, which has the general form:

$$c_{s\alpha}(r) = e^{\Xi_{s\alpha}(r) - B_{s\alpha}(r)} - h_{s\alpha}(r) + c_{s\alpha}(r) - 1, \tag{3}$$

where $\Xi_{s\alpha}(r) = -\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r)$, $u_{s\alpha}(r)$ is the atom-atom potential, $B_{s\alpha}(r)$ is a so-called *bridge function*,[28,29] $\beta = 1/k_B T$, k_B is the Boltzmann constant, and T is the temperature. The case $B(r) \equiv 0$ corresponds to the frequently used Hypernetted Chain closure[28,29]. However, the RISM equations with Hypernetted Chain closure do not converge for many molecules of chemical interest[28,30,31]. Therefore, to improve convergence, the Partially-Linearized Hypernetted Chain closure (PLHNC) is often used instead:[32,33]

$$c_{s\alpha}(r) = \begin{cases} e^{\Xi_{s\alpha}(r)} - h_{s\alpha}(r) + c_{s\alpha}(r) - 1 & \Xi_{s\alpha}(r) < 0 \\ -\beta u_{s\alpha}(r) & \Xi_{s\alpha}(r) > 0 \end{cases} \tag{4}$$

The intramolecular correlation functions $\omega_{ss'}(r)$ can be found from equation (1). The total correlation functions of the bulk solvent $h_{\alpha\xi}^{\mathrm{bulk}}(r)$ are normally obtained by solution of the solvent-solvent 1D RISM equations.[28,34]

The set of RISM equations (Equation (2)), together with the closure relation (Equation (4)), allow one to find the functions $h_{s\alpha}(r)$ and $c_{s\alpha}(r)$, which are illustrated for an example molecule in Figure 1. There are no known methods to solve the set of RISM equations analytically in the general case. Thus, the RISM equations are commonly solved numerically. In the current work we use the RISM-MOL solver, which is a Matlab implementation of a multi-grid algorithm for solving RISM equations[35].

Within the RISM theory, there are several expressions which allow one to obtain values of the hydration free energy from the total and direct correlation functions $h_{s\alpha}(r)$ and $c_{s\alpha}(r)$. Here we discuss four of the most popular free-energy expressions.[36–39]

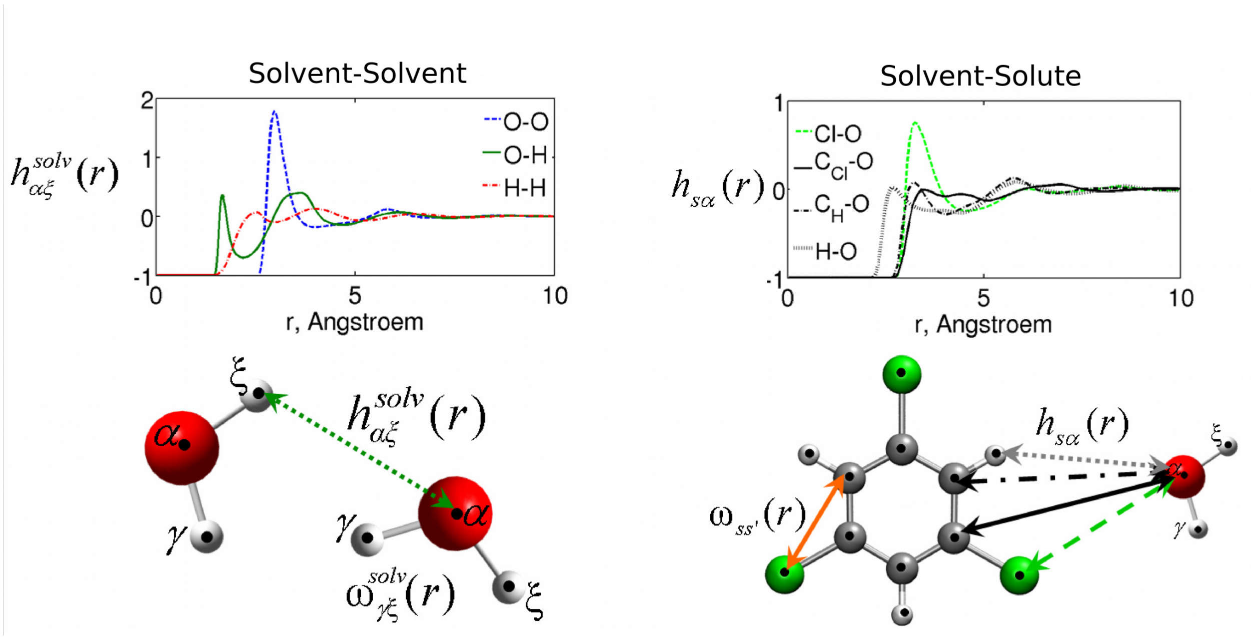


Fig. 1 Solvent-solvent and solute-solvent correlation functions in 1D RISM. $h(r)$ are total correlation functions, $\omega(r)$ are intramolecular correlation functions. Subscripts s and s' refer to solute sites (atoms), while greek letters refer to solvent sites (atoms).

The first expression is the Gaussian Fluctuations approximation (GF),[37,38,40] in which the free energy is given as:

$$\Delta G_{GF} = 2\pi\rho kT \sum_{s\alpha} \int_0^\infty \left(-2c_{s\alpha}(r) - c_{s\alpha}(r)h_{s\alpha}(r)\right) r^2 dr \tag{5}$$

The second hydration free energy equation we consider is the Kovalenko-Hirata expression (KH).[28]

$$\begin{aligned}\Delta G_{KH} =& 2\pi\rho kT \sum_{s\alpha} \int_0^\infty [-2c_{s\alpha}(r) - \\ & h_{s\alpha}(r)(c_{s\alpha}(r) - \Theta(-h_{s\alpha}(r)))]r^2 dr\end{aligned} \tag{6}$$

where $\Theta(x)$ is a Heaviside step function.

The third expression is the Hyper-Netted-Chain (HNC) approximation,[36] in which the formula for the hydration free energy is:

$$\begin{aligned}\Delta G_{HNC} =& 2\pi\rho kT \sum_{s\alpha} \int_0^\infty [-2c_{s\alpha}(r) - \\ & h_{s\alpha}(r)(c_{s\alpha}(r) - h_{s\alpha}(r))]r^2 dr\end{aligned} \tag{7}$$

The final hydration free energy expression is the Hyper-Netted-Chain with repulsive Bridge correction (*HNCB*), which has the form:[39]

$$\begin{aligned}\Delta G_{HNCB} =& \Delta G_{HNC} + \\ & 4\pi\rho kT \sum_{s\alpha} \int_0^\infty (h_{s\alpha}(r)+1)(e^{-B^R_{s\alpha}(r)}-1) r^2 dr\end{aligned} \tag{8}$$

The $\{B^R_{s\alpha}(r)\}$ in the HNCB functional are repulsive bridge functions, defined for each pair of solute s and solvent α atoms by the expression:

$$\exp(-B^R_{s\alpha}(r)) = \prod_{\nu\neq\alpha} \left\langle \omega^{\text{bulk}}_{\alpha\nu} * \exp\left(-\beta\varepsilon_{s\nu}\left(\frac{\sigma_{s\nu}}{r}\right)^{12}\right)\right\rangle \tag{9}$$

where $\omega^{\text{bulk}}_{\alpha\nu}(r)$ are the solvent intramolecular correlation functions, and $\sigma_{s\nu}$ and $\varepsilon_{s\nu}$ are the site-site parameters of the pair-wise Lennard-Jones potential.

When computing hydration free energies with either HNC of HNCB free energy functionals, the 1D RISM equations are normally solved using the HNC or HNCB closure relationships, respectively. The HNC closure is $B(r) \equiv 0$ as discussed previously. The HNCB closure has the following value when $\Xi_{s\alpha}(r) - B^R_{s\alpha}(r) < 0$:

$$c_{s\alpha}(r) = e^{\Xi_{s\alpha}(r)-B^R_{s\alpha}(r)} - h_{s\alpha}(r) + c_{s\alpha}(r) - 1 \tag{10}$$

whilst when $\Xi_{s\alpha}(r) - B^R_{s\alpha}(r) > 0$, it takes a value given by:

$$c_{s\alpha}(r) = -\beta u_{s\alpha}(r) - B^R_{s\alpha}(r) \tag{11}$$

Unfortunately, due to the many approximations inherent in 1D RISM theory, the standard 1D RISM free energy functionals (i.e GF, KH, HNC, or HNCB) give trivial results that are too inaccurate for most practical applications. However, it has previously been shown that more accurate predictions of hydration free energies can be obtained by combining RISM calculations with simple molecular descriptors in order to reduce systematic errors. Examples of these approaches include the partial wave correction[41], atomic descriptor correction (ADC)[42] and structural descriptor correction (SDC) models[43], which have a common functional form:

$$\Delta G_{corrected} = \Delta G_{uncorrected} + aV + \sum_i a_i x_i + c \tag{12}$$

where $\Delta G_{corrected}$ is the corrected hydration free energy, $\Delta G_{uncorrected}$ is the uncorrected hydration free energy computed by one of the previously mentioned free energy functionals (e.g. GF, KH, HNC, or HNCB), a and b_i are regression coefficients, V is the partial molar volume computed by 1D RISM (Equation 13), x_i is the count of a particular structural feature i, and c is a constant. The values of a, b_i and c may be obtained by multiple linear regression against experimental hydration free energy data. The success of these semi-empirical free energy functionals suggests that the total and direct correlation functions computed by 1D RISM contain useful information about solvation thermodynamics even if the standard free energy functionals do not give accurate results.

The excluded volume of a solute in infinitely dilute aqueous solution can be calculated as a limiting case of the partial molar volume[44] when the solute number density tends to zero:

$$V_{ex} = \frac{1}{\rho} + \frac{4\pi}{N_{solute}} \sum_s \int_0^\infty \left(h^{\text{bulk}}_{oo}(r) - h_{so}(r)\right) r^2 dr \tag{13}$$

where $h_{oo}^{bulk}(r)$ 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.

2.2 RISM-MOL-INF variables

The aim of this research is to develop a general method to predict the solution-phase properties of druglike molecules that are important in pharmaceutical research and development (here we test the method for the prediction of caco-2 cell permeability and HFE). Since solvation behaviour is known to be an important factor in determining the bioavailability of candidate drugs,[45] we hypothesise that solute-solvent correlation functions computed by 1D RISM combined with statistical or machine learning algorithms may provide a fast and general prediction method. More specifically, we propose that variables to quantify solvation and desolvation processes can be derived from the following functionals, which are based on the standard 1D RISM free energy functionals, but with the integration over r omitted.

Indeed, each of the functionals given in Equation 5 to Equation 8 can be re-written in a compact form:

$$\Delta G_{solv}^{RISM} = \int\limits_{0}^{\infty} w(r)dr, \tag{14}$$

where the integrand functionals $w(r)$ combine the $N_S \times N_\alpha$ total and direct correlation functions of a single solute into a single function of r. The form of one of these functionals ($gf\_w(r)$) is shown for four small organic molecules in Figure 2 (similar graphs for the $hncb\_w(r)$, $hnc\_w(r)$ and $kh\_w(r)$ functionals are provided in the Supporting Information).

$$gf\_w(r) = 2\pi\rho kT \sum_{s\alpha} [-2c_{s\alpha}(r) - c_{s\alpha}(r)h_{s\alpha}(r)] \tag{15}$$

$$\begin{aligned} kh\_w(r) =& 2\pi\rho kT \sum_{s\alpha}[-2c_{s\alpha}(r) - \\ & h_{s\alpha}(r)(c_{s\alpha}(r) - \Theta(-h_{s\alpha}(r)))] \end{aligned} \tag{16}$$

$$\begin{aligned} hnc\_w(r) =& 2\pi\rho kT \sum_{s\alpha}[-2c_{s\alpha}(r) - \\ & h_{s\alpha}(r)(c_{s\alpha}(r) - h_{s\alpha}(r))] \end{aligned} \tag{17}$$

$$\begin{aligned} hncb\_w(r) =& hnc\_w(r) + \\ & 4\pi\rho kT \sum_{s\alpha}[(h_{s\alpha}(r) + 1)(e^{-B_{s\alpha}^{R}(r)} - 1)] \end{aligned} \tag{18}$$

Each of the four functions considered here ($gf\_w(r)$, $hncb\_w(r)$, $hnc\_w(r)$ and $kh\_w(r)$) provide a very sensitive measure of the response of the solvent molecules to the solute.

We therefore hypothesise that a set of variables to quantify solvation and desolvation effects can be defined based on the numerical value of these functions at selected values of r . Statistical or machine learning algorithms will then be trained on these variables and the resulting model used to predict the property of interest, i.e. a molecular informatics based approach. We refer to this method as RISM-MOL-INF since it combines *RISM* with *MOL*ecular *INF*ormatics.

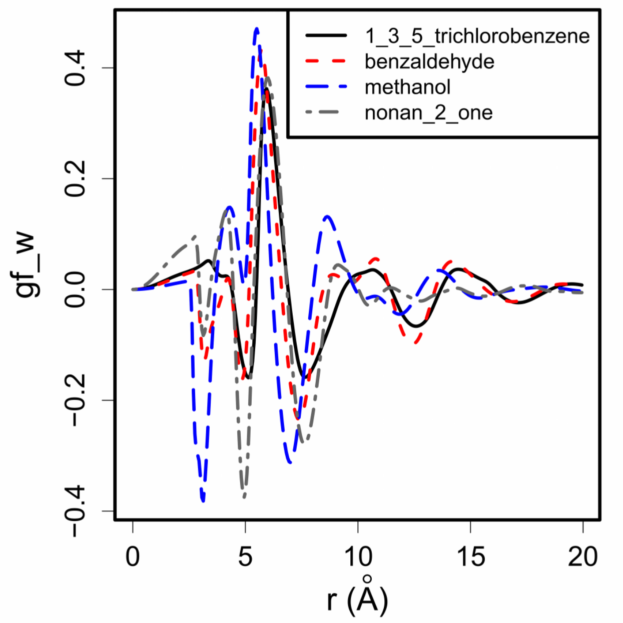


Fig. 2 The gf_w) function plotted for four different small organic molecules.

As an extra proof of our hypothesis, we refer to recent works that have used integrand functions in 3D RISM free energy functionals to study the detailed mechanisms of solute-solvent interactions and the association of molecules in complex solutions.[46–48] In these works the integrand functions in 3D RISM free energy functionals are used to determine the 3D Solvation Free Energy Density (3D-SFED) functions.[48] These functions can be used to characterize the intensity of the effective solute-solvent interactions in different 3D spatial regions at the solute surface and to indicate where these contribute the most/least to the entire solvation free energy.[47] Overall it was shown that 3D-SFED functions can provide useful information about different association/adsorption processes at complex biomolecule and supra-molecule aggregate surfaces.[47–49]

Since the total and direct correlation functions are normally represented on a fine grid when the 1D RISM equations are solved, the grid points provide a natural coordinate system in which to define the RISM-MOL-INF variables. However, the grid used to solve the 1D RISM equations has a narrower grid-spacing than is required for this purpose. For example, the total and direct correlation functions obtained here from the RISM-MOL program[35] are represented on a grid from $r = 0$ Å to $r = 120$ Å with grid-spacing, $\delta r = 0.00625$ Å. Simply taking the value of the selected function at each discrete grid point as a separate variable would lead to a large number of redundant variables for two reasons. Firstly, the variables corresponding to neighbouring grid points would be expected to be highly correlated. Secondly, grid points corresponding to $r \gtrsim 15$ would contain little useful information, since the total and direct correlation functions decay to single values for all molecules in this region. Therefore, for the purposes of testing, we work with variables defined at every tenth grid point from $r = 0$ Å to $r = 15$ Å. We denote each RISM-MOL-INF variable as m_w_n, where m is the 1D RISM free energy functional upon which the variables are based (i.e. m is gf, kh, hnc or $hncb$) and n is the grid point at which the variable is evaluated. The RISM-MOL-INF variables are given in lowercase to distinguish them from the standard 1D RISM free energy functionals, which are given in uppercase (e.g GF, HNCB, HNC, and KH). The grid point can be converted to a radial distance from solute site by $r = n \times \delta r$, where δr is the grid spacing. Hence, for example, the RISM-MOL-INF variable gf_w_1000 is computed from a function derived from the GF free energy functional at $r = 6.25$ Å. Since 1D RISM is known to overestimate excluded volume effects,[41] and this error has been shown to be correlated with the partial molar volume computed by 1D RISM,[26,41–43,50] we include this property as an additional descriptor in each of the sets of RISM-MOL-INF variables. The RISM-MOL-INF variables as defined here are not mutally orthogonal and, therefore, contain some redundant features. We have not attempted to make an *a priori* selection of a set of orthogonal descriptors for two reasons. Firstly, selecting a model from a pool of correlated descriptors is a standard problem in statistical modelling, for which there are many solutions.[51] Secondly, it is not possible to predict *a priori* what values of r will lead to the most useful variables for a given problem. Our decision to define a RISM-MOL-INF variable at every tenth grid point simply represents a balance between reducing the number of variables to allow efficient subset selection, while at the same time maintaining sufficient useful chemical information.

We use the RISM-MOL-INF variables as input to statistical or machine learning models to predict two important physico-chemical properties of organic molecules: hydration free energy and caco-2 cell permeability. The overall approach is similar to the cheminformatics models (i.e. Quantitative Structure-Property Relationships, (QSPRs)) that are widely used in pharmaceutical property prediction, but the RISM-MOL-INF variables characterize molecules by the effect they have on the density distribution of solvent molecules, rather than by molecular structure alone. Since our main aim is to demonstrate the potential of the RISM-MOL-INF variables, rather than to develop definitive QSPRs for either property, each model is trained on RISM-MOL-INF variables *only* and comparisons are made to selected state-of-the-art tools.

2.3 Statistical and Machine Learning Algorithms

Training the RISM-MOL-INF models involves finding a function that relates objects $x \in X$ and targets $y \in Y$ based solely on a sample $z = (x, y) = ((x_1, y_1), ..., (x_m, y_m)) \in (X \times Y)^m$ of size $m \in N$. In the following, the output space is a set of n real targets $Y \in R^n$, and the task is referred to as regression. We consider two different methods of regression: Partial-Least-Squares (PLS) and Random Forest (RF).

2.3.1 Partial-Least-Squares Regression

Partial least squares (PLS) is a method for linear regression that has been widely used in many different fields of research, including chemistry, biology, econometrics and social science. The PLS algorithm finds a linear regression model by projecting both the dependent and independent variables into a new mathematical space in which the covariance in the data structure can be explained by a small number of latent variables. As such PLS regression has some similarity to principal component regression, but the latent variables are selected for their ability to explain the variance in the dependent variable as well as in the independent variables. The algorithms used for PLS regression have been explained elsewhere.[52]

2.3.2 Random Forest

Random Forest is a method for classification and regression which was introduced by Breiman and Cutler.[53] The method is based upon an ensemble of decision trees, from which the prediction of a continuous variable is provided as the average of the predictions of all trees. Recent studies have suggested that Random Forest offers features which make it very attractive for statistical modelling studies.[54] These include relatively high accuracy of prediction, built-in variable selection, and a method for assessing the importance of each variable to the model.

In RF regression, an ensemble of regression trees is grown from separate bootstrap samples of the training data using the CART algorithm.[53] The branches in each tree continue to be subdivided while the minimum number of observations in each leaf is greater than a predetermined value. Unlike regression trees, the branches are not pruned back. Furthermore, the descriptor selected for branch splitting at any fork in any tree is not selected from the full set of possible descriptors but from a randomly selected subset of predetermined size. There are three possible training parameters for Random Forest: *ntree* - the number of trees in the Forest; *mtry* - the number of different variables tried at each split; and *nodesize* - the minimum node size below which leaves are not further subdivided.

The bootstrap sample used during tree growth is a random selection with replacement from the molecules in the data set. The molecules that are not used for tree growth are termed the *out-of-bag* sample. Each tree provides a prediction for its out-of-bag sample, and the average of these results for all trees provides an in situ cross-validation called the out-of-bag validation.

3 Methods

3.1 1D RISM calculations

The RISM-MOL program was used to solve the 1D RISM equations[35] using the dielectrically-consistent reference interaction site model (DRISM) formulism.[55,56] All calculations were performed assuming solute molecule in infinitely dilute aqueous solution. Using the HNC or HNCB closures to solve the 1D RISM equations often leads to convergence issues. Since we would like the method developed here to be as robust as possible, we use the numerically more stable PLHNC closure for all 1D RISM calculations.

3.1.1 Solvent Parameters

Solvent molecules were modeled using the Lue and Blankschtein version of the SPC/E model of water (MSPC/E).[57] This differs from the original SPC/E water model[58] by the addition of modified Lennard-Jones (LJ) potential parameters for the water hydrogen, which were altered to prevent possible divergence of the algorithm.[39,41,59,60] The Lorentz-Berthelot mixing rules were used to generate the solute-water LJ potential parameters[61], i.e. $\sigma_{s\alpha} = (\sigma_s + \sigma_\alpha)/2$ and $\epsilon_{s\alpha} = \sqrt{\epsilon_s \epsilon_\alpha}$. The following LJ parameters (for water hydrogen) were used to calculate the interactions between solute sites and water hydrogens: $\sigma^{LJ}_{H_w} = 1.1657$Å and $\epsilon^{LJ}_{H_w} = 0.0155$ kcal/mol.

3.1.2 Solute Parameters

For the hydration free energy dataset, atomic coordinates for each solute were obtained from Ratkova et al.[43] Since these structures had already been geometry optimized at the MP2/6-311G(d,p) level of theory, no further pre-processing was performed. For the caco-2 cell permeability dataset, molecular structures were taken from the Supporting Information of the article by Hou et al.[62] A single global minimum energy conformer was selected for each solute by performing a low-mode conformational search using the OPLS-2005 force-field[63] in Macromodel v.9.1.[64] The atom-atom potential parameters and atomic partial charges required as input to solve the 1D RISM equations were taken from the OPLS-2005 forcefield for all molecules in both the hydration free energy and caco-2 cell permeability datasets. These parameters were selected because they have performed well in previous 1D RISM studies.[42]

3.2 Calculation of RISM-MOL-INF variables

The method to compute RISM-MOL-INF variables was implemented using a locally modified version of the RISM-MOL software[35] and additional routines written in the R statistical computing environment.[65]

3.3 Partial-Least Squares

Partial-Least Squares regression models were trained using the *pls* library[66] in the R statistical computing environment.[65] The number of latent variables to include in each PLS model was selected by plotting the root-mean-square error for leave-one-out cross-validation against the number of latent variables as discussed later.

3.4 Random Forest

Random Forests were trained with the *randomForest* library[67] in the R statistical computing environment,[65] using standard parameters: $mtry = N/3$, $nodesize = 5$, and $ntree = 500$, where N is the number of input variables and $mtry$ is rounded down to the nearest integer. There is extensive evidence in the literature that the Random Forest algorithm is insensitive to training parameters,[68,69] so that variation of $mtry$ between 40 and N, of $ntree$ from 250 upward, and of $nodesize$ in the region 5 to 10 has little effect on prediction accuracy. As has been done previously, we use these standard Random Forest parameters without further optimization.[68,69]

3.5 Statistical Analysis

To compare calculated and experimental results for different computational models, a correlation coefficient and the root mean squared deviation ($RMSD$) were evaluated:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^i - y^i_{exp})^2}{\sum_{i=1}^{n}(y^i_{exp} - M(y^i_{exp}))^2}, \tag{19}$$

$$RMSD(y, y_{exp}) = \sqrt{\frac{1}{N}\sum_i \left(y^i - y^i_{exp}\right)^2} \tag{20}$$

where index i runs through the set of N selected molecules, and y^i and y^i_{exp} are the calculated and experimental values, respectively, for molecule i for a given property (i.e. ΔG_{hyd}, $logP_{eff}$, or similar). The total deviation can be split into two parts: bias (or mean displacement, M) and standard deviation (SD), which are calculated by the formulae:

$$bias = M(y - y_{exp}) = \frac{1}{N}\sum_{i\in S}\left(y^i - y^i_{exp}\right) \tag{21}$$

$$\sigma(y - y_{exp}) = \sqrt{\frac{1}{N}\sum_{i\in S}\left(y^{(i)} - y^{(i)}_{exp} - M(y - y_{\mathrm{exp}})\right)^2} \tag{22}$$

The bias gives the systematic error, which can be corrected by a simple constant term. The standard deviation gives the random error that is not explained by the model. One can see the connection between these three formulae:

$$RMSD(y,\, y_{\mathrm{exp}})^2 = M(y - y_{\mathrm{exp}})^2 + \sigma(y - y_{\mathrm{exp}})^2 \tag{23}$$

Models reporting $RMSE$ greater than the standard deviation of the experimental data offer less accurate predictions than the null model provided by the mean of the experimental data.

Statistical analyses were carried out in the R Statistical Computing Environment.[65] Python scripts were used to manipulate raw data files.

3.6 Experimental Datasets

3.6.1 Hydration Free Energy

We use experimental hydration free energy data for 185 small organic molecules that were originally published by Ratkova et al[43]. The use of this dataset facilitates a simple benchmark, since other methods to predict hydration free energy have been tested on the same molecules.[50,70] Experimental hydration free energies (in kcal/mol) are tabulated as $\Delta G_{hydr} = -RT\ln(c_{aq}/c_{gas})$, with concentrations in mol/L, which corresponds to the choice of standard states proposed by Ben-Naim.[71] The dataset was partitioned into a training dataset of 123 molecules and a testing dataset of 62 molecules by ranking all molecules by increasing molecular weight and placing every third molecule into the test set.

3.6.2 Caco2 Permiability

Experimental caco2 cell permeability data were obtained for 100 druglike molecules from the work of Hou et al.[62] Caco2 data were expressed as $logP_{eff}$, using a decadic logarithm with P_{eff} referred to units of $cm\ s^{-1}$ Experimental $logP_{eff}$ ranged from -6.96 to -4.11 with a mean value of -5.14 and a standard deviation of 0.77 $logP_{eff}$ units. The molecules in the dataset had molecular weights from 32.0 to 670.4 Dalton with a mean of 314.7 Dalton and a standard deviation of 115.7 Dalton. The data were partitioned into the training (77 molecules) and test (23 molecules) sets that were previously used by Hou et al.[62] The molecules in the test dataset are illustrated in Figure 3.

3.7 Benchmark calculations

To provide a comparison to the results obtained by the RISM-MOL-INF approach, the following methods were used to compute hydration free energies or caco-2 cell permeability.

3.7.1 Density Functional Theory calculations

Hydration free energies were computed with density functional theory using the M06-2X density functional, the 6-31G* basis set and the SMD implicit continuum model for solvent as implemented in GAMESS-US (version released on 1st May, 2013);[72] this method was selected because it performed well in a recent blind challenge for HFE calculation.[73–75] We note that this is also the recommended method for HFE calculation in both GAMESS-US and Gaussian09. Sample GAMESS input files are provided for these calculations in the Supporting Information.

3.7.2 3D RISM/UC calculations

Hydration free energies were computed using the 3D Reference Interaction Site Model with the Universal Correction hydration free energy functional (3DRISM/UC). The 3DRISM method has been described elsewhere;[50] here we provide only a brief overview of the 3D RISM/UC functional. Within the standard 3DRISM theory, hydration free energy can be computed using the Gaussian fluctuations (GF) HFE functional, which was adopted for 3DRISM from the 1DRISM case by Kovalenko and Hirata[28,40]:

$$\Delta G^{GF}_{hyd} = k_B T \sum_{\alpha=1}^{N_{solvent}} \rho_\alpha \int\limits_{R^3} \left[-c_\alpha(\mathbf{r}) - \frac{1}{2} c_\alpha(\mathbf{r}) h_\alpha(\mathbf{r}) \right] d\mathbf{r} \tag{24}$$

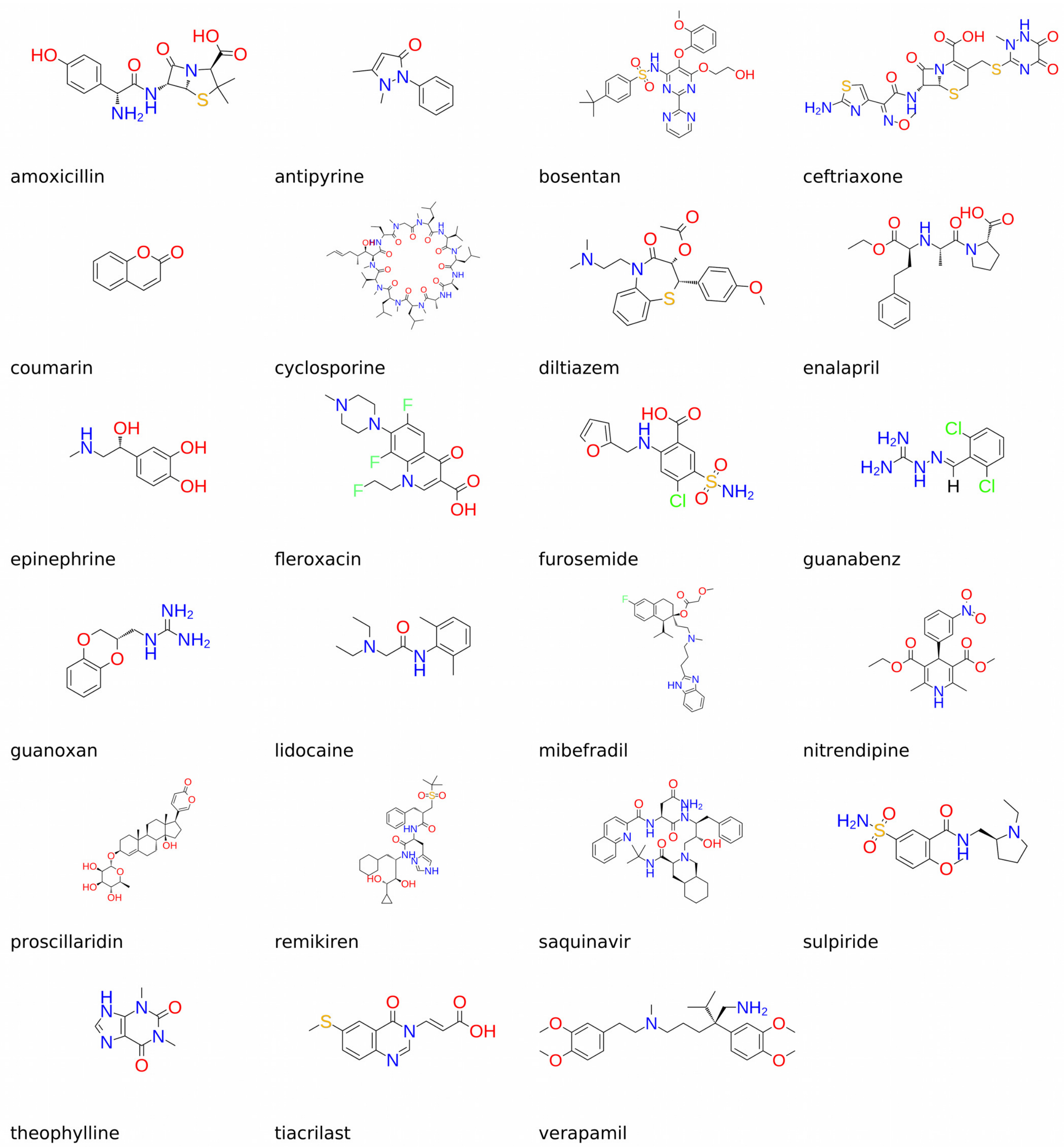


Fig. 3 Chemical structures of the molecules in the external test that was used to validate the models to predict caco-2 cell permeability

where ρ_α is the number density of solvent sites α. Unfortunately, HFEs calculated using the GF free energy functional have only a *qualitative* agreement with experiment. The error in hydration free energies calculated by the GF functional in 3D RISM is strongly correlated with the partial molar volume calculated by 3D RISM.[14,50,76] The 3D RISM/UC free energy functional developed from this observation is a linear combination of the ΔG_{hyd}^{GF}, the dimensionless partial molar contribution, ρV, and a bias correction, b (intercept):[50]

$$\Delta G_{hyd}^{3D-RISM/UC} = \Delta G_{hyd}^{GF} + a(\rho V) + b \tag{25}$$

where the values of the scaling coefficient a and intercept b are obtained by linear regression against experimental data for simple organic molecules. To provide a like-for-like comparison to the other methods tested here, we reparameterize the 3D RISM/UC model on the same dataset used to train the RISM-MOL-INF models (using ΔG_{hyd}^{GF} and ρV data reported previously[50]). A similar strategy has been employed successfully in previous studies.[24,76] The coefficients in the reparameterized 3D RISM/UC model have the values $a = -3.185$ kcal/mol and $b = 0.433$ kcal/mol. The 3D RISM is a more theoretically advanced method than 1D RISM, but is also $\sim$ 100-fold more computationally expensive, which makes it less suitable for in silico screening of large compound libraries.

We estimate the solute partial molar volume via *solute-solvent site* correlation functions using the standard 3D RISM theory expression[77,78]:

$$V = k_B T \eta \left(1 - \rho_\alpha \sum_{\alpha=1}^{N_{solvent}} \int_{R^3} c_\alpha(\mathbf{r}) d\mathbf{r} \right) \tag{26}$$

where η is the pure solvent isothermal compressibility, and ρ_α is the number density of solute sites α.

The 3D RISM/UC method has been shown to give accurate hydration free energies for both simple organic molecules and bioactive (druglike) molecules.[14,50,76] and has been successfully used in computing solubility[2] and protein-ligand binding free energies.[20]

3.7.3 ACD/Labs Software

Caco-2 permeability was predicted using version 12 of the commercial software released by ACD/Labs

3.7.4 QSPR models

As a benchmark for the RISM-MOL-INF results, QSPR models were built using Random Forest regression on a set of common 2D and 3D descriptors computed by the program PADEL. The PADEL software supports the calculation of a wide-variety of molecular descriptors, of which 355 2D/3D descriptors were considered here (the remaining descriptors had either low variance or non-numeric values for some molecules and were removed before regression models were built). The final set of descriptors included calculated physical properties (Moriguchi logP, Crippen's logP and molar refractivity), constitutional descriptors (counts of atoms and functional groups, counts of polar atoms/bonds), thermodynamic descriptors (molecular linear free energy relation descriptors), connectivity and topological indices (electrotopological state descriptors, extended topochemical atom descriptors, Kier/Wiener/Balaban/Zagreb indices), molecular flexibility descriptors (counts/fractions of rotatable bonds), pharmacophore feature counts (counts of hydrogen bond donors and acceptors), volume descriptors (McGowan volume), and molecular shape and surface area descriptors (solvent-accessible surface areas, total polar surface area), amongst other properties.

4 Results and Discussion

4.1 Hydration Free Energy

Four models were trained to predict hydration free energies using partial-least squares regression on each of the four sets of RISM-MOL-INF descriptors (gf_w, hnc_w, $hncb\_w$ and kh_w). Each partial-least squares regression model was trained using seven latent variables, where this number was selected by considering the graph of root-mean-square error for leave-one-out cross-validation ($RMSE(cv)$) against number of latent variables (see Supporting Information). The results for fit-to-the-training data, leave-one-out cross validation and prediction of an external test set are presented in Table 1.

All of the models trained on RISM-MOL-INF variables predict hydration free energy significantly more accurately than any of the standard 1D RISM free energy functionals (e.g. GF, HNCB, HNC and KH), which provides evidence to support the approach taken to develop the RISM-MOL descriptors. In this context, it is interesting to note that many of the RISM-MOL-INF descriptors are significantly more correlated with experimental HFE than are the calculated HFEs obtained using the standard 1D RISM free energy functionals. For example, Figure 4 shows that the kh_w_900 RISM-MOL-INF descriptor has a strong inversely linear correlation with experimental HFE ($R = -0.93$), even though the HFE calculated using the KH free energy functional Equation 6 has a very low correlation with experiment ($R = 0.36$). This observation explains why combining the RISM-MOL descriptors with PLS regression leads to successful predictions even when the standard 1D RISM functionals fail.

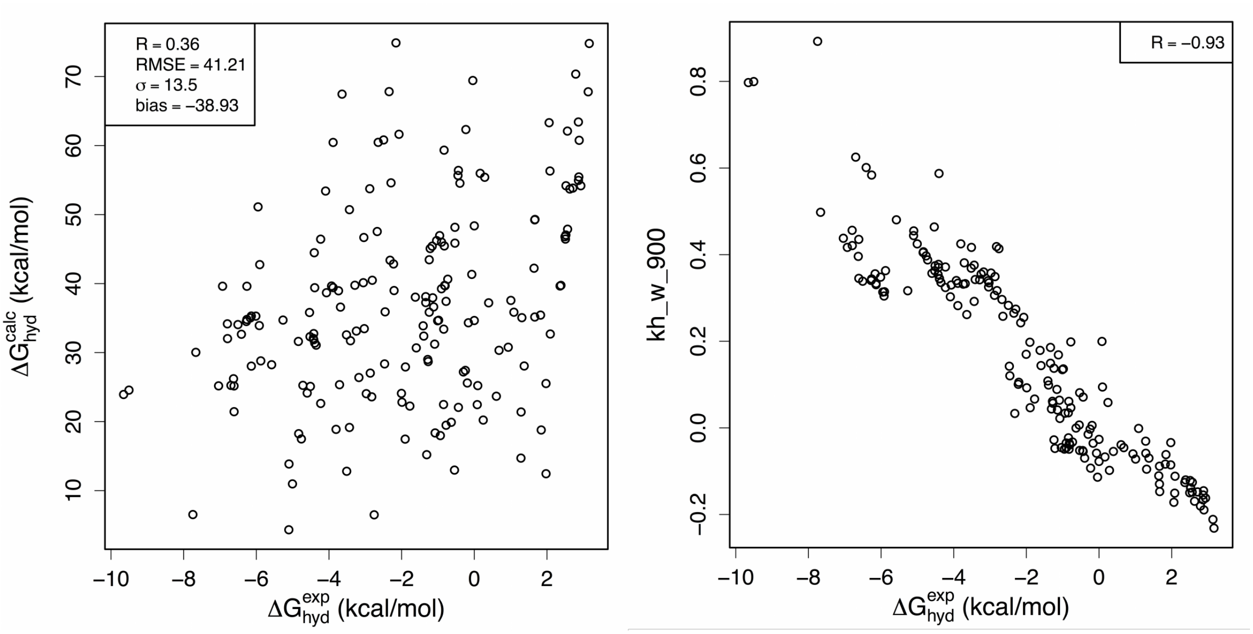


Fig. 4 (i) uncorrected values of hydration free energy calculated using the Kovalenko-Hirata free energy functional plotted against experimental HFE data (left-hand-side); (ii) the *kh_w* descriptor with the highest correlation to experimental hydration free energy data plotted against experimental hydration free energy data (right-hand-side).

From inspection of Table 1, the choice of free energy functional from which to compute RISM-MOL-INF variables does not strongly influence the accuracy of the models to predict HFE. The reason for this can be seen in the Figure provided in the Supporting Information, where the functions from which the different sets of RISM-MOL descriptors are computed are shown to have similar forms for four small organic molecules (even though the HFEs computed from the standard 1D RISM functionals GF, KH, HNC or HNCB differ by up to an order of magnitude).

The most accurate predictions of HFE for the external test set were obtained using the *kh_w* variables in a 7-latent-variable PLS regression model, which gave $R = 0.98$, RMSE= 0.60 kcal/mol (Table 1) and no significant outliers (Figure 5). Indeed, all of the RISM-MOL-INF variables gave relatively accurate results with $0.60 < \text{RMSE} < 0.63$ kcal/mol.

To provide a benchmark, HFEs were also computed with: (i) density functional theory using the M06-2X density functional combined with the 6-31G* basis set and the SMD implicit continuum model for aqueous solvent; (ii) the 3DRISM/UC model reparameterized using the molecules in the training set. As can be seen from inspection of Table 1, although the density functional and 3DRISM based approaches give accurate hydration free energies, the predictions obtained using the RISM-MOL-INF variables are significantly more accurate than those obtained using these benchmark methods, which provides encouraging proof-of-concept of the method proposed here.

Method	R(tr)	RMSE(tr)	bias(tr)	R(cv)	RMSE(cv)	bias(cv)	R(te)	RMSE(te)	bias(te)
gf_w	0.98	0.56	0.00	0.97	0.67	0.01	0.98	0.61	-0.12
hncb_w	0.98	0.58	0.00	0.97	0.71	0.02	0.98	0.63	-0.13
hnc_w	0.98	0.57	0.00	0.97	0.68	0.02	0.98	0.60	-0.11
kh_w	0.98	0.57	0.00	0.97	0.68	0.01	0.98	0.60	-0.11
GF	0.82	3.59	-2.37				0.83	3.70	-2.49
HNCB	0.85	5.93	4.67				0.82	5.79	4.46
HNC	0.36	42.43	-40.21				0.23	42.18	-40.17
KH	0.36	41.21	-38.93				0.23	40.95	-38.90
QM	0.96	0.87	-0.31				0.97	0.78	-0.31
3DRISM/UC	0.95	0.93	0.00				0.96	0.84	-0.16

Table 1 Prediction of the hydration free energy of an external test set of 62 small organic molecules. The top four lines of the table show the results obtained using the RISM-MOL-INF method with the *gf_w*, *hncb_w*, *hnc_w*, or *kh_w* variables. The following four lines show the results obtained using the GF, HNCB, HNC and KH free energy functionals. The remainder of the table provides the results obtained using density functional theory combined with a polarizable continuum model for solvent (M06-2X/6-31G*/PCM) or the 3D Reference Interaction Site Model with *Universal Correction* free energy functional (3DRISM/UC). The statistics reported in columns two to ten are the correlation coefficient (R), the root-mean-square-error (RMSE) and the bias and these are assessed for the training data (tr), the test data (te), and for 10-fold cross-validation (cv).

4.2 caco2

Accurate computational methods to predict the caco2 permeability of organic molecules are highly sought after in pharmaceutical research and development to assess the bioavailability of de novo designed drugs.[79] Despite recent progress, caco-2 cellpermeability remains a difficult property to calculate directly from molecular simulation.

To further validate the RISM-MOL-INF descriptors, we applied them to the prediction of caco2 permeability for a dataset of 100 druglike molecules (partitioned into a training dataset of 77 molecules and a test set of 23 molecules). For one molecule in the external test set (cyclosporin), the 1D RISM equations did not converge within a reasonable time period. Cyclosporin - a cyclic peptide containing a large ring system comprising 33 backbone atoms - has a significantly higher molecular weight (1202.61 Daltons) than most orally administered

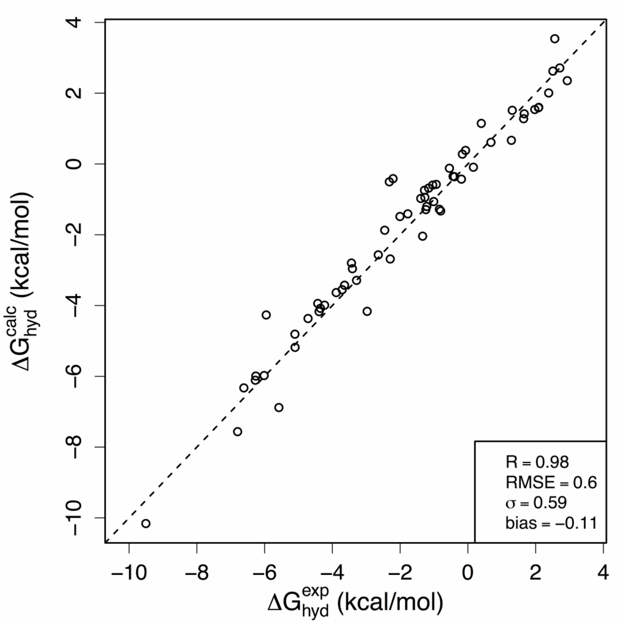


Fig. 5 Correlation plot of experimental versus predicted hydration free energy data for an external test set of 62 organic molecules. Predictions were made with a 7-latent variable partial-least squares model trained using *kh_w* descriptors.

pharmaceuticals. For the purposes of this work, cyclosporin was removed from the external test set and statistics are reported for 22 molecules.

Descriptors	Model	R(cv)	RMSE(cv)	bias(cv)	R(te)	RMSE(te)	bias(te)
gf_w	RF	0.79	0.45	0.02	0.91	0.39	-0.07
hncb_w	RF	0.78	0.45	0.03	0.91	0.41	-0.08
hnc_w	RF	0.77	0.46	0.03	0.91	0.41	-0.09
kh_w	RF	0.77	0.47	0.02	0.91	0.41	-0.09
Padel	RF	0.73	0.51	-0.01	0.75	0.60	-0.13
ACD					0.79	0.56	0.17
Hou et al.[62]					0.77	0.58	0.04
Ponce et al.[80]					0.80	1.43	0.52

Table 2 Prediction of caco-2 cell permeability (logP_{eff}) for a dataset of druglike molecules

Four models to predict caco-2 cell permeability were developed using the training dataset and Random Forest regression combined with the *gf_w*, *kh_w*, *hnc_w*, *hncb_w* RISM-MOL-INF variables. The predictive accuracy of the four different RISM-MOL-INF models for the external test set is shown in Table 2. All of the RISM-MOL-INF models tested here gave reasonably accurate predictions of caco-2 cell permeability with $R(te) > 0.9$ and $RMSE(te) \leq 0.41$. The best RISM-MOL predictions of caco-2 cell permeability were obtained using *gf_w* descriptors, which gave $R(te) = 0.91$, RMSE= 0.39 $logP_{eff}$ units (Table 2 and Figure 6).

Figure 7 shows a barplot of the Random Forest importance metric for each $gf\_w(r)$ variable (blue) overlaid on a line graph showing the form of the gf_w function for each molecule in the training dataset (grey) (similar graphs for *kh_w*, *hnc_w*, *hncb_w* are provided in the Supporting Information). Random Forest importance was assessed for each variable by measuring the increase in mean-square-error for cross-validation when the Random Forest was retrained with the selected variable replaced by Gaussian noise. Although the gf_w function for each molecule in Figure 7 is a combination of $N_S \times N_\alpha$ total and direct solute-solvent correlation functions, which means that the peaks in the gf_w function do not correspond directly to solvent shells, it is nonetheless interesting to note that in the region $3< r <10$ Å important variables are found in clusters (at solute-solvent distances of $\sim$ 3 Å, $\sim$ 4.5 Å, $\sim$ 5.5 Å,$\sim$ 7 Å and $\sim$ 9 Å) and that the average importance metric in each cluster decreases as solute-solvent distance increases. The most important variables are found at $\sim$ 3 Å, which corresponds approximately to the position of the first solvation shells observed in the pairwise total correlation functions between water oxygen atoms and solute heavy atoms in Figure 1. The importance to the model of a cluster of variables at r $\sim$ 11 Å(Figure 7) is surprising because solute-solvent correlations would be expected to be weak at this intermolecular separation. Long-range correlations are known to be poorly treated in standard 1D RISM methods.[28] If the model is retrained using only those variables corresponding to $r <$ 10Å the root-mean-square error for prediction of the test set increases a small amount from 0.39 to 0.42 $logP_{eff}$ units, which is still more accurate than any of the other approaches against which we have benchmarked the RISM-MOL-INF predictions. Therefore, in future, removing those descriptors corresponding to large values of r may provide a simple method to increase physical interpretability without significantly reducing predictivity.

Comparing the values of $R(te)$, $RMSE(te)$ and $bias(te)$ for the RISM-MOL-INF methods with those obtained using the two benchmark methods (i.e. a QSPR model based on PADEL descriptors and predictions made using ACD v12 software), it is clear that the RISM-MOL-INF method provides significantly more accurate predictions of caco-2 cell permeability (Table 2). Furthermore, the RISM-MOL-INF method also performs better than the QSPR models previously reported by Hou et al.[62] and Ponce et al.[80] (Table 2). These results provide a clear proof-of-concept of the RISM-MOL-INF methods proposed here.

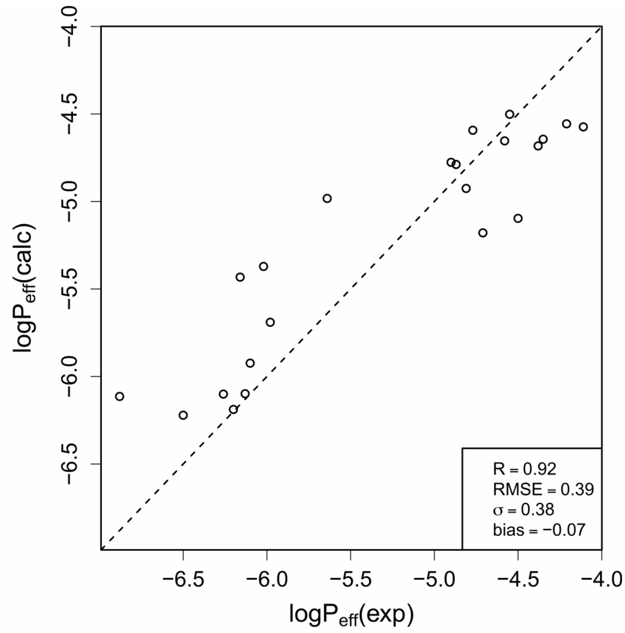


Fig. 6 Correlation plot of experimental versus predicted caco-2 cell permeability for an external test set of 22 druglike molecules. Predictions were made using Random Forest regression trained on *gf_w* variables

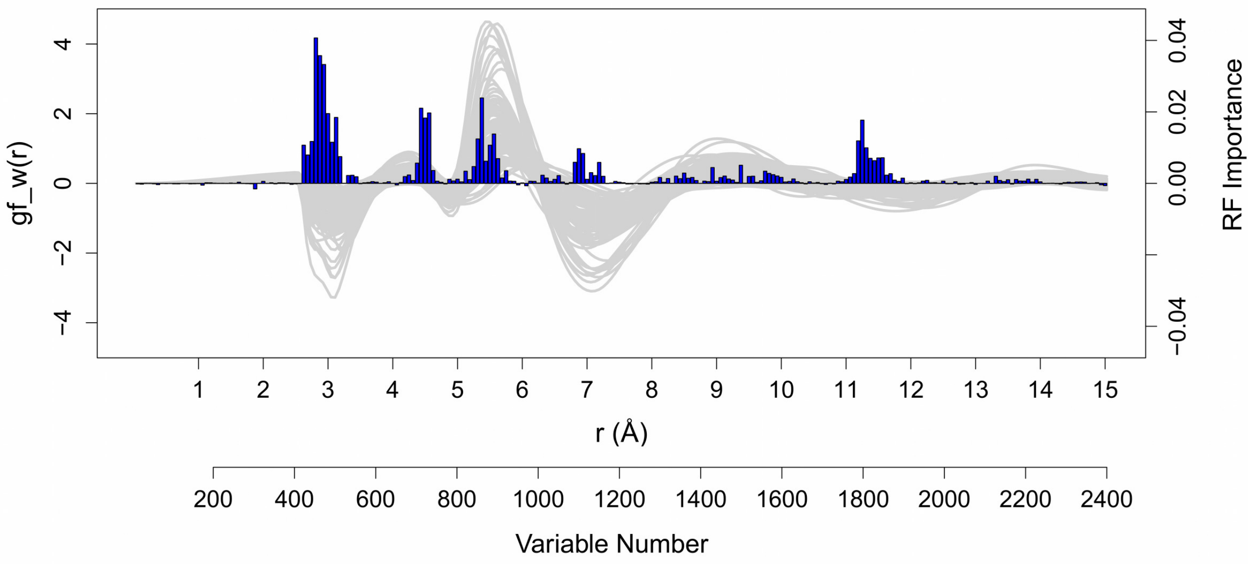


Fig. 7 Plot of the function from which gf_w variables are defined (grey lines) overlaid on a barpot showing the importance of each gf_w variable to the Random Forest model to predict caco-2 cell permeability (blue bars). Please see the text for a definition of the Random Forest importance metric.

4.3 Computational Expense

One of the possible applications of RISM-MOL-INF variables is *in silico* screening of large virtual libraries, where it is necessary to consider the computational expense of the selected method, as well as its accuracy. The calculations discussed in this paper were performed in serial on Dual Intel Xeon X5650 2.66 GHz processors at the ARCHIE-WeSt supercomputing centre located at the University of Strathclyde in Glasgow, Scotland. The most time-consuming step in computing the RISM-MOL-INF variables for a selected solute is solving the 1D RISM equations using the RISM-MOL program; the remaining steps require minimal computational expense. Figure 8 shows the time required for the RISM-MOL calculation plotted against number of solute atoms for a dataset of 1000 small organic molecules taken at random from the PUBCHEM database.[81] The mean time required for the RISM-MOL calculations was 3 minutes and 30 seconds. Since the RISM-MOL-INF variables can be calculated in a matter of minutes for most small organic and druglike molecules using existing software (RISM-MOL), it suggests that they may be useful for medium to high-throughput *in silico* screening of large virtual compound libraries.

5 Conclusions

We have proposed a method to compute variables to quantify solvation and desolvation processes in molecular informatics. The RISM-MOL-INF variables are derived from solvent density distribution functions computed by the 1D Reference Interaction Site Model of the Integral Equation Theory of Molecular Liquids. As such, they quantify the response of solvent molecules as a function of distance from the selected solute. The RISM-MOL-INF variables can be computed in a matter of minutes for most druglike solutes using existing software (RISM-MOL). We have shown that hydration free energy and caco2 cell permeability can be predicted accurately using RISM-MOL-INF variables only combined with statistical or machine learning algorithms.

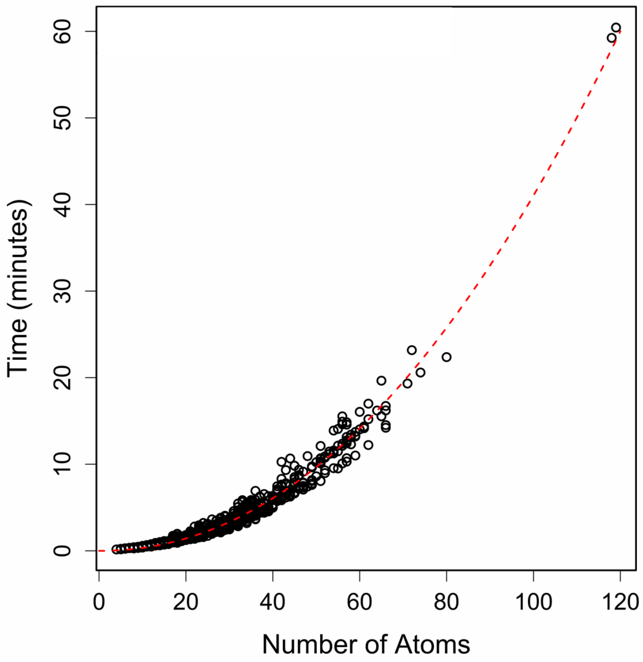


Fig. 8 The time required to solve the 1D RISM equations scales as $\sim N^2$, where N is the number of atoms in the molecule. The dashed-line shows an exponential fit of this dependence with parameters $y = 0.00278(\pm 0.00008)x^{2.08(\pm 0.007)}$.

There is clearly great scope to develop the methods presented here. From one side, the RISM-MOL-INF variables would be expected to benefit from developments in RISM theory. Open problems in this field include the design of bridge functionals,[82] free energy functionals,[14] and more efficient and robust algorithms for solving the RISM equations.[35,83] It should also be possible to derive useful descriptors from correlation functions computed by 3D RISM, which is a more computationally expensive, but more advanced model from RISM theory. The use of 3D RISM would also be more appropriate for modeling larger biomacromolecules (e.g peptides, proteins) whose complex solvation behaviour is known to be poorly represented by standard 1D RISM theory. Further work is required to define the domain of applicability of the RISM-MOL-INF method presented here and to test whether this can be extended using 3D RISM. From another side, it may be possible to derive variables that are more relevant to specific molecular informatics tasks than those discussed here. For example, by considering cosolutes or non-aqueous solvents, both of which are possible using existing RISM methods. The prediction of octanol-water distribution coefficient using RISM-MOL-INF variables computed separately for solute in octanol and solute in water is an obvious example of this idea. The performance of RISM-MOL-INF variables for standard cheminformatics applications such as assessing molecular diversity or classification tasks should also be investigated. Since solvation and desolvation effects are important in many biomolecular processes, we believe that the RISM-MOL-INF variables will find many applications in biophysical and biomedical property prediction. Further work to address the points raised in this paragraph is ongoing in our laboratory.

Acknowledgements DSP is grateful for funding from the European Commission through a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme (FP7-PEOPLE-2010-IEF. Grant agreement number: PIEF-GA-2010-272391). DSP and MVF thank the Scottish Universities Physics Alliance (SUPA) for funding through MVFs

SUPA2 start-up funds. DSP and MVF are also grateful for partial financial support from the Deutsche Forschungsgemeinschaft (DFG) - German Research Foundation, Research Grant FE 1156/2-1, and for the use of the computing facilities at the John von Neumann-Institut für Computing (NIC), Juelich Supercomputing Centre (JSC), Forschungszentrum Juelich GmbH, Germany - Project IDs: HLZ18, HLZ16, ESMI11. DSP and MVF are grateful for use of the EPSRC funded ARCHIE-WeSt High Performance Computer (www.archie-west.ac.uk, EPSRC grant no. EP/K000586/1). DSP thanks the University of Strathclyde for support through its Strategic Appointment and Investment Scheme.

6 Supporting Information

Further information about the regression models is provided in the supporting information. This information is available free of charge via the Internet at http://pubs.acs.org.

References

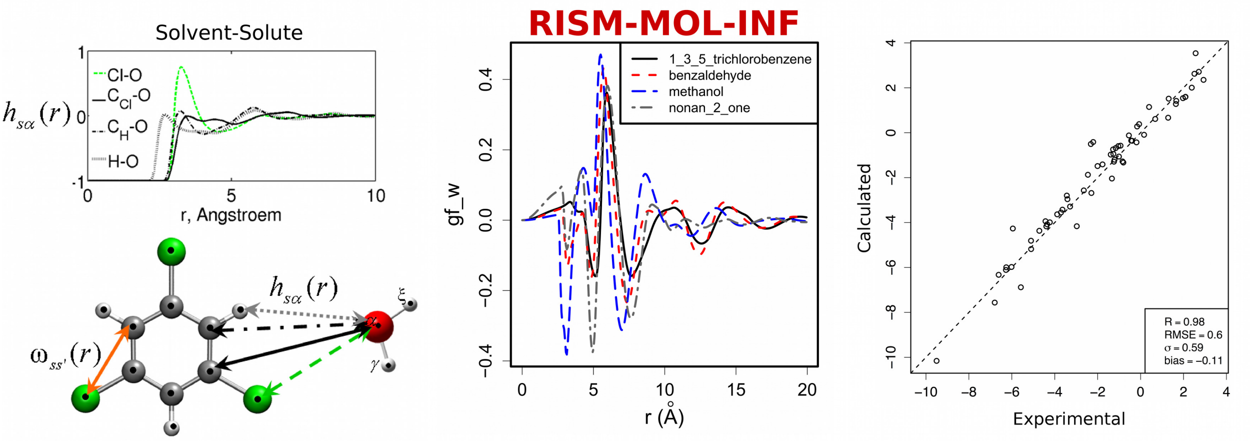
1. Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *ASAP. DOI:10.1021/cr5000283*.
2. Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; V., F. M. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory. Comput.* **2012**, *8*, 3322–3337.
3. Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A. Ligand Mapping on Protein Surfaces by the 3D-RISM Theory: Toward Computational Fragment-Based Drug Design. *J. Am. Chem. Soc.* **2009**, *131*, 12430–12440.
4. Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. Water molecules in a protein cavity detected by a statistical-mechanical theory. *J. Am. Chem. Soc.* **2005**, *127*, 15334–15335.
5. Imai, T.; Kinoshita, M.; Hirata, F. Salt effect on stability and solvation structure of peptide: An integral equation study. *Bull. Chem. Soc. Jpn.* **2000**, *73*, 1113–1122.
6. Kast, S. M.; Heil, J.; Gussregen, S.; Schmidt, K. F. Prediction of tautomer ratios by embedded-cluster integral equation theory. *J. Comput. Aided Mol. Des.* **2010**, *24*, 343–353.
7. Stumpe, M. C.; Blinov, N.; Wishart, D.; Kovalenko, A.; Pande, V. S. Calculation of Local Water Densities in Biological Systems: A Comparison of Molecular Dynamics Simulations and the 3D-RISM-KH Molecular Theory of Solvation. *J. Phys. Chem. B* **2011**, *115*, 319–328.
8. Miyata, T.; Hirata, F. Combination of molecular dynamics method and 3D-RISM theory for conformational sampling of large flexible molecules in solution. *J. Comput. Chem.* **2008**, *29*, 871–882.
9. Hilgers, A. R.; Conradi, R. A.; Burton, P. S. Caco-2 Cell Monolayers as a Model for Drug Transport Across the Intestinal Mucosa. *Pharm. Res.* **1990**, *7*, 902–910.
10. Artursson, P.; Karlsson, J. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochem. Bioph. Res. Co* **1991**, *175*, 880–885.
11. Steffansen, B., Brodin, B., Nielsen, C. U., Eds. *Molecular Biopharmaceutics*, 1st ed.; Pharmaceutical Press: London, 2010.
12. Matubayasi, N.; Nakahara, M. An approach to the solvation free energy in terms of the distribution functions of the solute-solvent interaction energy. *J. Mol. Liq.* **2005**, *119*, 23–29.
13. Matubayasi, N. Free-Energy Analysis of Solvation with the Method of Energy Representation. *Front. Biosci.* **2009**, *14*, 3536–3549.
14. Palmer, D. S.; Chuev, G. N.; Ratkova, E. L.; Fedorov, M. V. In silico screening of bioactive and biomimetic solutes by Integral Equation Theory. *Curr. Pharm. Des.* **2011**, *17*, 1695–1708.
15. Perlovich, G. L.; Rodionov, S. V.; Bauer-Brandl, A. Thermodynamics of solubility, sublimation and solvation processes of parabens. *Eur. J. Pharm. Sci.* **2005**, *24*, 25–33.
16. Perlovich, G.; Bauer-Brandl, A. Solvation of Drugs as a Key for Understanding Partitioning and Passive Transport Exemplified by NSAIDs. *Curr. Drug Deliv.* **2004**, *1*, 213–226.
17. Docherty, R.; Pencheva, K.; Abramov, Y. A. Low solubility in drug development: de-convoluting the relative importance of solvation and crystal packing. *J. Pharm. Pharmacol.* **2015**, *67*, 847–856.
18. Abramov, Y. A. Major Source of Error in QSPR Prediction of Intrinsic Thermodynamic Solubility of Drugs: Solid vs Nonsolid State Contributions? *Mol. Pharmaceutics* **2015**, *12*, 2126–2141.
19. Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol. Pharmaceutics* **2008**, *5*, 266–279.
20. Palmer, D. S.; Sørensen, J.; Schiøtt, B.; Fedorov, M. V. Solvent Binding Analysis and Computational Alanine Scanning of the Bovine Chymosin – Bovine κ-Casein Complex using Molecular Integral Equation Theory. *J. Chem. Theory Comput.* **2013**, *9*, 5706–5717.
21. Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and rationalization of protein pK(a) values using QM and QM/MM methods. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

22. Garrido, N. M.; Queimada, A. J.; Jorge, M.; Macedo, E. A.; Economou, I. G. 1-Octanol/Water Partition Coefficients of n-Alkanes from Molecular Simulations of Absolute Solvation Free Energies. *J. Chem. Theory Comput.* **2009**, *5*, 2436–2446.
23. Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J. Chem. Inf. Model.* **2008**, *48*, 220–232.
24. Huang, W.; Blinov, N.; Kovalenko, A. Octanol-Water Partition Coefficient from 3D-RISM-KH Molecular Theory of Solvation with Partial Molar Volume Correction. *J. Phys. Chem. B* **2015**, *119*, 5588–5597.
25. Mackay, D.; Shiu, W. Y.; Sutherland, R. P. Determination of air-water Henry's law constants for hydrophobic pollutants. *Environ. Sci. Technol.* **1979**, *13*, 333–337.
26. Ratkova, E. L.; Fedorov, M. V. Combination of RISM and Cheminformatics for Efficient Predictions of Hydration Free Energy of Polyfragment Molecules: Application to a Set of Organic Pollutants. *J. Chem. Theory Comput.* **2011**, *7*, 1450–1457.
27. Chandler, D.; Andersen, H. C. Optimized cluster expansions for classical fluids. 2. Theory of molecular liquids. *J. Chem. Phys.* **1972**, *57*, 1930–1937.
28. Hirata, F., Ed. *Molecular theory of solvation*; Kluwer Academic Publishers, Dordrecht, Netherlands, 2003.
29. Hansen, J.-P.; McDonald, I. R. *Theory of Simple Liquids, 4th ed*; Elsevier Academic Press, Amsterdam, The Netherlands, 2000.
30. Kinoshita, M.; Okamoto, Y.; Hirata, F. Calculation of solvation free energy using RISM theory for peptide in salt solution. *J. Comput. Chem.* **1998**, *19*, 1724–1735.
31. Kinoshita, M.; Okamoto, Y.; Hirata, F. Calculation of hydration free energy for a solute with many atomic sites using the RISM theory: A robust and efficient algorithm. *J. Comput. Chem.* **1997**, *18*, 1320–1326.
32. Kovalenko, A.; Hirata, F. Potential of mean force between two molecular ions in a polar molecular solvent: A study by the three-dimensional reference interaction site model. *J. Phys. Chem. B* **1999**, *103*, 7942–7957.
33. Kovalenko, A.; Hirata, F. Self-consistent description of a metal-water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **1999**, *110*, 10095–10112.
34. Chuev, G. N.; Fedorov, M. V. Wavelet algorithm for solving integral equations of molecular liquids. A test for the reference interaction site model. *J. Comput. Chem.* **2004**, *25*, 1369–1377.
35. Sergiievskyi, V. P.; Hackbusch, W.; Fedorov, M. V. Multigrid Solver for the Reference Interaction Site Model of Molecular Liquids Theory. *J. Comput. Chem.* **2011**, *32*, 1982–1992.
36. Singer, S. J.; Chandler, D. Free-energy Functions in the Extended RISM Approximation. *Mol. Phys.* **1985**, *55*, 621–625.
37. Ten-no, S. Free energy of solvation for the reference interaction site model: Critical comparison of expressions. *J. Chem. Phys.* **2001**, *115*, 3724–3731.
38. Sato, K.; Chuman, H.; Ten-no, S. Comparative study on solvation free energy expressions in reference interaction site model integral equation theory. *J. Phys. Chem. B* **2005**, *109*, 17290–17295.
39. Kovalenko, A.; Hirata, F. Hydration free energy of hydrophobic solutes studied by a reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method. *J. Chem. Phys.* **2000**, *113*, 2793–2805.
40. Chandler, D.; Singh, Y.; Richardson, D. M. Excess Electrons In Simple Fluids .1. General Equilibrium-Theory For Classical Hard-Sphere Solvents. *J. Chem. Phys.* **1984**, *81*, 1975–1982.
41. Chuev, G.; Fedorov, M.; Crain, J. Improved estimates for hydration free energy obtained by the reference interaction site model. *Chem. Phys. Lett.* **2007**, *448*, 198–202.
42. Palmer, D. S.; Sergiievskyi, V. P.; Jensen, F.; Fedorov, M. V. Accurate calculations of the hydration free energies of druglike molecules using the reference interaction site model. *J. Chem. Phys.* **2010**, *133*, 044104.
43. Ratkova, E. L.; Chuev, G. N.; Sergiievskyi, V. P.; Fedorov, M. V. An Accurate Prediction of Hydration Free Energies by Combination of Molecular Integral Equations Theory with Structural Descriptors. *J. Phys. Chem. B* **2010**, *114*, 12068–12079.
44. Kirkwood, J. G.; Buff, F. P. The Statistical Mechanical Theory Of Solutions .1. *J. Chem. Phys.* **1951**, *19*, 774–777.
45. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
46. Yamazaki, T.; Kovalenko, A. Spatial Decomposition of Solvation Free Energy Based on the 3D Integral Equation Theory of Molecular Liquid: Application to Miniproteins. *J. Phys. Chem. B* **2011**, *115*, 310–318.
47. Silveira, R. L.; Stoyanov, S. R.; Gusarov, S.; Skaf, M. S.; Kovalenko, A. Plant Biomass Recalcitrance: Effect of Hemicellulose Composition on Nanoscale Forces that Control Cell Wall Strength. *J. Am. Chem. Soc.* **2013**, *135*, 19048–19051.
48. Silveira, R. L.; Stoyanov, S. R.; Gusarov, S.; Skaf, M. S.; Kovalenko, A. Supramolecular Interactions in Secondary Plant Cell Walls: Effect of Lignin Chemical Composition Revealed with the Molecular Theory of Solvation. *J. Phys. Chem. Lett.* **2015**, *6*, 206–211.
49. Huang, W.; Dedzo, G. K.; Stoyanov, S. R.; Lyubimova, O.; Gusarov, S.; Singh, S.; Lao, H.; Kovalenko, A.; Detellier, C. Molecule–Surface Recognition between Heterocyclic Aromatic Compounds and Kaolinite in Toluene Investigated by Molecular Theory of Solvation and Thermodynamic and Kinetic Experiments. *J. Phys. Chem. C* **2014**, *118*, 23821–23834.
50. Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards a universal method to calculate hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys. Cond. Matt.* **2010**, *22*, 492101.

51. O'Boyle, N. M.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Simultaneous feature selection and parameter optimisation using an artificial ant colony: case study of melting point prediction. *Chem. Cent. J.* **2008**, *2*, 21.
52. Freedman, D. A. *Statistical Models: Theory and Practice*, 2nd ed.; Cambridge University Press: Cambridge ; New York, 2009.
53. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
54. Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.
55. Perkyns, J.; Pettitt, B. M. A Site Site Theory For Finite Concentration Saline Solutions. *J. Chem. Phys.* **1992**, *97*, 7656–7666.
56. Perkyns, J. S.; Pettitt, B. M. A Dielectrically Consistent Interaction Site Theory For Solvent Electrolyte Mixtures. *Chem. Phys. Lett.* **1992**, *190*, 626–630.
57. Lue, L.; Blankschtein, D. Liquid-state theory of hydrocarbon water-systems: application to methane, ethane, and propane. *J. Phys. Chem.* **1992**, *96*, 8582–8594.
58. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
59. Hirata, F.; Rossky, P. J. An Extended RISM Equation For Molecular Polar Fluids. *Chem. Phys. Lett.* **1981**, *83*, 329–334.
60. Lee, P. H.; Maggiora, G. M. Solvation Thermodynamics Of Polar-Molecules In Aqueous-Solution By The XRISM Method. *J. Phys. Chem.* **1993**, *97*, 10175–10185.
61. Allen, M. P., Tildesley, D. J., Eds. *Computer Simulation of Liquids*; Clarendon Press, Oxford, 1987.
62. Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *J. Chem. Inf. and Comput. Sci.* **2004**, *44*, 1585–1600, PMID: 15446816.
63. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
64. Schrodinger LLC (2008) Schrodinger Suite 2008. Maestro Version 8.5. MacroModel Version 9.6.
65. R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2015.
66. Mevik, B.-H.; Wehrens, R.; Liland, K. H. pls: Partial Least Squares and Principal Component regression. 2013; R package version 2.4-3.
67. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
68. Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Choosing Feature Selection and Learning Algorithms in QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 837–843.
69. Palmer, D. S.; Mitchell, J. B. O. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Mol. Pharmaceutics* **2014**, *11*, 2962–2972.
70. Frolov, A. I.; Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Hydration Thermodynamics using the Reference Interaction Site Model: Speed or Accuracy? *J. Phys. Chem. B* **2011**, *115*, 6011–6022.
71. Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016–2027.
72. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
73. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 4538–4543.
74. Nicholls, A.; Wlodek, S.; Grant, J. SAMPL2 and continuum modeling. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 293–306.
75. Meunier, A.; Truchon, J.-F. Predictions of hydration free energies from continuum solvent with solute polarizable models: the SAMPL2 blind challenge. *J. Comput-Aided Mol. Des.* **2010**, *24*, 361–372.
76. Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Toward a Universal Model To Calculate the Solvation Thermodynamics of Druglike Molecules: The Importance of New Experimental Databases. *Mol. Pharmaceutics* **2011**, *8*, 1423–1429.
77. Harano, Y.; Imai, T.; Kovalenko, A.; Kinoshita, M.; Hirata, F. Theoretical study for partial molar volume of amino acids and polypeptides by the three-dimensional reference interaction site model. *J. Chem. Phys.* **2001**, *114*, 9506–9511.
78. Imai, T.; Harano, Y.; Kovalenko, A.; Hirata, F. Theoretical study for volume changes associated with the helix-coil transition of peptides. *Biopolymers* **2001**, *59*, 512–519.
79. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.
80. Ponce, Y. M.; Perez, M. A. C.; Zaldivar, V. R.; Ofori, E.; Montero, L. A. Total and Local Quadratic Indices of the Molecular Pseudographs Adjacency Matrix. Application to Prediction of Caco-2 Permeability of Drugs. *Int. J. Mol. Sci.* **2003**, *4*, 512–536.
81. Bolton, E.; Wang, Y.; Thiessen, P.; Bryant, S. *Annual Reports in Computational Chemistry*; American Chemical Society, Washington, DC, 2008; Vol. 4.
82. Chuev, G. N.; Vyalov, I.; Georgi, N. Extraction of atom-atom bridge and direct correlation functions from molecular simulations: A test for ambient water. *Chem. Phys. Lett.* **2013**, *561-562*, 175–178.
83. Sergiievskyi, V. P.; Fedorov, M. V. *J. Chem. Theory Comput.* **2012**, *8*, 2062–2070.

7 Table of Contents graphic

For table of contents use only for the manuscript entitled, "Fast and General Method to Predict the Physico-Chemical Properties of Druglike Molecules using the Integral Equation Theory of Molecular Liquids", by David S. Palmer, Maksim Mišin, Maxim V. Fedorov, and Antonio Llinas.



8 Keywords

ADME, ADMET, caco-2, hydration free energy, solvation free energy, permeability, drug discovery, bioavailability, integral equation theory of molecular liquids, statistical mechanics, QSPR, QSAR, druglike molecules, Random Forest, RISM, IET, reference interaction site model