Multi-Solvent Models for Solvation Free Energy Predictions using 3D-RISM Hydration Thermodynamic Descriptors

Vigneshwari Subramanian[†,§], Ekaterina Ratkova[‡], David Palmer[§], Ola Engkvist[||], Maxim Fedorov[#,††], Antonio Llinas*[,†]

[†]Drug Metabolism and Pharmacokinetics, Research and Early Development - Respiratory, Inflammation and Autoimmune, Biopharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, SE-431 83, Mölndal, Sweden

[‡]Medicinal Chemistry, Research and Early Development - Cardiovascular, Renal and Metabolism, Biopharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, SE-431 83, Mölndal, Sweden

[§]Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland G1 1XL, U.K.

[||]Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Pepparedsleden 1, SE-431 83, Mölndal, Sweden

[#]Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow, 143026, Russia.

[††]Department of Physics, Scottish Universities Physics Alliance (SUPA), University of Strathclyde, John Anderson Building, 107 Rottenrow, Glasgow, Scotland G4 0NG, U.K.

Abstract

The potential to predict Solvation Free Energies (SFEs) in any solvent using a machine learning (ML) model based on thermodynamic output, extracted exclusively from 3D-RISM simulations in water is investigated. The models on multiple solvents take into account both the solute and solvent description and offer the possibility to predict SFEs of any solute in any solvent with root mean squared errors less than 1 kcal/mol. Validations that involve exclusion of fractions or clusters of the solutes or solvents exemplify the model's capability to predict SFEs of novel solutes and solvents with diverse chemical profiles. In addition to being predictive, our models can identify the solute and solvent features that influence SFE predictions. Furthermore, using 3D-RISM hydration thermodynamic output to predict SFEs in any organic solvent reduces the need to run 3D-RISM simulations in all these solvents. Altogether, our multi-solvent models for SFE predictions that take advantage of the solvation effects are expected to have an impact in the property prediction space.

Introduction

In the last decade, many pharmaceutical companies have heavily invested in reducing the drug attrition rate, especially in the late phases. Frontloading Absorption, Distribution, Metabolism, and Excretion (ADME) considerations in Drug Discovery has had a significant impact in reducing the drug attrition from approximately 40% in 1990 to 10% in 2000[1]. The drug design teams have been able to shift attrition to earlier phases by focusing on "quality" rather than "quantity" reducing times and costs[2] and dramatically increasing the number of positive Proof of Mechanisms (PoMs) achieved[3,4] These improvements can be attributed to the increased understanding of human pharmacokinetics (PK), pharmacodynamics (PD) and ADME properties of leads and candidates. Early prediction of physicochemical properties is therefore vital and allow design teams to drive the design of new molecules focusing on quantitative PK scaling and human early dose prediction (eD2M). Some of the most important physicochemical and Drug Metabolism and Pharmacokinetics (DMPK) properties, which need to be understood and optimized as early as possible to allow a compound to meet the required quality are solubility, permeability, oral bioavailability, clearance, volume of distribution and half-life. These properties are all dependent on partition coefficients, which could be derived from Solvation Free Energy (SFE), a thermodynamic quantity [5-9]. Improving our capability to estimate SFEs accurately will have a profound effect on our ability to predict these key properties and ultimately design more efficient drugs, in a shorter time frame with a reduced attrition.

SFE estimation methods can be grouped into two major categories, namely implicit and explicit solvent models. Implicit models, including the solvation models (SM)[10-13] polarizable continuum models (PCM)[14,15] and conductor-like screening model for realistic solvation (COSMO – RS)[16] were shown to work well for simple systems. However, the dependence of SFEs on various factors including temperature, pressure, concentration and cosolvent effects[17] together with the limited molecular-level information considered by implicit models frequently makes it challenging to use this approach for SFE calculations of more complex systems. Explicit solvent models, including Monte Carlo (MC) algorithms and Molecular Dynamics (MD) simulations could be an alternative to implicit models. Earlier studies have shown that MD could be used to predict Hydration Free Energies (HFEs) of a large set of chemically-diverse compounds with a root mean squared error of ~1 kcal/mol[18,19]. Nevertheless, the accuracy in MD comes at the expense of huge computational costs.

An alternative approach, Reference Interaction Site Model (RISM) complements implicit/explicit solvent models and it uses a set of integral equations to model the solvation effects in terms of correlation functions[20,21]. Apart from being reliable in terms of HFE estimations, RISM is computationally less expensive than MD simulations, which could be attributed to the use of correlation functions. RISM has

been frequently used for computing solvation free energies[17,22-27] predominantly in water. Palmer et al. showed a successful application of 1D version of RISM to predict permeabilities of drug-like molecules[28]. However, 1D-RISM does not correctly account for the spatial orientation of molecules in solution, which is often crucial for predicting permeabilities or other pharmacokinetic properties of large conformationally flexible molecules. The limited scope of using 1D RISM for modeling the solvation effects of large molecules has encouraged the use of more advanced 3D version of RISM theory in recent studies. It has been demonstrated that 3D-RISM can compute HFEs, as accurately as MD[17].

Despite the fact that quite a few studies have reported the use of 3D-RISM for theoretical computation of HFEs, it is often demanding to use 3D-RISM to compute SFEs in other organic solvents. Misin et al. have developed a method in which solvents were considered as Lennard Jones spheres and parameters derived from critical points were used to estimate SFEs in several nonaqueous solvents[17]. Later, Roy et al. published a guideline for 3D-RISM parameter optimizations related to estimation of SFEs in n-Octanol[29]. Despite these successful applications, technical challenges related to convergence issues and difficulties in optimizing the force field parameters for organic solvents frequently restricts the usage of 3D-RISM in SFE estimations. To overcome these limitations, ML approaches have been widely used to predict SFEs in pure organic solvents[30-35]. Katritzky and his co-workers have shown the application of Multiple Linear Regression (MLR) models to predict SFEs of multiple solutes in one solvent[30] and one solute in multiple solvents[31] by making use of the structural descriptors computed by CODESSA PRO. Yet another study by Delgano et al. demonstrates the use of MLR in predicting SFEs in octanol, using molecular descriptors generated by CODESSA[32]. Though several studies have reported the generation of predictive models specific to a solute or a solvent, Borhani et al. attempted to build hybrid Quantitative Structure Property Relationship (QSPR) models on SFE data corresponding to 295 solutes and 210 organic solvents. These hybrid Partial Least Squares (PLS) regression and MLR models that relied on the quantum mechanical descriptors of the solutes and the experimental solvent descriptors could predict the SFEs of 356 solute/solvent pairs in the test set with Root Mean Squared Errors (RMSEs) of ~0.52 – 0.58 kcal/mol[34]. Another recent study on Minnesota database is an example of a hybrid model that showed how deep learning models derived from SMILES-based encodings of solutes and solvents can predict SFEs of left out solute/solvent pairs during cross validation with RMSE of 0.57 kcal/mol[35]. Besides the successful applications of using 3D-RISM for HFE/SFE estimations and using ML models derived from structural and quantum mechanical descriptors for SFE predictions, some recent studies have combined the two approaches in the property prediction space. Recently, some of us reported accurate predictions of bioaccumulation factor using a 3D convolution neural network trained on 3D-RISM correlation functions[36]. Roy et al. have since shown accurate classification of blood-brain-barrier (BBB) permeability using a support vector machine trained on traditional 2D molecular descriptors and 3D-RISM thermodynamic output; a sensitivity of 0.99 and a specificity of 0.95 were reported for a single training and test set split[37]. We believe that the success of these models is related

to the accurate description of solvation effects with 3D-RISM. Being motivated by the fruitful combination of 3D-RISM and ML approaches, we aim to extend this application to predict SFEs in various organic solvents.

In this work, we developed and tested a multi-solvent model that utilizes not only the 3D-RISM hydration thermodynamics, but also solvent properties as descriptors to predict SFEs. We assess the validity and extrapolative capability of the models by conducting several extensive validations including Leave One Solvent Out (LOSO), Leave One Cluster Out (LOCO) and Leave a Fraction of Solutes/Solvents Out (LFSO). Besides generating predictive models, we also focus on highlighting the solute and solvent features, relevant for SFE predictions.

3D-RISM - Theory

3D-RISM allows modeling of the solvent density distribution in a system via 3D correlation functions[20,21]:

$$h_{\alpha}(\boldsymbol{r}) = \sum \int c_{\propto}(\boldsymbol{r})\, \chi_{\alpha\alpha'}(r - r')d\boldsymbol{r}_{\alpha'} \quad (1)$$

where $h_{\alpha}(\boldsymbol{r})$ represents the total correlation function, which is related to radial distribution function: $(h(\boldsymbol{r}) = g(\boldsymbol{r}) - 1)$, $c_{\alpha}(\mathbf{r})$ is the direct correlation function, α and α' correspond to the index of solvent sites (atoms), and $\chi_{\alpha\alpha'}$ is the bulk solvent susceptibility function, describing the solvent atom-atom correlations, obtained from 1D RISM. The 3D-RISM equation (Eq. 1) requires a closure relation, which can be expressed as[21,38,39]:

$$h(\boldsymbol{r}) = \exp\big(-\beta u(\boldsymbol{r}) + \gamma(\boldsymbol{r}) + B(\boldsymbol{r})\big) - 1 \quad (2)$$

where $\beta=1/(K_BT)$, K_B is the Boltzmann constant and T is the temperature. $u(\mathbf{r})$ is the solute-solvent interaction potential, $\gamma(\mathbf{r})$ is the indirect correlation function [$\gamma(\mathbf{r}) = h(\mathbf{r}) - c(\mathbf{r})$] and $B(\mathbf{r})$ is the bridge function. When B(r) = 0 in Eq. 2, it corresponds to the HyperNetted Chain (HNC) approximation of the closure relation (Eq. 3)[20]. However, the usage of HNC in 3D-RISM simulations is often limited by the poor convergence[40]. The most widely used closure is Partial Series Expansion (PSEn)[41] (n = 1 corresponds to Kovalenko Hirata (KH) closure[42]):

$$h_{\alpha}(\boldsymbol{r})_{HNC} = \exp\big(-\beta u_{\alpha}(\boldsymbol{r}) + \gamma_{\alpha}(\boldsymbol{r})\big) - 1 \quad (3)$$

$$h_{\alpha}(\boldsymbol{r})_{PSE-n} = \begin{cases} \exp\big(\Xi_{\alpha}(\boldsymbol{r})\big) - 1, if\ \Xi_{\alpha}(\boldsymbol{r}) \leq 0 \\ \sum_{i=0}^{n} \frac{(\Xi_{\alpha}(\boldsymbol{r}))^{i}}{i!} - 1, if\ \Xi_{\alpha}(\boldsymbol{r}) > 0 \end{cases} \quad (4)$$

where $\Xi_{\alpha}(\boldsymbol{r}) = -\beta u_{\alpha}(\boldsymbol{r}) + \gamma_{\alpha}(\boldsymbol{r})$

Within the RISM framework, the SFE functionals can be derived from Kirkwood's equation[43] using correlation functions[44]. SFE functionals corresponding to HNC and PSE closures can be represented as:

$$\Delta G^{HNC} = K_B T \sum \rho_\alpha \int [\tfrac{1}{2} h_\alpha^2(\boldsymbol{r}) - c_\alpha(\boldsymbol{r}) - \tfrac{1}{2} c_\alpha(\boldsymbol{r}) h_\alpha(\boldsymbol{r})]\, d\boldsymbol{r} \quad (5)$$

$$\Delta G^{PSE-n} = \Delta G^{HNC} - K_B T \sum \rho_\alpha \int [\tfrac{\Theta(\Xi_\alpha)(\Xi_\alpha(\boldsymbol{r}))^{n+1}}{(n+1)!}]\, d\boldsymbol{r} \quad (6)$$

where Θ is the Heavyside step function.

$$\Theta(x) = \begin{cases} 1 \text{ for } x > 0 \\ 0 \text{ for } x \leq 0 \end{cases}$$

In a number of works, it was shown that HFEs computed using theory-derived SFE functionals (Eq. 5-6) often have large deviations from the experimental data[45-48]. To solve this problem, several corrections were introduced, including Gaussian fluctuation (GF)[49,50]:

$$\Delta G^{GF} = K_B T \sum \rho_\alpha \int [- c_\alpha(\boldsymbol{r}) - \tfrac{1}{2} c_\alpha(\boldsymbol{r}) h_\alpha(\boldsymbol{r})]\, d\boldsymbol{r} \quad (7)$$

as well as corrections based on pressure and/or partial molar volume: PC/PC+[48] and UC[51]:

$$\Delta G^{X-PC} = \Delta G^{X} + \tfrac{1}{2}\rho k_B T \left(1 - \frac{1}{\rho k_B T Xi}\right)\overline{V} - (2\rho k_B T)\overline{V} \quad (8)$$

$$\Delta G^{X-PC+} = \Delta G^{X} + \tfrac{1}{2}\rho k_B T \left(1 - \frac{1}{\rho k_B T Xi}\right)\bar{V} - (\rho k_B T)\bar{V} \quad (9)$$

where *X* indicates the uncorrected 3D-RISM functional used to compute SFE; here HNC or PSE-n. $k_BTXi=(R/N) \times T \times c$; *R* is the gas constant, *N* is the Avagadro's number, *T* is the temperature and *c* is the compressibility extracted from 1D-RISM calculations on bulk solvent.

$$\Delta G^{X-UC} = \Delta G^{X} + a.PMV + b \quad (10)$$

where *PMV* is the partial molar volume and the empirical coefficients *a* and *b* in UC corrections are obtained by linear regression, using a dataset of molecules with known experimental SFE data. In the original UC model, PMV was multiplied by the bulk solvent number density, to make the coefficient "a" unit less. In this work, this step has been skipped and the original units are retained .

Eq. (8) - (10) have been shown to considerably improve the accuracies of the estimated HFEs (Standard Deviation of error: ~1 kcal/mol)[48,51].

Materials and Methods

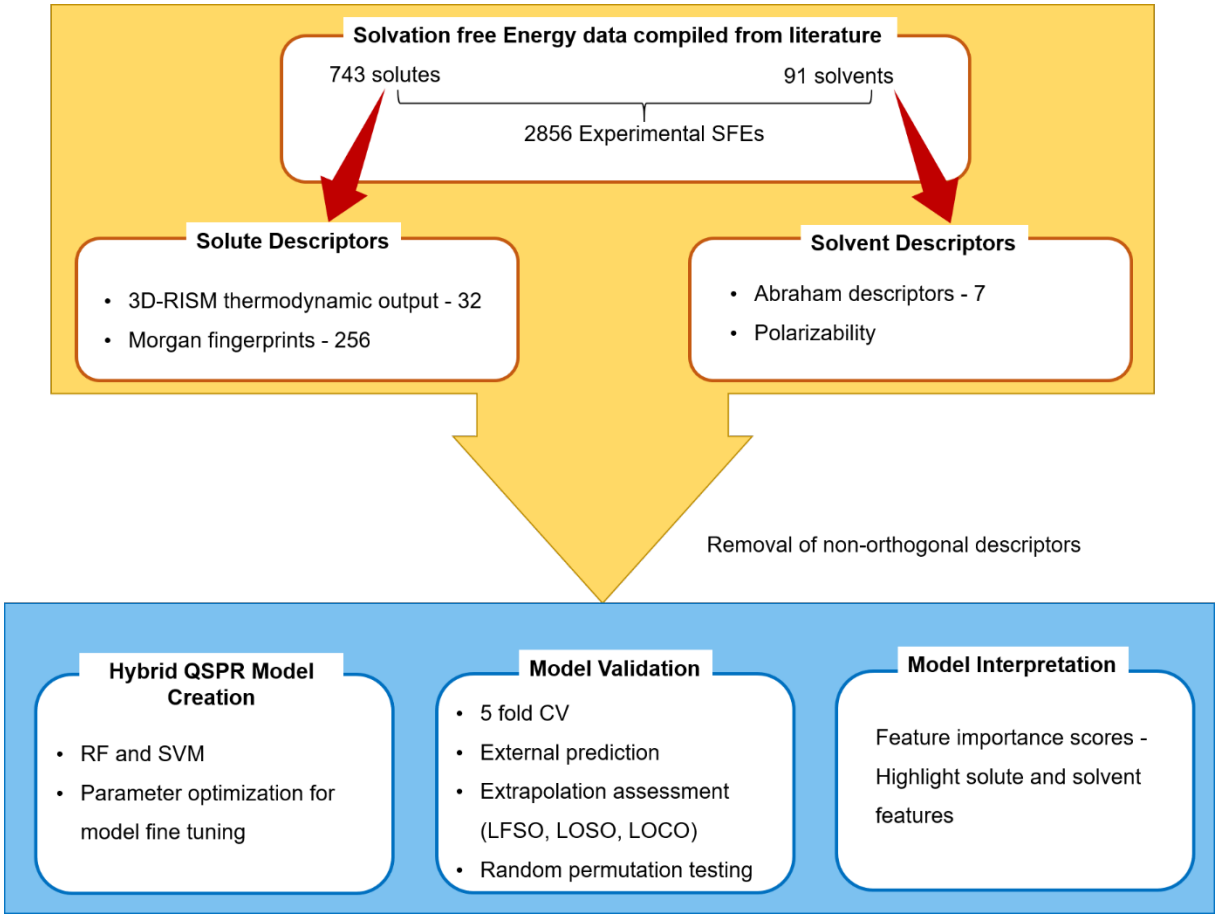


Figure 1. Workflow describing the steps involved in hybrid QSPR model generation

Experimental data

The solvation free energy dataset was compiled from multiple public sources[26,52,53]. Initially, we extracted 642 HFEs from FreeSolv[52], 822 HFEs from the dataset published by Roy et al.[26] and 3037 SFEs corresponding to 663 solutes and 106 solvents from Minnesota database[53]. All duplicates were removed and for those solutes with multiple HFE measurements across different sources, average values were used. Only SFEs of the neutral forms of the solutes were considered.

The final dataset covers 743 solutes and 91 pure solvents, but it is sparsely populated with only 2856 experimentally measured SFEs out of 743 × 91 = 67613 SFEs, which corresponds to 4.2% coverage. HFEs are available for 722 of the 743 solutes, but the coverage of non-aqueous SFEs is lower with 2134 SFEs reported for 315 solutes in 90 organic solvents.

Diversity of the dataset in terms of SFEs, dielectric constants of the solvents, heavy atom counts and SlogPs of the solutes is shown on Figures S1 - S3.

Molecular structures

For solutes in FreeSolv database, structures were generated in Maestro based on the available SMILES notation[54]. For solutes extracted from Roy et al., IUPAC names were used to extract the structures from PubChem[55]. Whereas, in Minnesota database, the structures provided in XYZ format were used as such[53]. All the 2D structures obtained from different sources were processed by using the ligprep module of Schrodinger package with the default settings, except that the neutral ionization states provided in the input were used and no tautomers were generated[56]. Multiple low energy conformers were generated and minimized by ligprep, but only the lowest energy conformer was used for 3D-RISM calculations.

For solvents, the structures were extracted from Minnesota database[53].

3D-RISM calculations

3D-RISM calculations were performed using the rism3d.snglpnt utility in AmberTools17[57]. All simulations were conducted using water as the solvent. A wide range of input parameters, including force fields, partial charges, water models, grid spacing, buffer distance and closures were optimized using FreeSolv data (See Figure 2 for details about the parameters tested). Only the 3D-RISM outcomes corresponding to the best set of parameters were considered for further analysis (see "Optimization of 3D-RISM input parameters for accurate hydration thermodynamics estimations" section for more information). Prior to RISM simulations, several preprocessing steps were carried out to assign charges and force field parameters with customized scripts. GAFF force field parameters and AM1-BCC / RESP charges were assigned using antechamber program in Amber[57]. OPLS 2005 charges and the corresponding force field parameters were designated using the Ligprep module[56] and the FFLD server utilities of Schrodinger software. In case of CM5 charges, Gaussian 16 was used for computations, using B3LYP/6-31G* basis set[58]. The charges were then extracted and assigned in Schrodinger software[54].

Besides the theory-derived SFE functionals (HNC, KH and PSE-3), we also employed the GF, PC/PC+ and UC corrected functionals for HFE calculations.

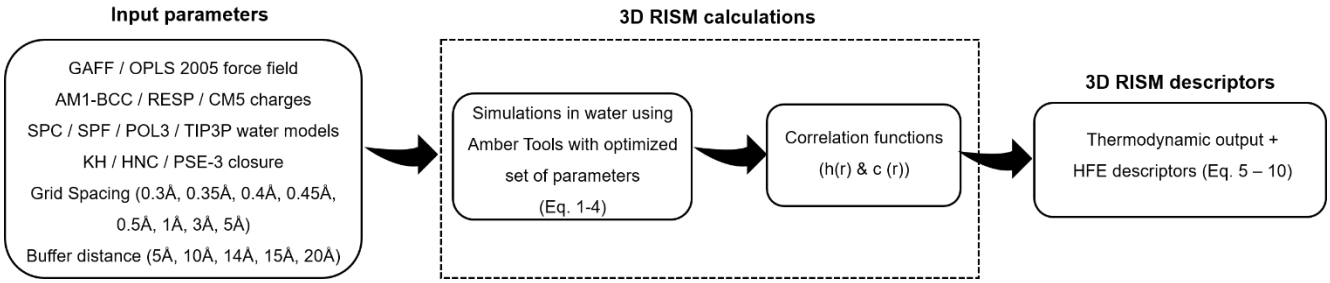


Figure 2. 3D-RISM simulations: From parameter optimization to descriptor generation

Solute and solvent descriptors

Thermodynamic output generated by 3D-RISM calculations (Figure 2) were used as solute descriptors (Table 1) for the ML models. Additionally, solutes were described by Morgan fingerprints, where the environment of every atom in a molecule is explored based on a user-specified diameter and the corresponding structural fragments are encoded in the form of bits[59]. Fingerprints were calculated by using the RDKIT modules in Python with the diameter set to 4 and the number of bits set to 256[60]. Solvents were characterized by Abraham descriptors[61-63] provided in Minnesota database[53] together with the polarizabilities calculated by Certara's D360 platform, based on the approach described in Miller, K.J., 1990[64]. Some of the Abraham descriptors used in our ML models include dielectric constant (eps), refractive index at the wavelength of Na line (n), hydrogen bond acidity (alpha), hydrogen bond basicity (beta, beta^2), macroscopic molecular surface tension (gamma), square of the fraction of non-hydrogen atoms that are either aromatic carbons or F/Cl/Br (phi^2, psi^2)

Highly correlated descriptors were removed based on a threshold set to 95%, prior to model building (see Table S1 in the Supporting Information for statistics regarding removed descriptors).

Abbreviations used	Solute descriptors
PSE-3, GF and PSE-3 / PC+	Excess chemical potentials based on different SFE functionals: PSE-3, Gaussian Fluctuations and pressure correction plus
Polar_PMV, Apolar_PMV, Polar_PMV_dT, Apolar_PMV_dT	Partial Polar Volume (Polar, Apolar and Temperature derivatives)
Solute_PE, Solute_PE_LJ, Solute_PE_LJ14, Solute_PE_Coulumb, Solute_PE_Coulumb14	Solute Potential Energy (Lennard Jones and Coulumb components)
Solvent_PE	Solvent Potential Energy
Polar_DirectCor_Integral, Apolar_DirectCor_Integral, Polar_DirectCor_Integral_dT, Apolar_DirectCor_Integral_dT	Direct Correlation Function integral (Polar, Apolar and Temperature derivatives)

Polar_TotalCor_Integral, Apolar_TotalCor_Integral, Polar_TotalCor_Integral_dT, Apolar_TotalCor_Integral_dT	Total Correlation Function integral (Polar, Apolar and Temperature derivatives)
Polar_TS_term, Apolar_TS_term	-Temperature*Solvation Entropy term (Polar and Apolar)
Bit_1…. Bit_256	Morgan Fingerprints

Table 1. Summary of solute descriptors used in our models

Model building and validation

We focused on building ML models that relied on both solute and solvent descriptors for predicting SFEs. The complete dataset was randomly split into training (80%) and test sets (20%) using the *train_test_split* function in scikit-learn module of python[65]. Random Forests (RF)[66] and Support Vector Machine (SVM)[67] models were built only on the training data and the test set was used as an external validation set. *RandomForestRegressor* and *SVR* functions available in the scikit-learn module of python were used for model building[65]. For RF models, a wide range of parameters including the number of trees (200 – 2000 with increment steps of 200), maximum depth (10 – 100 with increment steps of 10), minimum number of samples per split (2, 5, 10), minimum number of samples per leaf (1, 2, 4) and maximum number of features for the best split (auto, log2) were optimized, using a random search as implemented by the *RandomizedSearchCV* function in scikit-learn module in python[65]. 20 iterations of the random search algorithm were performed to minimize the mean squared error between predicted and experimental data, as computed from 5-fold cross-validation on the training data. This corresponds to setting the *scoring*, *cv* and *n_iter* parameters in the RandomizedSearchCV function to *neg_mean_squared_error*, *5* and *20*, respectively. For SVM models, cost (10^{-3} - 10^{8} with incremental powers of 10) and gamma (10^{-5} - 10^{4} with incremental powers of 10) parameters were optimized in a similar fashion. All SVM models were built with a Radial Basis Function kernel and solute/solvent descriptors were centered and scaled by subtracting their mean and dividing by their standard deviation, prior to model building. The means and standard deviations were computed from the training data for each model and were then used to center and scale the data to be predicted.

The optimized models derived from the best combination of parameters were then used to predict the external test set. Model performances were estimated by coefficient of determination (R^2) and Root Mean Squared Errors of Prediction (RMSEP) of the fitted data (R^2_{Tr}), cross validated data (R^2_{CV}, $RMSEP_{CV}$) and external test set (R^2_{Test}, $RMSEP_{Test}$).

Model training and validations were repeated 10 times by considering different training and test set splits to assess the stability of the models. Thus, the overall process used to build and validate the models

corresponds to a nested cross-validation in which 5-fold cross-validation was used for the inner loop and 10-fold Monte Carlo cross-validation was used for the outer loop. Overall, 9 RF and 9 SVM multi-solvent models were built with different combinations of solute and solvent descriptors.

External validations on 20% of the data excluded randomly, only shows the model's potential to predict unseen solute/solvent combinations. The model's extrapolative capabilities in predicting new solutes or solvents, were assessed by conducting other stringent validations, such as LFSO, LOSO and LOCO. In these validations, all observations corresponding to a specific solute or solvent were excluded completely from model building. In LFSO validations, models were built on the observations corresponding to 80% of the solutes or 80% of the solvents or 80% of the solutes and the solvents. Model performances were evaluated based on their ability to predict the left-out fractions (20%) of the solutes and the solvents. In LOSO validations, all observations pertaining to one solvent were excluded at a time. Models were built on the data for other solvents and the excluded solvent was used as a test set. LOSO was conducted only on the solvent space and not on the solutes. For each solute, SFEs are available only for 4 solvents on average, and for nearly 60% of the solutes, SFEs are known only for 1 solvent. The inadequate number of datapoints makes it unsuitable for extending the LOSO validations to solute space. In case of LOCO validations, all observations that belong to one cluster were excluded and models built by considering the observations from remaining clusters were used to predict the left-out cluster. Solutes were clustered based on their 3D-RISM thermodynamic output / Morgan Fingerprints. Whereas, solvent clusters were generated using the Minnesota solvent descriptors and polarizabilities. A k-means clustering approach was used with the *kmeans* function in the R Statistical Computing Environment with *nstart* (random number of samples) and *iter.max* (maximum iterations) parameters set to 10 and 50, respectively. Clustering was repeated with the cluster numbers being set from 2 to 50 for solutes and 2 to 15 for solvents. Calinski Harabasz (CH)[68] index was used as the criterion to choose the optimal number of clusters, which in turn resulted in 8 solute and 14 solvent clusters (Figure S4 and S5).

Besides these validations, the models were also subjected to random permutation testing by randomly shuffling the SFE values and refitting the models 20 times. Performances of the original and random models were then compared to assess overfitting.

Model interpretation

Solute and solvent descriptors that influence HFE / SFE predictions were analyzed based on the feature importance scores (mean decrease in impurity based on the variance explained) obtained from RF models. The scores across 10 splits were averaged for each descriptor and the mean values were sorted

to identify the 10 most relevant features. Considering the performances, only the RF models based on Morgan fingerprints and 3D-RISM thermodynamic output were interpreted.

Results and Discussion

Optimization of 3D-RISM input parameters for accurate hydration thermodynamics estimations

We conducted a systematic analysis on the FreeSolv dataset to analyze the parameters that are expected to influence the theoretically computed HFEs. The combination of settings used in 3D-RISM calculations for parameter optimizations are listed in Table S2. To identify the best closure approximation, we performed 3D-RISM calculations with KH, HNC and PSE-3 closures by using the default AM1-BCC charge and GAFF force field parameters. RISM calculations with HNC closure failed to converge for nearly 12% of the solutes, which is in line with literature data[22]. Therefore, we restricted our comparisons to KH and PSE-3 closures. On comparing the experimental HFEs to the theoretically computed values based on different corrections, we found that using the PSE-3 closure resulted in low RMSEs for most of the SFE functionals, when compared to the KH closure. The only exception being the GF/UC correction, for which the HFEs computed by PSE-3 closure have comparatively large deviations from the experimental HFEs, thereby resulting in high RMSEs. (Table S3). Taking advantage of the accurate HFE estimations provided by higher order closures, together with the minimalistic convergence issues, we chose PSE-3 closure for all subsequent analyses.

On investigating the force field and partial charge parameters (Table S4), we found that some combinations of force fields and charges yield better results (high R^2 and low RMSEs) than the others for selected SFE functionals. These trends are often inconsistent, which makes it challenging to choose the appropriate force field/charge. Nevertheless, considering the accuracy of OPLS force fields shown in previous benchmarking studies[69] and the lower dependence of CM5 charges on conformations and basis sets[70] motivated us to use OPLS force field and CM5 charges in all our future analyses. Further, assessing the influence of water models on HFE computations showed that there is no remarkable difference in terms of HFE estimations, irrespective of the water model used (Table S5). We, therefore chose the SPC water model that has been widely used in the previous works[24-26]. We also analyzed the grid parameters such as spacing (distance between grid points) and buffer (minimum distance between solute surface and grid box edge), which influence the way that the correlation functions [h(r) and c(r)] defined in RISM equation (Eq. 1) are represented. We found that changing the buffer size and spacing from the default 14Å and 0.5Å does not significantly influence the theoretically computed values, as their correlations with the experimental HFEs remain nearly the same (Tables S6 and S7). An exception to this trend is, increasing the grid spacing beyond 1Å leads to a significant drop in correlations and the error grows exponentially.

Overall, HFE correlations and RMSEs resulting from the RISM calculations with the optimal set of parameters (OPLS force field, CM5 charges, default buffer and grid spacing, SPC water model and PSE-3 closure) shows that the corrected functionals (PC, PC+ and UC) perform significantly better than the theory derived PSE-3 and GF functionals. Among the correction methods, UC functional provides the most accurate HFE estimations (Table 2 and Figure 3).

Free Energy Functional	**R^2**	**RMSE**
PSE-3	0.02	26.88
GF	0.42	10.18
PSE-3 / PC	0.77	2.79
PSE-3 / PC+	0.88	2.34
GF / UC	0.90	1.23
PSE-3 / UC	0.88	1.60

Table 2. Comparison of HFEs computed from 3D-RISM calculations on the optimal set of input parameters with that of the experimental values; Coefficient of determination and Root Mean Squared errors in kcal/mol are reported.

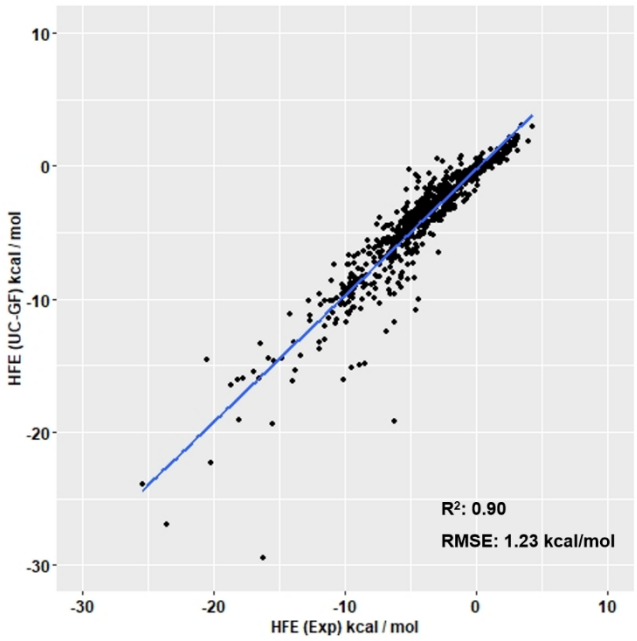


Figure 3. Scatter plot of the experimental HFEs and Universal corrections applied to GF functional (UC-GF), computed from 3D-RISM based on the optimal set of input parameters

Prediction of SFEs using machine learning

To understand the relevance of different types of solute descriptors as well as add-ons by solvent descriptors in predictive modeling, we created and tested multi-solvent models based on (i) solute descriptors only and (ii) in combination with solvent descriptors. Additionally, we generated models based on a null set that relies only on solvent descriptors to check for spurious correlations in the dataset. Comparisons of all of the RF and SVM models show that RF consistently outperforms SVM models both in terms of internal and external validations. We therefore used RF for all further comparisons, validations and interpretations. As would be expected, the performance of the models based only on solvent descriptors (Table 3, "S" descriptors) is very poor: R^2_{Tr}=0.11, R^2_{CV}=0.08, R^2_{Test}=0.08 and $RMSEP_{cv}$ = 2.85 kcal/mol, $RMSEP_{Test}$ = 2.78 kcal/mol. On the other hand, models derived from solute descriptors only (Morgan Fingerprints and 3D-RISM thermodynamic output) perform significantly better (Table 3, "F+R" descriptors), but their accuracy is not fully satisfactory with R^2_{CV} and R^2_{Test}, of 0.72 and 0.71, respectively. $RMSEP_{cv}$ and $RMSEP_{Test}$ of these models are 1.58 kcal/mol and 1.56 kcal/mol, respectively. The Morgan Fingerprints do not contribute much to these models, since models trained on 3D-RISM thermodynamic descriptors (Table 3, "R" descriptors) only give similar results; Whereas models trained on Morgan Fingerprints (Table 3, "F" descriptors) only are less predictive (R^2_{Test}: 0.61 and $RMSEP_{Test}$: 1.81 kcal/mol). However, combining the solvent descriptors with all of the solute descriptors (Table 3, "S+F+R" descriptors) significantly improves the predictions with R^2_{Test} increasing from 0.71 to 0.93 and $RMSEP_{Test}$ decreasing from 1.56 kcal/mol to 0.76 kcal/mol, thereby

demonstrating the value of using both solute and solvent descriptors for predicting SFEs in multiple solvents.

On assessing the performance of different solute descriptor sets combined with the solvent descriptors, we found that 3D-RISM thermodynamic output improved the CV and Test set correlations by about 15% (Table 3, "S+R" descriptors), when compared to the Morgan fingerprint-based models (Table 3, "S+F" descriptors). On the other hand, using the 2 descriptors as a combination resulted in nearly the same performance, as that of the models derived from 3D-RISM thermodynamic output with only minor variations in RMSEPs. Despite the fact that there are no significant improvements in terms of model performances, we used the descriptor combinations (3D-RISM thermodynamic output + Morgan Fingerprints) as solute descriptors in all our validations. We anticipate that incorporating the structural information in the form of fingerprints would have an impact on the SFE predictions in organic solvents. Our solvent-specific models on Octanol support this claim, as is evident from the interpretation of these models, where a fingerprint is identified as one of the top 10 features for SFE predictions in Octanol (Figure S8). The best performing RF models on Morgan Fingerprints/3D-RISM thermodynamic output along with the solvent descriptors (Table 3, "S+F+R" descriptors) resulted in R^2_{Test} of 0.93 and $RMSEP_{Test}$ of 0.76 kcal/mol.

In order to evaluate the strength of using 3D-RISM hydration thermodynamic output in predicting SFEs, we built a model considering the experimental data from all organic solvents, except water. These models (R^2_{Test}: 0.94; $RMSEP_{Test}$: 0.59 kcal/mol) have better predictive performances than the hybrid models (R^2_{Test}: 0.93; $RMSEP_{Test}$: 0.76 kcal/mol) that included water-specific experimental data (Table 3 and Figure 4). The results clearly support our hypothesis of predicting SFEs of any solute in any organic solvent by exclusively using the 3D-RISM thermodynamic output from water.

To further validate the hybrid QSPR model that uses both solute and solvent descriptors, we compared its performance to that of solvent-specific models, where only the solute descriptors are used. Solvent specific models were built only for water, octanol and hexadecane, as these are the solvents with reasonable number of datapoints (at least 150). In terms of model performances, solvent specific models have RMSEPs between 0.94 kcal/mol and 1.33 kcal/mol (Table 3 and S8). The inability to generate a reasonable QSPR model for majority of the solvents, the limited applicability domain of the models resulting from fewer datapoints, and the high RMSEPs of the external test sets clearly demonstrate the need for a hybrid QSPR model that relies on both solute and solvent descriptors.

Descriptor sets	**RF**					**SVM**				
	R^2_{Tr}	**R^2_{cv}**	**$RMSEP_{cv}$**	**R^2_{Test}**	**$RMSEP_{Test}$**	**R^2_{Tr}**	**R^2_{cv}**	**$RMSEP_{cv}$**	**R^2_{Test}**	**$RMSEP_{Test}$**

All solvents models (only solute / only solvent descriptors)										
S	0.11	0.08	2.85	0.08	2.78	0.08	0.06	2.88	0.07	2.80
F	0.75	0.62	1.84	0.61	1.81	0.68	0.52	2.06	0.51	2.04
R	0.83	0.72	1.57	0.71	1.56	0.73	0.66	1.73	0.65	1.72
F + R	0.84	0.72	1.58	0.71	1.56	0.72	0.61	1.84	0.61	1.82
All solvents hybrid models										
S + F	0.96	0.77	1.42	0.77	1.38	0.90	0.68	1.68	0.68	1.65
S + R	0.99	0.93	0.82	0.93	0.77	0.95	0.92	0.85	0.92	0.82
S + F + R	0.99	0.92	0.83	0.93	0.76	0.97	0.83	1.23	0.84	1.15
Organic Solvents hybrid models										
S + F + R	0.99	0.93	0.62	0.94	0.59	0.96	0.86	0.77	0.87	0.85
Water models (Solvent-specific models with only solute descriptors)										
F + R	0.98	0.90	1.33	0.92	1.16	0.97	0.87	1.48	0.88	1.37

Table 3. Performances of SFE prediction models based on 2 ML approaches (RF: Random Forests; SVM: Support Vector Machines) and different descriptor combinations: S – Solvent Descriptors (ones from Minnesota + Polarizability); F – Morgan fingerprints as solute descriptors; R – 3D-RISM hydration theromodynamic output as solute descriptors. Coefficient of determination (R^2_{Tr}: Trained models; R^2_{CV}: 5-fold Cross Validation; R^2_{Test}: Test set) and Root Mean Squared errors ($RMSEP_{cv}$: Cross Validation; $RMSEP_{Test}$: Test set) resulting from internal and external validations are reported.

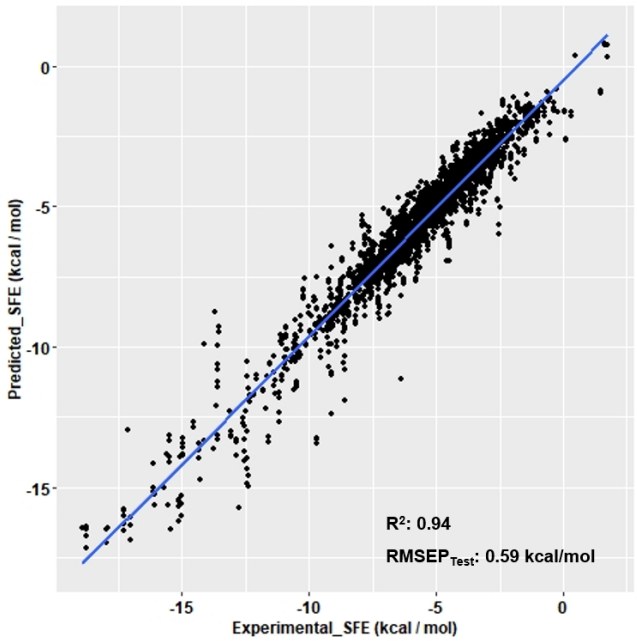


Figure 4: Scatter plot of the experimental SFEs and average test set predictions resulting from the RF models for organic solvents (HFE excluded).

In addition to internal and external validations, we also evaluated the effects of overfitting by means of y-scrambling, where we rebuilt the models 20 times with randomly shuffled experimental SFEs and estimated their performances. Comparing the Pearson correlation coefficients between the actual SFE values and the randomly shuffled values against R^2_{Tr} / R^2_{cv} shows that R^2_{Tr} of the null models was always less than 0.2 and R^2_{cv} remained close to 0 or negative. This further confirms that our models are not prone to overfitting and can be used validly for all subsequent analyses and interpretation.

Leave a Fraction of Solutes/Solvents Out (LFSO) validation: An assessment of the model's extrapolation capability

To minimize the efforts to experimentally measure SFEs of all solutes in all solvents, it is useful to generate a model with reasonable extrapolative capabilities. We therefore investigated our model's potential to predict SFEs of unseen solutes and solvents by building models that excluded a subset of them and predicted the left-out solutes/solvents (Table 4). RF models built on observations pertaining to 80% of the solutes (594) resulted in R^2_{Test} predictions of 0.87. On the other hand, when 20% of the solvents were excluded, R^2_{Test} predictions dropped to 0.58. Further, when 20% of both the solutes and the solvents were excluded from the models, the predictions were even worse with R^2_{Test} dropping further by 13% to 0.45 and $RMSEP_{Test}$ increasing to about 2 kcal /mol. Exploratory analysis on the individual models from different splits showed that when observations corresponding to water were excluded from

the modeling, the model performances dropped significantly. Cluster analysis of the solvent descriptor space further confirms this tendency (Figure S6). Water remains as a singleton cluster and it is rather distant from most of the solvents. Though tetrahydrothiophenedioxide and methyl formamide solvent clusters remain slightly close to water, they have limited SFE data for only a few solutes. Therefore, it is quite challenging to predict HFEs with models exclusively based on other organic solvents. However, HFEs /SFEs of excluded solutes can be predicted much more accurately, as the solutes are quite similar in terms of the descriptor space. Yet another issue with predicting both new solutes and solvents is the sparse solute-solvent data matrix, which in turn limits the extrapolative power of the models.

Prediction set	**Data used for Modeling**			**RF**	
	Solutes	**Solvents**	**Data points (Avg. across 10 splits)**	**R^2_{Test}**	**$RMSEP_{Test}$**
New solutes	594	91	2306	0.87	1.13
New Solvents	743	73	2180	0.58	1.77
New Solutes and new solvents	594	73	1760	0.45	1.97

Table 4. Performances of SFE prediction models based on a subset of solutes and solvents using Morgan fingerprints and 3D-RISM thermodynamic output. Coefficient of determination (R^2_{Test}) and Root Mean Squared errors ($RMSEP_{Test}$) resulting from the SFE predictions of new solutes and new solvents are reported.

Leave One Solvent Out (LOSO) validation: An assessment of model's sensitivity to solvent data

To assess the model's robustness with the exclusion of different solvents, and to analyze the model's capability to predict SFEs in unseen solvents, we left out one solvent at a time, built models on the observations corresponding to remaining solvents and predicted the excluded solvent. Model performances remained nearly the same (Table S9), as that of the models based on all solvents (Table 3) with R^2_{Tr} always being 0.99, R^2_{CV} ~0.93 and $RMSEP_{CV}$ <1 kcal/mol. This trend was observed, irrespective of the solvent or the number of data points excluded. However, there is a lot of variation with respect to the predictions of the excluded solvent ($RMSEP_{Sol}$: 0.11-3.91 kcal/mol). Nearly 95% of the solvents can be predicted with $RMSEP_{Sol}$ < 1 kcal/mol. An exception to this trend is the predictions from octanol ($RMSEP_{Sol}$: 1.58 kcal/mol) and water ($RMSEP_{Sol}$: 3.91 kcal/mol). Of the 91 solvents, acetic acid, ethanol, isopropanol and nitroethane have R^2_{Sol} <0.5, despite having RMSEPs below 1 kcal/mol.

The low R^2_{Sol} could be attributed partly to the narrow range of SFE values for these solvents. The different behavior of water, when compared to other solvents and exclusion of a considerable number of datapoints makes it quite challenging for the models to predict HFEs. These results further support the limitations of our models used to predict new solvents and new solutes/new solvents.

Leave One Cluster Out (LOCO) validation: Probing applicability domain

To evaluate the effectiveness of the models in predicting solutes and solvents with different chemical profiles and properties, we conducted LOCO validation. Training performances of models built by excluding clusters of solutes or solvents were on par with the models relying on the entire dataset (Tables S10 and S11). With reference to predictions of excluded solvent clusters, we found that SFEs in any organic solvent could be predicted with $RMSEP_{Test} \leq 1$ kcal/mol. An exception being water, whose $RMSEP_{Test}$ is ~3.9 kcal/mol. Challenges with predicting SFEs in water have been discussed previously in LFSO and LOSO validation sections. As compared to the solvent clusters, it is quite demanding to predict SFEs of excluded solute clusters. Nearly 40% of the solute clusters, when excluded have $RMSEP_{Test}$ of over 2 kcal/mol. Clusters 2, 3 and 8, which had minimal overlap with the solutes from other clusters were among the poorly predicted ones (Figure S7). Furthermore, a comprehensive analysis of these clusters revealed that the poor predictions could be attributed to a few solutes containing sulphur dioxide, anthraquinone and thiophosphate groups in cluster 8. In case of cluster 3 with many solutes containing carbonyl groups and halogens linked to aromatic rings, the error is evenly distributed with nearly 80% of the solutes having SFE deviations of more than 1 kcal/mol, as against their experimental values. On analyzing cluster 2, where the majority of the solutes have less than 5 heavy atoms, we found that the amide containing solutes are the worst predicted ones. An interesting observation is SFEs of water in any solvent are predicted poorly with deviations of about 10 kcal/mol from the experimental values. Altogether, our results suggest that it might be challenging to predict SFEs of water in other solvents or SFEs of other solvents in water, when water is excluded completely from model building.

Analysis of feature relevance

Considering the model performance and the ease to decipher features related to SFE predictions, we focused on interpreting the RF models based on solvent descriptors and Morgan fingerprints/3D-RISM thermodynamic output. In case of both all solvent models and models relying on organic solvents, the high importance of solute-solvent interaction energies, PMV terms and dielectric constants of solvents show that these descriptors are important for predicting SFEs in any solvent. The relevance of pressure corrected HFEs together with the solvent descriptors like dielectric constant, polarizability, molecular surface tension, hydrogen bond acidity and refractive index in all solvent models reveal that these

features influence the SFE predictions in water (Figure 5a). On the contrary, for predicting SFEs in organic solvents, direct / total correlation function integrals, Lennard-Jones component of the solute potential energy, HFEs computed from Gaussian fluctuation functionals and hydrogen bond basicity of the solvents serve to be the most important contributors (Figure 5b).

In addition, we analyzed the solvent-specific models, to compare the features influencing SFE predictions in different solvents. In water models, we found that the pressure corrected HFE is the major contributor for SFE predictions (Figure 5c). We have shown before that HFEs based on PSE-3/PC+ correlate reasonably with the experimental values (R^2: 0.88). Therefore, it is not surprising that this solute descriptor is the most relevant one for predicting SFEs in water. By contrast, for predicting SFEs in octanol, solute-solvent interaction energy and the temperature derivative of the PMV term are the most relevant contributors (Figure S8). All the other solute descriptors have meager contributions to SFE predictions in both water and octanol. In case of hexadecane, RISM thermodynamic parameters with the top 10 importance scores are similar to the ones identified by models on all organic solvents (Figure S9). The consensus between the features highlighted to be important in multi-solvent and solvent-specific models demonstrate the model's ability to capture the right descriptors for SFE predictions.

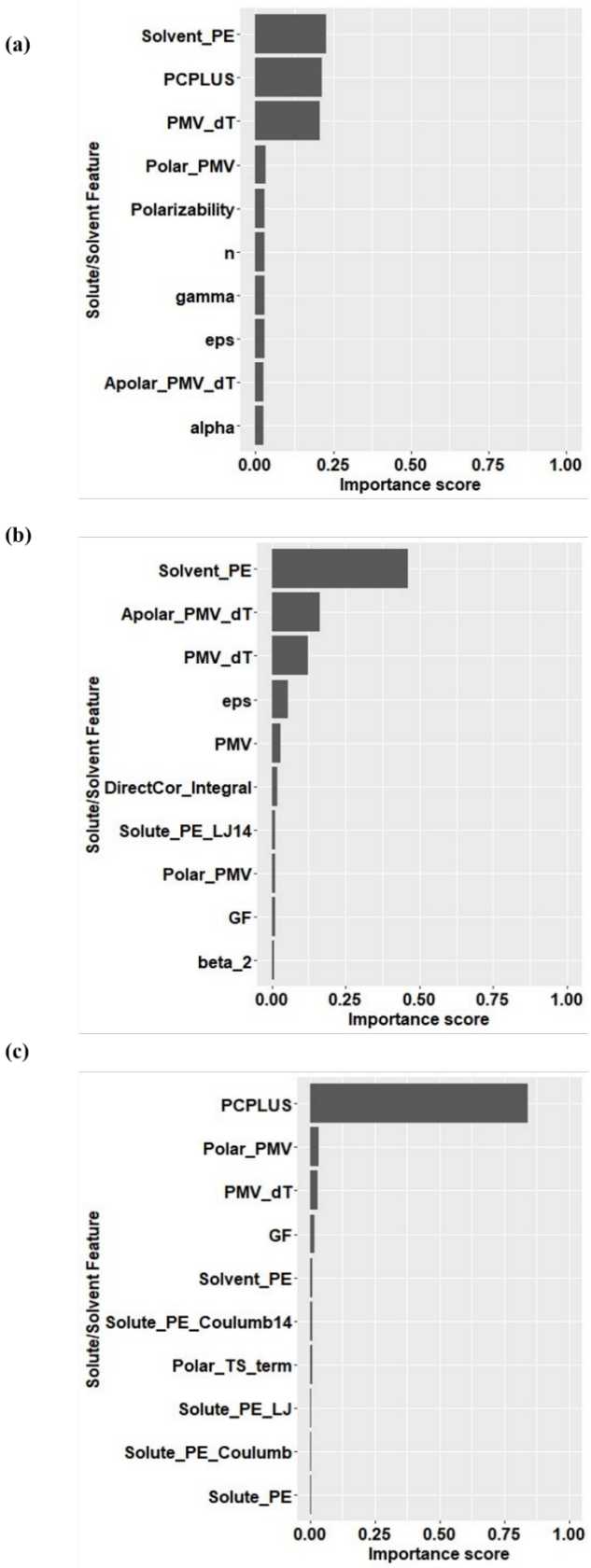


Figure 5. Feature importance score distributions of the top 10 features identified to be relevant by the RF model based on solvent descriptors and Morgan Fingerprints / 3D-RISM thermodynamic output. **(a)** Models on all

available solvent data **(b)** Models on organic solvents **(c)** Models on water. For more information on the features, refer Table 1 and "Solute and Solvent descriptors" section under Materials and Methods.

Comparison of predictive modeling to state of the art

It has been shown previously that theoretically computed HFEs from 3D-RISM have low discrepancies with the experimental values, when pressure or PMV based corrections are applied[48,51]. In our studies, HFEs that resulted from universal corrections applied to GF functionals correlated strongly with the experimental measurements (R^2: 0.90; RMSE: 1.23). In case of predictive modeling based on 3D-RISM thermodynamic output and Morgan fingerprints, this correlation remains nearly the same as that of our solvent-specific water models trained using ML ($R^2_{cv:}$ 0.90; R^2_{Test}: 0.92; $RMSEP_{Test}$: 1.16). However, our hybrid models on multiple solutes and solvents suggest that ML can be used to predict SFEs in any solvent efficiently (R^2_{cv}: 0.92; R^2_{Test}: 0.93; $RMSEP_{Test}$: 0.76). On the whole, our results show that predictive modeling using ML algorithms could be a great add-on to the 3D-RISM based theoretical computation of SFEs in any solvent, especially for those solvents that encounter convergence and force field parameterization issues in 3D-RISM simulations.

QM solvation models have been used quite extensively for estimating SFEs. We therefore compared our results on predictive modeling with the recently published SM12 solvation model, that influences SFE by introducing approximations to the electrostatic contributions[13]. Only the Minnesota solute/solvent pairs from our dataset have been taken into consideration. For aqueous data, the Mean Absolute Errors (MAEs) from SM12 models and our solvent-specific water models are 0.59-0.63 kcal/mol (Errors vary by the Density Functional Theory levels used) and 0.79 kcal/mol, respectively. On the other hand, for non-aqueous data, the predictive models offer better estimations of SFEs by reducing the MAEs from 0.54 kcal/mol in SM12 models to 0.37 kcal/mol. However, it is quite challenging to compare the two studies due to the variation in datapoints used for modeling. This issue can be connected to the fact that our test set splits are random and some of the solute/solvent pairs might never be included in the test set and are therefore not predicted; Whereas, some of them might be included twice in the different splits. The errors, we report here are the average values based on the solute/solvent pairs, included in the10 test set splits. Although the outcomes are not directly comparable, we anticipate that the SM12 solvation models and our predictive hybrid QSPR models would complement each other.

Yet another example of using predictive modeling in SFE estimations is Delfos, a deep learning model on all Minnesota solutes and solvents developed using SMILES based encodings[35]. It is not ideal to compare our models based on 3D-RISM hydration thermodynamic output and Morgan fingerprints with Delfos, as the validation strategies are different. Delfos assesses the models using 10-fold cross-validation; Whereas, we have adopted a double cross-validation approach. Our prediction errors on

excluded solute clusters (RMSEP: 1.54 kcal/mol) are in line with the ones reported by Delfos (RMSEP: 1.61 kcal/mol). On the other hand, we could predict the left-out solvent clusters much more accurately (RMSEP: 0.87 kcal/mol) with errors reduced by about 50%, when compared to Delfos (RMSEP:1.45 kcal/mol). Nevertheless, it is important to remember that the number of solute and solvent clusters used in the two studies are different. Overall, our results show that a RF model with 3D-RISM output and fingerprints can be as predictive as a sophisticated deep learning model, which is often associated with tuning of several hyperparameters and huge computational costs.

A more recent study, which is an exemplification of the hybrid QSPR modelling approach attempted to predict SFEs of 295 solutes in 210 organic solvents[34]. The test set predictions of these linear models (RMSE: ~0.52 – 0.58 kcal/mol) based on solute's quantum mechanical descriptors and experimental solvent descriptors are in line with the predictions of our RF models on all organic solvents (RMSE: ~0.59 kcal/mol). Our models that utilize the 3D-RISM hydration thermodynamic output / Morgan Fingerprints and Abraham descriptors complement the previously generated hybrid QSPR models by including aqueous data and by using the 3D-RISM output as solute descriptors, which takes solvation effects into account. In addition, the extensive validations on solute and solvent space demonstrate our model's extrapolative capabilities to predict SFEs of unseen solutes / solvents. Further, our study serves as a proof of concept for combining 3D-RISM with ML to predict physicochemical properties.

Conclusions

We have shown that a multi-solvent model derived from combination of ML approaches and advanced descriptors (i.e. solutes' 3D-RISM hydration thermodynamics together with solvent descriptors) can predict solvation free energies much more effectively than the theoretical approaches. The possibilities to predict solvation free energies in any organic solvent merely by using the 3D-RISM thermodynamic output from water opens up new avenues for extending this approach to predict other physicochemical properties, for instance, solubilities, permeabilities and partition coefficients. Our current models, being focused on small molecules, we avoided the problem of sophisticated conformational analysis. In the future work, focus will be laid on the exploration of conformational space that would not only improve the outcome of 3D-RISM simulations, but also support generation of models, suitable for permeability predictions of beyond-rule-of-5 compounds such as macrocycles and other new modalities.

Author Information

Corresponding Author

*Phone: (+46) 31 706 4132; Fax: (+46) 31 776 3748; E-mail: Antonio.Llinas@ astrazeneca.com

Notes

The authors declare no competing financial interest

Acknowledgements

Vigneshwari Subramanian is a fellow of the AstraZeneca R&D postdoc program.

References

(1) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discov.* **2004**, *3*, 711-716.

(2) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge *Nat. Rev. Drug Discov.* **2010**, *9*, 203-214.

(3) Cook, D.; Brown, D.; Alexander, R.; March, R.; Morgan, P. Lessons Learned from the Fate of AstraZeneca's Drug Pipeline: A Five-Dimensional Framework. *Nat. Rev. Drug Discov.* **2014**, *13*, 419–431.

(4) Morgan, P.; Brown, D. G.; Lennard, S.; Anderton, M. J.; Barrett, J. C.; Eriksson, U.; Fidock, M.; Hamrén, B.; Johnson, A.; March, R. E.; Matcham, J.; Mettetal, J.; Nicholls, D.J.; Platz, S.; Rees, S.; Snowde, M.A.; Pangalos, M.N. Impact of a Five-Dimensional Framework on R&D Productivity at AstraZeneca. *Nat. Rev. Drug Discov.* **2018**, *17*, 167–181.

(5) Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; Van Mourik, T.; Fedorov, M. V. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 3322–3337.

(6) Roy, D.; Hinge, V. K.; Kovalenko, A. Predicting Blood-Brain Partitioning of Small Molecules Using a Novel Minimalistic Descriptor-Based Approach via the 3D-RISM-KH Molecular Solvation Theory. *ACS Omega* **2019**, *4*, 3055–3060.

(7) Hinge, V. K.; Roy, D.; Kovalenko, A. Predicting Skin Permeability Using the 3D-RISM-KH Theory Based Solvation Energy Descriptors for a Diverse Class of Compounds. *J. Comput. Aided. Mol. Des.* **2019**, *33*, 605–611.

(8) Jensen, J. H. Predicting Accurate Absolute Binding Energies in Aqueous Solution: Thermodynamic Considerations for Electronic Structure Methods. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12441–12451.

(9) Best, S. A.; Merz, K. M.; Reynolds, C. H. Free Energy Perturbation Study of Octanol/Water Partition Coefficients: Comparison with Continuum GB/SA Calculations. *J. Phys. Chem. B* **1999**, *103*, 714–726.

(10) Cramer, C. J.; Truhlar, D. G. General Parameterized SCF Model for Free Energies of Solvation in Aqueous Solution. *J. Am. Chem. Soc.* **1991**, *113*, 8305–8311.

(11) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760–768.

(12) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on the Generalized Born Approximation with Asymmetric Descreening. *J. Chem. Theory Comput.* **2009**, 5, 2447–2464.

(13) Marenich, A. V; Cramer, C. J.; Truhlar, D. G. Generalized Born Solvation Model SM12. *J. Chem. Theory Comput.* **2013**, *9*, 609-620.

(14) Tomasi, J.; Mennucci, B.; Cance, E. The IEF Version of the PCM Solvation Method: An Overview of a New Method Addressed to Study Molecular Solutes at the QM Ab Initio Level. *J. Mol. Struct. THEOCHEM*, **1999**, *464*, 211 - 226.

(15) Lin, S.T.; Hsieh, C.M. Efficient and Accurate Solvation Energy Calculation from Polarizable Continuum Models. *J. Chem. Phys.* **2006**, *125*, 124103.

(16) Klamt, A.; Eckert, F.; Reinisch, J.; Wichmann, K. Prediction of Cyclohexane-water Distribution Coefficients with COSMO-RS on the SAMPL5 data set. *J. Comput. Aid. Mol. Des.* **2016**, *30 (11)*, 959-967.

(17) Misin, M.; Palmer, D. S.; Fedorov, M. V. Predicting Solvation Free Energies Using Parameter-Free Solvent Models. *J. Phys. Chem. B* **2016**, *120*, 5724–5731.

(18) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(19) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies Using Molecular Dynamics Free Energy Perturbation and The OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.

(20) Beglov, D.; Roux, B. An integral equation to describe the solvation of polar molecules in liquid water. *J. Phys. Chem.* **1997**, *101*, 7821–7826.

(21) Hirata, F., Ed. Molecular theory of solvation; Kluwer Academic Publishers: Dordrecht, Netherlands, 2003.

(22) 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*, 1–11.

(23) Fedorov, M. V; Palmer, D. S.; Frolov, A. I.; Ratkova, E. L. Towards a Universal Model to Calculate the Solvation Thermodynamics of Druglike Molecules: The Importance of New Experimental Databases. *Mol. Pharm.* **2011**, *8*, 1423 – 1429.

(24) 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**, *115 (13)*, 6312–6356.

(25) Misin, M.; Fedorov, M. V.; Palmer, D. S. Hydration Free Energies of Molecular Ions from Theory and Simulation. *J. Phys. Chem. B* **2016**, *120*, 975–983.

(26) Roy, D.; Kovalenko, A. Performance of 3D-RISM-KH in Predicting Hydration Free Energy: Effect of Solute Parameters. *J. Phys. Chem. A* **2019**, *123*, 4087–4093.

(27) Tanimoto, S.; Yoshida, N.; Yamaguchi, T.; Ten-no, S. L.; Nakano, H. Effect of Molecular Orientational Correlations on Solvation Free Energy Computed by Reference Interaction Site Model Theory. *J. Chem. Inf. Model.* **2019**, *59*, 3770 – 3781.

(28) Palmer, D. S.; Mišin, M.; Fedorov, M. V.; Llinas, A. Fast and General Method to Predict the Physicochemical Properties of Druglike Molecules Using the Integral Equation Theory of Molecular Liquids. *Mol. Pharm.* **2015**, *12*, 3420–3432.

(29) Roy, D.; Blinov, N.; Kovalenko, A. Predicting Accurate Solvation Free Energy in N-Octanol Using 3D-RISM-KH Molecular Theory of Solvation: Making Right Choices. *J. Phys. Chem. B* **2017**, *121*, 9268–9273.

(30) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806–1814.

(31) Delgado, E. J.; Jaña, G. A. Quantitative Prediction of Solvation Free Energy in Octanol of Organic Compounds. *Int. J. Mol. Sci.* **2009**, *10*, 1031–1044.

(32) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110*, 5714–5789.

(33) Zhang, P.; Shen, L.; Yang, W. Solvation Free Energy Calculations with Quantum Mechanics/Molecular Mechanics and Machine Learning Models. *J. Phys. Chem. B* **2019**, *123*, 901–908.

(34) Lim, H.; Jung, Y. Delfos: Deep Learning Model for Predicting Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306-8315.

(35) Borhani, T.N.; Garcia-Munoz, S.; Luciani, C.V. Galindo, A.; Adjiman, C.S. Hybrid QSPR Models for the Prediction of the Free Energy of Solvation of Organic Solute / Solvent Pairs. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13706–13720.

(36) Sosnin, S.; Misin, M.; Palmer, D. S.; Fedorov, M. V. 3D Matters! 3D-RISM and 3D Convolutional Neural Network for Accurate Bioaccumulation Prediction. *J. Phys. Condens. Matter* **2018**, *30*, 32LT03.

(37) Roy, D.; Hinge, V. K.; Kovalenko, A. To Pass or Not to Pass: Predicting the Blood-Brain Barrier Permeability with the 3D-RISM-KH Molecular Solvation Theory. *ACS Omega* **2019**, *4*, 16774–16780.

(38) Hansen, J.-P.; McDonald, I. R. Theory of Simple Liquids, 4th ed; Elsevier Academic Press: Amsterdam, The Netherlands, 2000.

(39) Duh, D. M.; Haymet, A. D. J. Integral-Equation Theory For Uncharged Liquids: The Lennard-Jones Fluid And The Bridge Function. *J. Chem. Phys.* **1995**, *103*, 2625–2633.

(40) Misin, M.; Fedorov, M. V.; Palmer, D. S. Communication: Accurate Hydration Free Energies at a Wide Range of Temperatures from 3D-RISM. *J. Chem. Phys.* **2015**, *142*, 091105 (1-6).

(41) Kast, S. M.; Kloss, T. Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J. Chem. Phys.* **2008**, *129*, 236101(1–3).

(42) 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.

(43) Frenkel, D.; Smit, B. Understanding molecular simulation; Academic Press, **2002**; p 672.

(44) Singer, S. J.; Chandler, D. Free-energy Functions in the Extended Rism Approximation. *Mol. Phys.* **1985**, *55*, 621–625.

(45) Guthrie, J.P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem.* B **2009**, *113*, 4501–4507.

(46) 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.

(47) Chuev, G. N.; Fedorov, M. V.; Crain, J. Improved estimates for hydration free energy obtained by the reference interaction site model. *Chem. Phys. Lett.* **2007**, *448*, 198–202.

(48) Sergiievskyi, V.; Jeanmairet, G.; Levesque, M.; Borgis, D. Solvation Free-Energy Pressure Corrections in the Three Dimensional Reference Interaction Site Model. *J. Chem. Phys.* **2015**, *143*, 184116 (1-6).

(49) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. An MM/3D-RISM Approach for Ligand Binding Affinities. *J. Phys. Chem. B* **2010**, *114*, 8505–8516.

(50) Chandler, D.; Singh, Y.; Richardson, D. M. Excess Electrons In Simple Fluids 0.1. General Equilibrium-Theory For Classical Hard- Sphere Solvents. *J. Chem. Phys.* **1984**, *81*, 1975–1982.

(51) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards a Universal Method for Calculating Hydration Free Energies: A 3D Reference Interaction Site Model with Partial Molar Volume Correction. *J. Phys. Condens. Matter* **2010**, *22*, 492101.

(52) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided. Mol. Des.* **2014**, *28*, 711–720.

(53) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database-Version 2012. *Univ. Minnesota, Minneap.* **2012**.

(54) Maestro 11.4, Schrödinger, LLC, New York, NY, 2017.

(55) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E.E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1).

(56) LigPrep, Schrödinger, LLC, New York, NY, 2017.

(57) Case, D.A. ; Cerutti, D.S. ; Cheatham, T.E.; Darden, T.A.; Duke, R.E.; Giese, T.J.; Gohlke, H.; Goetz, A.W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T.S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K.M.; Monard, G.; Nguyen, H.; Omelyan, I. ; Onufriev, A.; Pan, F.; Qi, R.; Roe, D.R.; Roitberg, A.; Sagui, C.; Simmerling, C.L.; Botello-Smith, W.M.; Swails, J.; Walker, R.C.; Wang, J.; Wolf, R.M.; Wu, X.; Xiao, L.; York, D.M. and Kollman, P.A. (2017), AMBER 2017, University of California, San Francisco.

(58) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B. and Fox, D. J. (2016), Gaussian 16, Revision C.01, Gaussian, Inc., Wallingford CT, 2016.

(59) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures — A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, 5, 107-113.

(60) Landrum, G. RDKit: Open-source cheminformatics Software; http://www.rdkit.org (accessed Aug 8, 2019).

(61) Abraham, M. H.; Whiting, G. S.; Doherty, R. M.; Shuely, W. J. Hydrogen Bonding. XVI. A New Solute Salvation Parameter, Π2H, from Gas Chromatographic Data. *J. Chromatogr. A* **1991**, *587*, 213–228.

(62) Abraham, M. H. Hydrogen Bonding. 31. Construction of a scale of solute effective or summation hydrogen-bond basicity. *J. Phys. Org. Chem.* **1993**, *6*, 660 – 684.

(63) Abraham, M. H. Application of Solvation Equations to Chemical and Biochemical Processes. *Pure Appl. Chem.* **1993**, *65*, 2503–2512.

(64) Miller, K. J. Calculation of the Molecular Polarizability Tensor. *J. Am. Chem. Soc.* **1990**, *112*, 8543–8551.

(65) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. **2011**, *12*, 2825–2830.

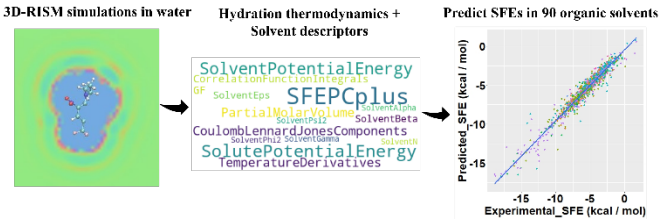
(66) Breiman, L. E. O. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(67) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.

(68) Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27.

(69) Caleman, C.; Maaren, P. J. Van; Hong, M.; Hub, J. S.; Costa, L. T.; Spoel, D. Van Der. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Com- Pressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. TheoryComput.* **2012**, *8*, 61–74.

(70) Marenich, A. V; Jerome, S. V; Cramer, C. J.; Truhlar, D. G. Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. TheoryComput.* **2012**, *8*, 527 – 541.


3D-RISM simulations in water
Hydration thermodynamics + Solvent descriptors
Predict SFEs in 90 organic solvents
SolventPotentialEnergy
CorrelationFunctionIntegrals
GF
SolventEps
SFEPCplus
PartialMolarVolume
SolventAlpha
SolventPsi2
SolventBeta
CoulombLennardJonesComponents
SolventPhi2
SolventGamma
SolventN
SoluteEnergy
TemperatureDerivatives
Predicted_SFE (kcal / mol)
0
-5
-10
-15
-15
-10
-5
0
Experimental_SFE (kcal / mol)