liquids

MDPI

Article

A Machine Learning Free Energy Functional for the 1D Reference Interaction Site Model: Towards Prediction of Solvation Free Energy for All Solvent Systems

Jonathan G. M. Conn, Abdullah Ahmad and David S. Palmer *

Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow G1 1XL, UK
* Correspondence: david.palmer@strath.ac.uk

Abstract: Understanding the interactions between solutes and solvents is vital in many areas of the chemical sciences. Solvation free energy (SFE) is an important thermodynamic property in characterising molecular solvation and so accurate prediction of this property is sought after. The One-Dimensional Reference Interaction Site Model (RISM) is a well-established method for modelling solvation, but it is known to yield large errors in the calculation of SFE. In this work, we show that a single machine learning free energy functional for RISM can accurately model solvation thermodynamics in multiple solvents. A convolutional neural network is trained on solvation free energy density functions calculated by RISM for small organic molecules in approximately 100 different solvent systems. We achieve an average RMSE of 1.41 kcal/mol and an R^2 of 0.89 across all solvent systems. We also compare the performance for the most and least commonly represented solvents and show that higher accuracy is generally seen with higher volumes of data, with RMSE values of 0.69–1.29 kcal/mol and R^2 values of 0.78–0.97 for solvents with more than 50 data points. We have shown that machine learning can greatly improve solvation free energy predictions in RISM, while demonstrating that the methodology is generalisable across solvent systems. This represents a significant step towards a universal machine learning SFE functional for RISM.

Keywords: solvation free energy; RISM; machine learning; solvent systems; convolutional neural network

Citation: Conn, J.G.M.; Ahmad, A.; Palmer, D.S. A Machine Learning Free Energy Functional for the 1D Reference Interaction Site Model: Towards Prediction of Solvation Free Energy for All Solvent Systems. *Liquids* **2024**, *4*, 710–731. https://doi.org/10.3390/liquids4040040

Academic Editors: William E. Acree, Jr., Juan Ortega Saavedra and Enrico Bodo

Received: 28 June 2024
Revised: 14 October 2024
Accepted: 19 October 2024
Published: 8 November 2024

1. Introduction

Solvation Free Energy

The process of solvation, as defined by Ben-Naim, is the transfer of one solute molecule from a fixed position in the ideal gas phase to a fixed position in the liquid phase at a given temperature and pressure [1]. The solvation free energy (SFE) is the reversible work associated with the transfer of the solute molecule from the gas phase to the liquid phase. SFE is important in many areas, such as drug discovery, as it may be used in the predictions of solubility and octanol-water partition coefficient (logP) [2–4], and environmental chemistry, as it is useful in understanding pollutant distributions in different aqueous environments [5,6].

Generally, the SFE of solutes with low vapour pressures must be obtained indirectly from separate measurements of the solubility and pure compound vapour pressure by a thermodynamic cycle via the gas-phase. Experimental SFE determination is often a lengthy and difficult process, so many computational approaches have been developed for its prediction.

SFE prediction methods can be separated into implicit and explicit methods. Implicit methods are characterised by a simplified representation of the solvent in which no solvent particles are explicitly modelled. Continuum models are the most common form of implicit methods, in which the solvent is represented by a homogeneous polarisable

medium described by a dielectric constant. A commonly used implicit model was developed by Cramer et al. [7], a universal model named SMD, which uses full solute electron density in place of atomic partial charges. The SFE consists of two components: bulk electrostatics calculated using the nonhomogeneous Poisson equation, and a cavity-dispersion term, which calculates the free energy required for forming a cavity in the solvent in which to place the solute, and short-range interactions between solute and solvent within the region of the first solvation shell. Cramer et al. also proposed the SMx models, the latest of which is SM12 [8–10], which implements the Generalised-Born (GB) approximation in conjunction with SASA to compute solvent effects. This model was parameterised for many different thermodynamic properties, including SFE predictions for both neutral and ionised species in both aqueous and organic solvents. The PCM by Tomasi et al. [11] implements a non-uniform dielectric treatment by accounting for the polarisability of the solvent and local polarisation effects around the solute. The PCM has been further developed in many ways, including the conductor-like PCM [12], which implements a boundary condition for conductor-like solvation screening, and the integral equation formalism [13,14], which allows bespoke dielectric treatment depending on the nature of chemical species. Additionally, the COSMO [15] approach calculates interaction energies between the solute and the dielectric continuum by determining the response of the continuum to screening charge distributions formed at the solvent accessible surface of the cavity formed around the solute. This boundary is discretised and the charge distributions are calculated, with exact solutions achievable for electric conductors.

Whilst implicit methods do provide reasonable accuracy and are computationally efficient, many physical phenomena which affect the solvation process, such as the reorientation of particles, are crudely approximated. This results in failure to accurately capture many local effects. The dielectric constant treatment may exhibit shortcomings in capturing local interactions which would be observed between solute and solvent molecules, such as electrostatic interactions at the solute–solvent interface, charge distribution changes in solvent molecules, and spatially induced changes to the solvent dielectric constant due to solute behaviour.

Explicit methods offer a more rigorous approach to simulating chemical systems by explicitly modelling both the solute and solvent particles. This allows the simulation of specific solute–solvent interactions in which local effects may be thoroughly probed. This additional level of molecular representation allows for a more physically realistic approach to the simulation of chemical systems. Populating a virtual space with a large number of additional simulated particles, i.e., solvent molecules, increases the computational expense considerably when compared to implicit models. Explicit solvent simulations are not generally amenable to high-throughout computing because of their computational expense. Additionally, the values of structural and thermodynamic data obtained from simulation may depend on the extent to which the configurational space has been sampled, with the result that there is a trade-off between efficiency (simulation time) and repeatability that introduces statistical noise to property estimates.

Explicit solvent methods for the calculation of SFE are commonly based on molecular dynamics (MD) or Monte Carlo simulations. Alchemical free energy calculations can be used to compute SFE. Sherman et al. [16] employed MD/free energy perturbation simulations to calculate the SFE as a transfer energy of a solute molecule transitioning from the ideal gas phase to the solvent, reporting an error of 1.1 kcal/mol. Leung et al. [17] used ab initio MD alongside thermodynamic integration to calculate the SFE of various ions to within approximately 2.5% of experimental observations. Although methods based on explicit solvent simulation (MD/MC) can provide accurate estimates of solvation free energies for small organic solutes in favourable circumstances, they are less reliable for larger, polyfunctional organic molecules, where issues caused by inaccurate forcefields or incomplete sampling of conformational/configurational space become more pronounced. A good example of these issues is provided by the results of the SAMPL challenges, in which entrants were asked to make blinded predictions of the HFEs of pharmaceutically

and environmentally relevant molecules [18–20]. The best predictions were in the range RMSE = 2.5–3.5 kcal/mol, which equates to an ∼2 log unit error in the related equilibrium physicochemical property (e.g., solubility, pKa, etc.). As the authors of these challenges observed, "Accurate calculations of more complex properties or events will remain over the horizon until these more basic values [HFEs] can be predicted with greater accuracy" [18]. Moreover, computing power limits the size of the system that can be reliably modelled using explicit solvent simulation, and even for smaller systems may limit the number of simulations that can be run. Given that every solute and solvent molecule within the system is modelled, many interactions must be computed at any given system state, particularly when compared to implicit methods. This computational expense increases with larger systems, requiring the use of more powerful processing units. For large systems, additional considerations are often required to improve the computational feasibility, and techniques such as coarse-graining have proven effective in this endeavour [21,22].

Comparisons between the use of implicit and explicit models for the prediction of SFE have been the focus of many studies. Shivakumar et al. [23] calculated the absolute hydration free energies of 239 ligands commonly used in drug-like compounds using both explicit and implicit water models. It was found that the explicit method generally outperformed the implicit method, but that the implicit method had good agreement with the explicit method when using the Poisson–Boltzman (PB) model. Additionally, the implicit method was much more sensitive to parameter changes, such as atomic charges and dielectric coefficients. The ESE-GB-DNN method proposed by Vyboishchikov employs terms calculated from a Generalized-Born implicit solvent model as input to a deep neural network for the calculation of SFE [24]. Steinmann et al. [25] focused on an organic adsorbate chemisorbed at a metal–liquid interface and compared the SFE calculations of an implicit and explicit approach. It was found that each method yielded the same general conclusions in terms of the stability profiles of different chemical species, but that the explicit method had generally better agreement with experimental findings. Interestingly, neither method accurately described the adsorption strength of water on the metal surface. VanderSpoel et al. [26] explored the use of implicit and explicit approaches to calculate the SFE of organic solutes in organic solvents, using various different methods. It was observed that the implicit methods were less accurate than explicit methods, but that the PB implicit method was in much closer agreement to both the explicit and experimental results than the GB variants. Errors of 15 kJ/mol and 6 kJ/mol were observed for the implicit GB and PB methods, respectively, while the explicit method yielded an error of 4.8 kJ/mol.

An alternative approach for modelling molecular solvation is the Reference Interaction Site Model (RISM), initially proposed by Chandler et al. [27], which captures specific solute–solvent interactions while circumventing the need for large-scale modelling of many solute and solvent molecules. The RISM models are a method derived from integral equation theory (IET), which captures how a solute and solvent interact in such a way that the solvent density distribution around a solute molecule can be calculated. RISM comes in the form of 1D-RISM and 3D-RISM, which differ in the dimensionality of the integral equations involved. The approach of 1D-RISM solves integral equations between solute and solvent sites (atoms) over a radial distance, while 3D-RISM solves integral equations over a set of 3D coordinates, which allows for modelling of the spatial distributions of solvent sites around a solute molecule. With 1D-RISM models, both the solute and the solvent molecules are represented as a set of sites (atoms), while 3D-RISM considers the full solute molecule as a single entity. In recent years, 3D-RISM has been more commonly used than 1D-RISM.

In 2010, Palmer et al. [28] showed that 3D-RISM lacked the ability to accurately predict SFE, resulting in errors of approximately 17 kcal/mol, and so proposed a universal correction. Based on earlier findings by Chuev et al. [29], Palmer et al. [30], and Ratkova et al. [31], where it was found that the partial molar volume correlated with the difference between experimental SFE and that predicted by 1D-RISM methods, the partial molar volume was used with a scaling coefficient and a bias correction term to improve the 3D-RISM SFE predictions. Truchon et al. [32] showed that the non-polar component of 3D-

RISM is the cause for inaccurate SFE prediction, while the electrostatic component requires no correction. They implemented a correction factor such that the non-polar component better describes repulsions between the oxygen in water and the solute such that solvent cavity formation is better described. Sergiievsky et al. [33] developed a pressure correction (PC) and an expansion upon this correction (PC+), originally for molecular density functional theory, in order to compensate for the overpressure resulting from homogeneous reference fluid approximation. Misin et al. then applied these corrections to 3D-RISM calculations in SFE prediction, finding that the accuracy did increase with the corrections [34]. This correction was later applied alongside solvent LJ approximations for coarse-grained solvent models and non-polar or weakly polar solvents [35], followed by an additional study in the use of this pressure correction as a means to describe the effects that salt has on solutes [36].

The standard 1D-RISM methods commonly produce errors of up to 20 kcal/mol in the prediction of SFE and so are not routinely used for the calculation of solvation thermodynamics. Ratkova and Fedorov [37] developed a model in which hydration free energy (HFE) calculations using 1D-RISM are improved using a cheminformatics-based correction. A multilinear regression model is trained on molecular features, such as the partial molar volume and various structural descriptors, to offset the error between the experimental and 1D-RISM-calculated HFE values. Fowles et al. [38] developed a machine learning approach trained on 1D-RISM calculations to predict the SFE in water, chloroform, and carbon tetrachloride. RMSE values with the 1D-RISM calculations themselves ranged between approximately 16 and 44 kcal/mol for two of three investigated approaches, with the third approach showing much lower error of approximately 1.8 to 6 kcal/mol. After training a machine learning model on these data, the RMSE values then dropped to below 1 kcal/mol. This approach was later applied to a multi-output model for the simultaneous prediction of the free energy, enthalpy, and entropy of hydration [39].

In this work, we use a machine learning model trained on the output of 1D-RISM calculations to accurately predict the SFE of various solute/solvent combinations. Additionally, we expand the number of solvents with the aim to make it feasible to make accurate SFE predictions for any combination of solute and solvent using this approach.

2. Theory

2.1. 1D-RISM

RISM is a method derived from the Integral Equation Theory (IET) of liquids, which is an implicit solvent model which uniquely describes interactions between solute and solvent particles to yield radial distribution functions (RDFs), which describe solvent density around the solute. This approach is effective as it provides a sufficient molecular description of solute molecules while being less computationally intensive than explicit solvent simulation, allowing for fast calculations of SFE.

The IET of atomic liquids is described by the Ornstein–Zernike (OZ) equation, which defines the total correlation function between two spherically symmetric particles in a homogeneous solvent, shown in Equation (1):

$$h(r_{12}) = c(r_{12}) + \rho \int c(r_{13})h(r_{32})dr_3, \tag{1}$$

where $h(r_{ij})$ and $c(r_{ij})$ are the total and direct correlation functions, respectively, between particles i and j, and ρ is the solvent density. The OZ equation is only applicable to simple atomic liquids, which is insufficient for chemical applications where more complex liquids are routinely used. Hence, the OZ equation has been generalised to the Molecular Ornstein–Zernike (MOZ) equation, which considers non-spherical molecules and is shown in Equation (2):

$$h(r_{12}, \Theta_1, \Theta_2) = c(r_{12}, \Theta_1, \Theta_2) + \frac{\rho}{\mathbb{Z}} \int c(r_{13}, \Theta_1, \Theta_3)h(r_{32}, \Theta_3, \Theta_2)dr_3 d\Theta_3, \tag{2}$$

where r_{12} and Θ_1 and Θ_2 are the displacement and the orientation, respectively, of particles 1 and 2. In this case, r is defined by a set of coordinates, (x, y, z), and Θ is defined by three Euler angles, (ψ, θ, φ). $\mathbb{Z}$ is equal to 4π for linear molecules, where only two angles are required to describe the molecule, or $8\pi^2$ for non-linear molecules, where three angles are required.

The 1D-RISM approach uses a one-dimensional approximation of the MOZ equation, based on spherically symmetric site–site correlation functions. In this approach, a modelled molecule is considered as a set of sites, each of which is spherically symmetric, where every atom is a single site. There are three types of correlation function present in the RISM theory: intramolecular correlation functions, $\omega(r)$, total correlation functions, $h(r)$, and direct correlation functions, $c(r)$, each of which depends only on the radial distance, r, between sites and has no angular dependencies.

In RISM calculations, the derived integral equations are solved in combination with a closure relation. The closure relation imposes limitations on the derived equations and ensures that there is self-consistency between pair correlation functions and the direct correlation function. To obtain solutions which are accurate beyond the mean-field approximation, it may be necessary for the integration over an effectively infinite series of integrals, which is not computationally feasible. Hence, a bridge functional is introduced to the closure relation, which quantifies the spatial contributions in the solvent density distribution and accounts for the correlations between solvent molecules. This reduces the necessity for an infinitely expansive integral series by helping to describe the part of the direct correlation function which cannot be easily determined from the solvent structure.

The structure of a molecule is described by the intramolecular correlation functions, $\omega(r)$. For two sites in a given molecule, denoted u and u', the intramolecular correlation function is shown in Equation (3):

$$\omega_{uu'}(r) = \frac{\delta(r - r_{uu'})}{4\pi r_{uu'}^2}, \tag{3}$$

where $r_{uu'}$ is the distance between the two sites and $\delta(r - r_{uu'})$ is the Dirac delta function. A single molecule will be described by a set of intramolecular correlation functions as these functions are site-pairwise in their description.

Intermolecular correlations between solute and solvent molecules are described by both pairwise total correlation functions and direct correlation functions, and are shown in Equation (4) for the solute site u and the solvent site v:

$$h_{uv}(r) = g_{uv} - 1, \tag{4}$$

where $h_{uv}(r)$ is the total correlation function and $g_{uv}(r)$ is the RDF of the solvent site around the solute site. Additionally, a total correlation function between sites of different solvent molecules, $h_{vv'}(r)$, is used to describe the distribution of solvent sites, v', of one molecule around the solvent site, v, of another solvent molecule. Hence, both solute–solvent and solvent–solvent total correlation functions are calculated.

In 1D-RISM, the total and direct correlation functions are related to one another through the derived integral equations, shown in Equation (5):

$$h_{uv}(r) = \sum_{u'=1}^{M} \sum_{v'=1}^{N} \int_{R^3} \int_{R^3} \omega_{uu'}(|\mathbf{r}_1 - \mathbf{r}'|) \times c_{u'v'}(|\mathbf{r}' - \mathbf{r}''|) \chi_{vv'}(|\mathbf{r}'' - \mathbf{r}_2|) d\mathbf{r}' d\mathbf{r}'', \tag{5}$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$ and $\chi_{vv'}(r)$ are the bulk solvent susceptibility functions and M and N are the number of sites of the solute and solvent, respectively. Bulk solvent sites have mutual correlations which are described by $\chi_{vv'}(r)$, which may be obtained via the solvent total correlation function, $h_{vv'}(r)$, and a 3D structure of the solvent molecule, and

therefore, the intramolecular correlation function, $\omega_{vv'}(r)$. Equation (6) shows the mutual correlation function:

$$\chi_{vv'}(r) = \omega_{vv'}(r) + \rho h_{vv'}(r), \tag{6}$$

where ρ is the bulk number density of the solvent.

To allow the total and direct correlation functions to be obtained by numerical solution of the RISM equations, $M \times N$ closure relations are introduced, the general form of which is shown in Equation (7):

$$h_{uv}(r) = e^{(-\beta u_{uv}(r) + \gamma_{uv}(r) + B_{uv}(r))} - 1, \tag{7}$$

where $u = 1, ..., M$, $v = 1, ..., N$, $\beta = 1/k_B T$, $u_{uv}(r)$ is a pair interaction potential between solute and solvent sites, and $B_{uv}(r)$ is a site–site bridge functional.

Generally, the interaction potential consists of short-range and long-range components. It is common for the short-range interactions to be described by an LJ potential, while the long-range interactions are described by an electrostatic term. Equation (8) shows the components of the interaction potential:

$$\begin{aligned} u_{uv}(r) &= u_{uv}^{el}(r) + u_{uv}^{LJ}(r), \\ u_{uv}^{el}(r) &= \frac{q_u q_v}{r}, \\ u_{uv}^{LJ}(r) &= 4\varepsilon_{uv}^{LJ}\left[\left(\frac{\sigma_{uv}^{LJ}}{r}\right)^{12} - \left(\frac{\sigma_{uv}^{LJ}}{r}\right)^{6}\right], \end{aligned} \tag{8}$$

where q_u and q_v are the partial charges of the solute and solvent sites, respectively, of interest, and ε_{uv}^{LJ} and σ_{uv}^{LJ} are the solute–solvent LJ parameters.

A bridge functional can be selected to increase the accuracy of the closure relations by accounting for higher-order correlations that are not considered in the mean-field approximation. The simplest bridge functional is the Hypernetted-chain (HNC) approximation, which sets $B_{uv}(r) = 0$. While this may be the most computationally inexpensive approximation, this results in a lesser degree of control in the exponent of Equation (7), which often leads to convergence issues, in which case, no solution is found. An improvement to this approximation is the Kovalenko–Hirata (KH) closure, which linearises the exponent above a threshold constant, C. The linearisation is shown in Equation (9):

$$h_{uv}(r) = \begin{cases} e^{\Xi_{uv}(r)} - 1 & \Xi_{uv}(r) \leq C, \\ \Xi_{uv}(r) + e^{C} - C - 1 & \Xi_{uv}(r) > C, \end{cases} \tag{9}$$

where $\Xi_{uv}(r) = -\beta u_{uv}(r) + \gamma_{uv}(r)$. When C tends to infinity, the KH closure becomes equal to the HNC closure. This approach allows the capturing of short-range interactions, while still setting them to 0 via the HNC closure in cases where observations of short-range interactions are not likely.

2.2. Solvation Free Energy Functionals

In a given system, the free energy of solvation may be obtained analytically through single-point calculations after the total and direct correlations functions have been computed by RISM. The HNC and KH closures have been used to derive SFE functionals for 1D-RISM, shown by Equations (10) and (11), respectively:

$$\Delta G_{solv}^{HNC} = 2\pi\rho k_B T \sum_{uv} \int_0^\infty [-2c_{uv}(r) - c_{uv}(r)h_{uv}(r) + h_{uv}^2(r)]r^2 dr, \tag{10}$$

$$\Delta G_{solv}^{KH} = 2\pi\rho k_B T \sum_{s=1}^{N} \sum_{a=1}^{M} \int_0^\infty [-2c_{uv}(r) - c_{uv}(r)h_{uv}(r) + h_{uv}^2(r)\Theta(-h_{uv}(r))]r^2 dr, \tag{11}$$

where ρ is the solvent bulk number density, k_B is the Boltzmann constant, T is the temperature, and $h(r)$ and $c(r)$ are the total and direct correlation functions, respectively. It has been found that the error of these functionals is too high for common use, with observed errors of the order of 20 kcal/mol. An alternative method is the Gaussian Fluctuations (GF) functional, shown in Equation (12):

$$\Delta G_{solv}^{GF} = 2\pi\rho k_B T \sum_{s=1}^{N}\sum_{a=1}^{M}\int_{o}^{\infty}[-2c_{uv}(r) - c_{uv}(r)h_{uv}(r)]r^2 dr. \quad (12)$$

Additionally, an improvement to the HNC functional was proposed: the repulsive bridge extension, shown in Equation (13):

$$\Delta G_{solv}^{HNCB} = \Delta G_{solv}^{HNC} + 2\pi\rho k_B T \sum_{uv}\int_{0}^{\infty}(h_{uv}(r)+1)(e^{-B_{uv}^{R}(r)-1)})r^2 dr, \quad (13)$$

where $B_{uv}^{R}(r)$ is the repulsive bridge correction function. The bridge functional for a given pair of solute, u, and solvent, s, sites is shown in Equation (14):

$$e^{-B_{uv}^{R}(r)} = \prod_{v'\neq v}\left\langle \omega_{vv'} \times e^{(\beta\varepsilon_{uv'}(\frac{\sigma_{uv'}}{r})^{12})}\right\rangle, \quad (14)$$

where $\omega_{uv'}$ are intramolecular correlation functions, and $\varepsilon_{uv'}$ and $\sigma_{uv'}$ are the LJ parameters for the site–site pairwise potential. Both the GF functional and the repulsive bridge extension to the HNC functional provide improved descriptions of the solvation thermodynamics compared to standard HNC and KH functionals [38,40].

2.3. Solvation Free Energy Densities

To define a machine learning SFE functional, we begin by noting that the standard RISM SFE functionals as described previously may be reduced into a general form, shown by Equation (15):

$$\Delta G_{RISM} = \int_{0}^{\infty} w(r)dr, \quad (15)$$

where the integrand, $w(r)$, consists of a function of the total and direct correlation functions of a single solute, as well as the prefactor, $2\pi\rho kT$. $w(r)$ may be used to discern a set of variables which quantify the solvation effects between solute and solvent molecules at a set distance, r, from the solute. This function, known as the solvation free energy density (SFED), is the input to the ML SFE functional. SFED functions may be derived from the previously mentioned functionals, HNC, KH, and GF, shown in Equations (16)–(18), respectively:

$$HNC_w(r) = 2\pi\rho kT \times \sum_{uv}[-2c_{uv}(r) - h_{uv}(r)(c_{uv}(r) - h_{uv}(r))], \quad (16)$$

$$KH_w(r) = 2\pi\rho kT \times \sum_{uv}[-2c_{uv}(r) - h_{uv}(r)c_{uv}(r) + h_{uv}^2\Theta(-h_{uv}(r))], \quad (17)$$

$$GF_w(r) = 2\pi\rho kT \times \sum_{uv}[-2c_{uv}(r) - h_{uv}(r)c_{uv}(r)]. \quad (18)$$

2.4. Model Performance

The predictive accuracy of a model is typically quantified by statistical analysis. Once a set of predicted values are output, they can be compared to the known target values and the agreement between these two sets can be measured in many ways.

Equations (19) and (20) show the formulae for the coefficient by determination (R^2) and the root mean squared error (RMSE), respectively.

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

$$RMSE(y, y_{exp}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y^i - y^i_{exp})^2}, \tag{20}$$

where index i runs through the set of N selected samples, $M()$ indicates a mean, and y^i and y^i_{exp} are the predicted and experimental target values, respectively. R^2 provides a measure of how well the output predictions match the target values with respect to the $y = x$ line, as a perfect set of predictions will reside on this line. RMSE provides a measure of the average difference between the predicted and target values, and also describes the combination of both the systematic and random error. The total deviation can be split into two parts: bias (or mean displacement, M), and the standard deviation of the error of prediction (SDEP), shown in Equations (21) and (22), respectively.

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

$$SDEP = \sigma(y - y_{exp}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y^i - y^i_{exp} - M(y - y_{exp}))^2} \tag{22}$$

The bias gives the systematic error, which can be corrected by the addition of a simple constant term in the final model. The SDEP gives the random error which is not explained by the model and cannot be corrected as bias can. The bias and SDEP are connected through the RMSE, given by Equation (23).

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

Models which report an 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.

3. Materials and Methods

3.1. Overview

Previously, work was undertaken to predict the SFE of small organic molecules in four different solvents; water, methanol, chloroform, and carbon tetrachloride [39]. A convolutional neural network (CNN) was trained on SFED data that resulted from 1D-RISM calculations for the accurate prediction of SFE. The focus of the work detailed in this paper is to expand this approach to a larger number of solvents such that the SFE of any solvent/solute combination can be predicted. Hence, SFE predictions have been made using a reoptimised CNN for over 100 different solvent systems with a variety of different small organic solutes. A workflow was developed for this to produce the required input files for the RISM calculations, as well as performing the RISM calculations and CNN predictions.

3.2. Dataset Compilation

Vermeire et al. [41] compiled several datasets containing various experimental SFE and enthalpy values at different temperatures. The work discussed in this paper made use of the CombiSolv-Exp (CSE) dataset. After applying structural filters and completing RISM

calculations, the resulting dataset consisted of 2698 samples, each consisting of a unique combination of solute and solvent.

In order to perform 1D-RISM calculations, the generation of several inputs was required. Forcefield parameters and structure files were required for both the solute and solvent molecule for a given pair, and the density of the pure solvent was also required. The general AMBER forcefield (GAFF) was used to model both the solute and solvent molecules and non-bonded parameters were obtained using Antechamber and tLEaP [42]. GAFF was used because it is a high-throughput method able to create a consistent set of parameters for all samples. The structure files were generated by converting SMILES strings to pdb files using OpenBabel [43] for all solutes and solvents. Solvent densities were obtained from various freely available libraries.

3.3. 1D RISM Calculations

PyRISM [44] was used to carry out 1D-RISM calculations using the KH closure within a system consisting of 16384 grid points over a 20.48 Å radius from the solute. The extended reference interaction site model (XRISM) was used for all calculations to allow for the same treatment of both organic and non-organic solvents.

The SFE calculations were performed using the KH, HNC, and GF free energy functionals, with the assumption of infinite dilution. The modified direct inversion of the iterative subspace (MDIIS) solver was used for these calculations as it achieves convergence quickly when compared to other solvers. The following parameters were used to solve the equations: lambda = 10, picard damping = 0.1, depth = 16, and tolerance = 10^{-5}. The lambda parameter dictates the extent of discretisation in the RISM calculation. With a lambda value of 1, a single calculation is performed before an output is given, while a lambda value of 10 means the calculation is performed a total of 10 times. A random guess, as described earlier, is used to perform the first calculation, while the output serves as the initial guess for the next calculation. This allows the calculation to iterate over several initial guesses to allow for a more robust methodology at the cost of increased calculation time. The Picard damping parameter applies an operator to the output of the RISM calculations to manipulate the extent to which the resulting $c(r)$ deviates from the initial $c(r)$. The higher the Picard damping value, the more the initial $c(r)$ is retained, resulting in slower convergence and decreased chance of divergence. With the MDIIS solver, a solution plane is explored and an optimal orthogonal plane is identified to increase the iterative subspace and increase chances of converging on a solution. The depth parameter sets the number of previous solutions that are saved, along with residuals, which are the difference between consecutive solutions. The residuals are then compared to the tolerance parameter. If the tolerance is met, then the calculation ends and produces an output, otherwise the calculation will continue until either the tolerance is met or it is determined that convergence is not possible.

3.4. SFED Processing

The SFED is a functional of the total and direct correlation functions between a solute and solvent molecule, dependent on the systems conditions (such as temperature) and the choice of RISM bridge and SFE functional. Here, SFED was calculated from Equation (18) based on the GF SFE functional. The choice of SFE functional has previously been shown to have little influence on the obtained form of the SFED function [38]. System conditions were kept constant throughout all calculations.

Post-processing of the SFED outputs involved grid reduction to remove highly correlated data. Every 40th grid point up to an 8Å radius was retained, which additionally removed noisy data corresponding to long-range separation between the solute and solvent. This resulted in a feature vector of length 160 that was used in the training of a machine learning model.

Figure 1 shows the total and direct correlation functions between solute and solvent sites. Figure 1a shows a selection of the solute–solvent site pairs between which total

correlation functions, $h_{s\alpha}$, are calculated. In practice, a total correlation function is calculated between every possible solute–solvent site pair, though in practice some may be identical due to symmetry in the solute or solvent. Hence, in the example of methanol and water, there are 18 total correlation functions calculated. The figure also illustrates an intramolecular correlation function, $\omega_{ss'}(r)$, in the methanol molecule. Each molecular structure involved in the RISM calculation will be described by intramolecular correlation functions, where one of these functions will be calculated for each site-pair within the molecule, e.g., water will be described by three intramolecular correlation functions; $\omega_{OH1}(r)$, $\omega_{OH2}(r)$, and $\omega_{H1H2}(r)$. Figure 1b shows the direct correlation functions between different particle sites. Figure 1c shows the total correlation functions highlighted by Figure 1a, while Figure 1d shows the direct correlation functions for the same solute–solvent site pairs, all plotted against the distance, r, between solute–solvent sites in Angstroms. Figure 1e shows the SFED calculated from the data in Figure 1c,d with Equation (15).

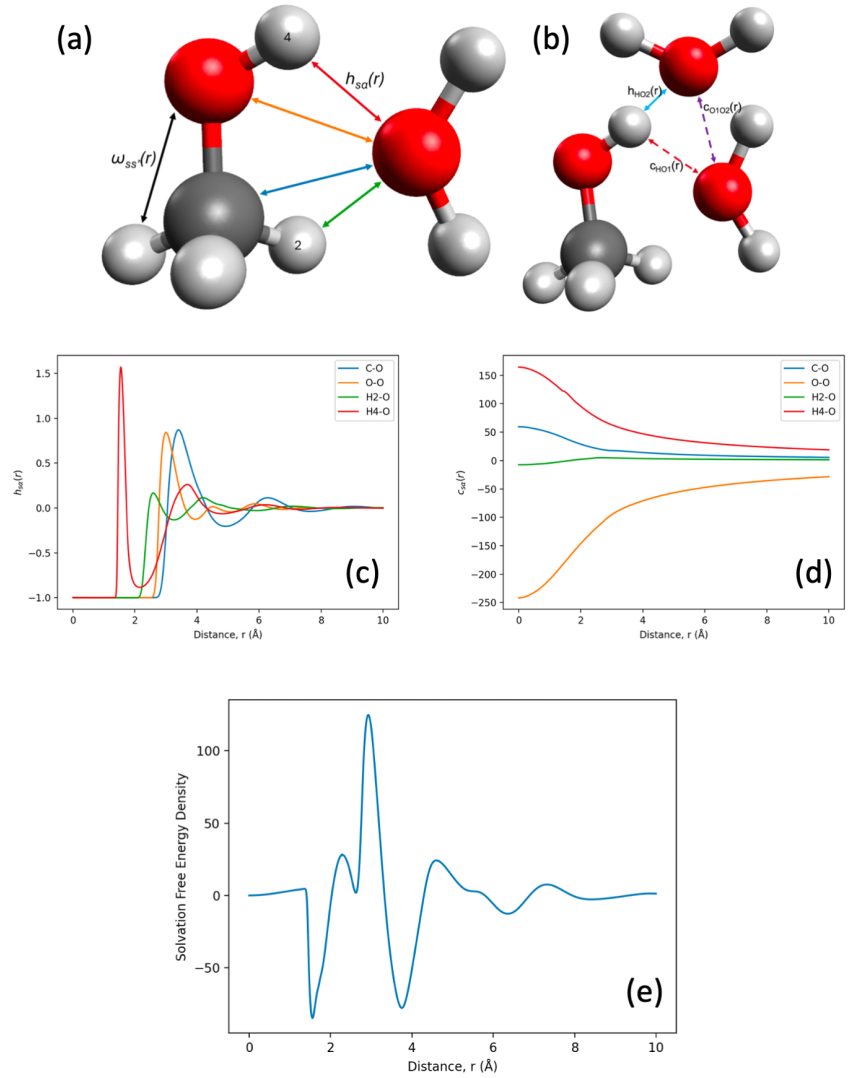

Figure 1. 1D-RISM correlation functions. (**a**) Intermolecular total correlation functions, $h_{s\alpha}(r)$, between a selection of solute sites and solvent sites. A total correlation function between every solute–solvent site-pair is calculated. Also shown is an intramolecular correlation function, $\omega_{ss'}(r)$, between two sites of the methanol. (**b**) Decomposition of $h_{OH4}(r)$ in terms of direct correlation functions $c_{s\alpha}(r)$ involving a third particle. (**c**) Total correlation functions corresponding to those shown in (**a**). (**d**) Direct correlation functions between the same solute–solvent site-pairs described in (**c**). (**e**) SFED calculated for methanol–water using $h_{s\alpha}(r)$ and $c_{s\alpha}(r)$, as well as a prefactor, $2\pi\rho kT$.

3.5. Final Dataset Analysis

The dataset contained 2689 data points comprising binary combinations of 108 unique solvents and 427 unique solutes. Figure 2a shows scores plots for principal component analysis (PCA) against the SFEDs calculated for all solute–solvent systems, while Figures 2b and 2c show the explained variance ratio (EVR) and cumulative explained variance (CEV), respectively.

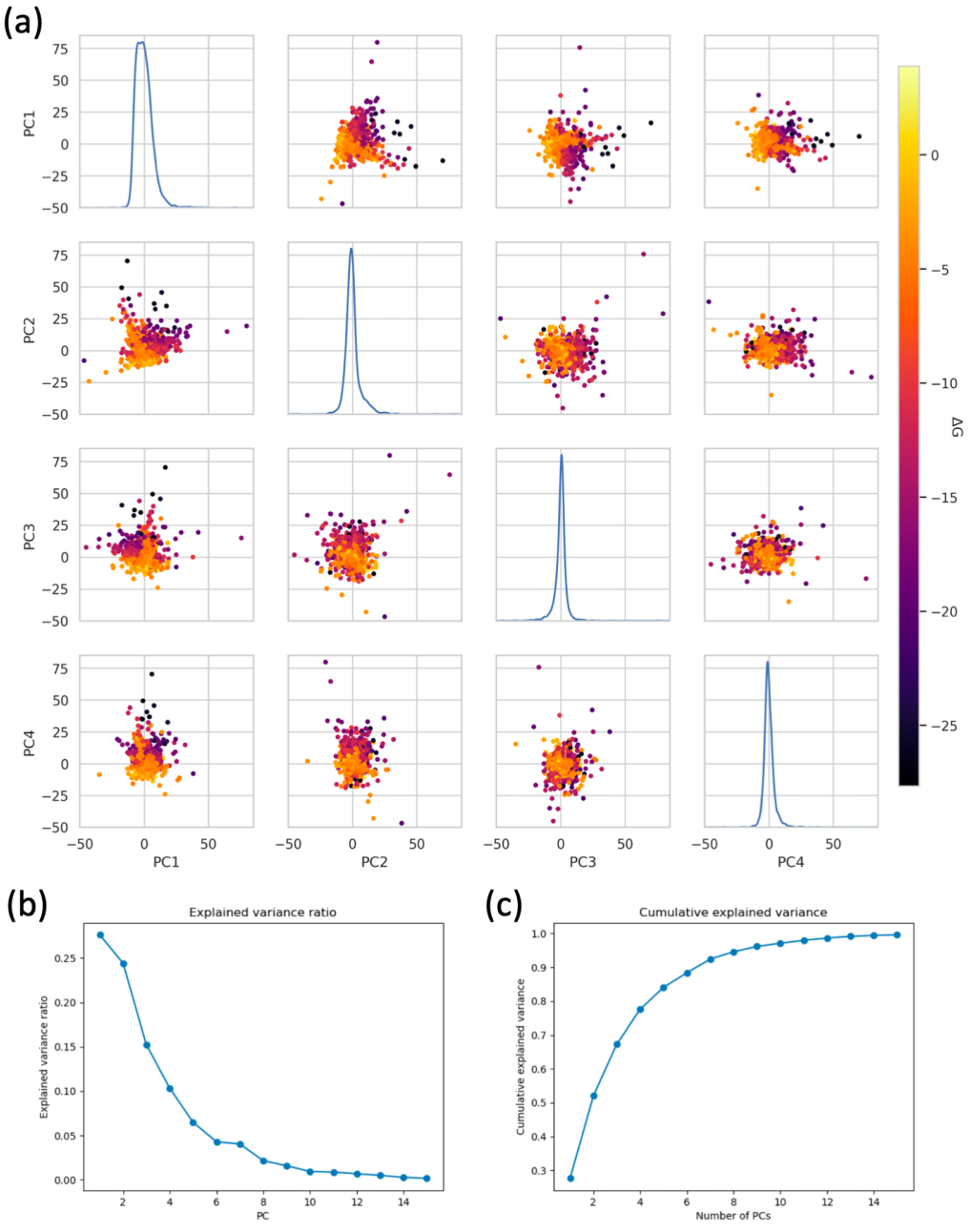

Figure 2. PCA plots showing the chemical space in which the final compiled dataset resides, with distribution plots of the first four PCs shown in the diagonal (**a**). Plots of the fraction of variance described by each principal component (**b**) and the cumulative variance explained by multiple principal components (**c**).

The EVR shows the variance captured by each PC, while the CEV shows the cumulative variance that is captured by all PCs. It is worth noting that because the PCA was carried

out against SFED functions, it captures information about the diversity of solute–solvent systems. The first two PCs capture approximately 28% and 25% variance, respectively, with a decrease in variance captured by each consecutive PC. Approximately 80% of the variance is captured cumulatively by the first four PCs, and the plots of Fig. (a) show each of these PCs plotted against one another. In each plot, it can be seen that the majority of samples reside within a tight chemical space, with a few outliers, which include solute–solvent systems with a wide range of values of ΔG_{solv}, as indicated by the shading.

Figures 3 and 4 show the distributions of experimental SFE values for those solvents and solutes, respectively, that appear in the dataset most commonly.

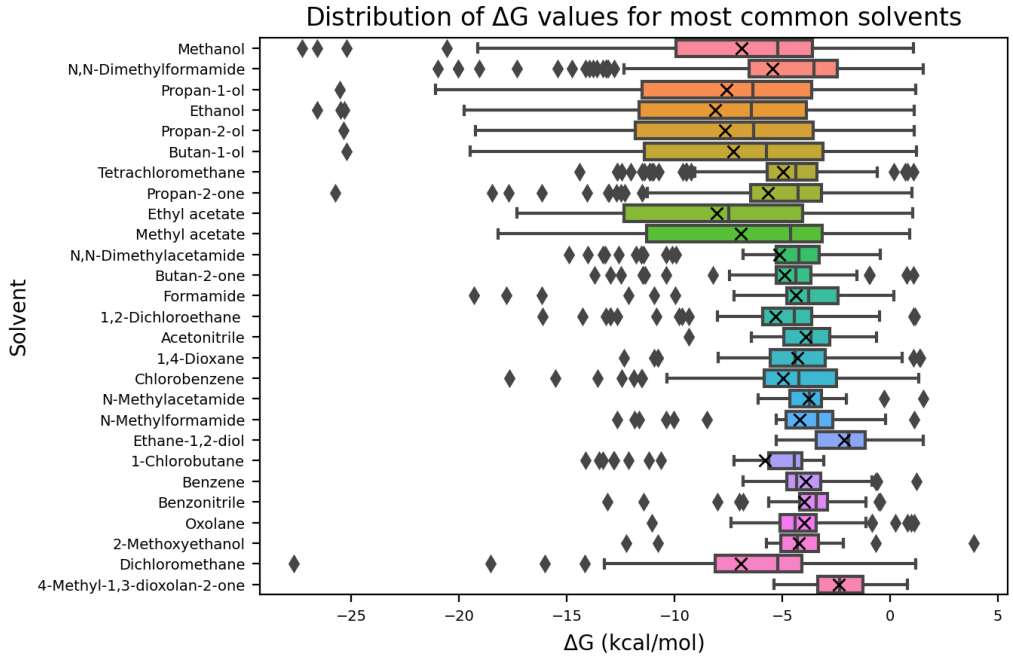

Figure 3. Boxplots showing the SFE distribution for the top 25% of solvents with respect to how frequently they appear in the dataset. The mean SFE value for each solvent is shown by a cross.

The SFE distributions for each solute (in different solvents) are much tighter than those for each solvent (with different solutes). This arises partly due to the sparsity of the dataset in which experimental SFEs were only available for some of the possible solute–solvent pairs. The most common solvent, methanol, appears 150 times, while the most common solute, benzene, appears only 52 times. There are many more solutes than solvents that appear only a small number of times, or even once. While there is a large degree of overlap between the SFE distributions observed between solvents, there is much more variation in the common solutes. Pyrene has measured ΔG values from -13.34 to -11.20 kcal/mol, while molecular nitrogen has measured ΔG values from 0.75 to 1.54 kcal/mol amongst 19 and 23 instances, respectively. The solvent distributions show a much higher number of outliers when compared to the solutes. This may be due to the fact that common solvents will often be paired with solutes for which there are very little data, while conversely, many solutes which appear a small number of times may be paired with a solvent for which a lot of data are present.

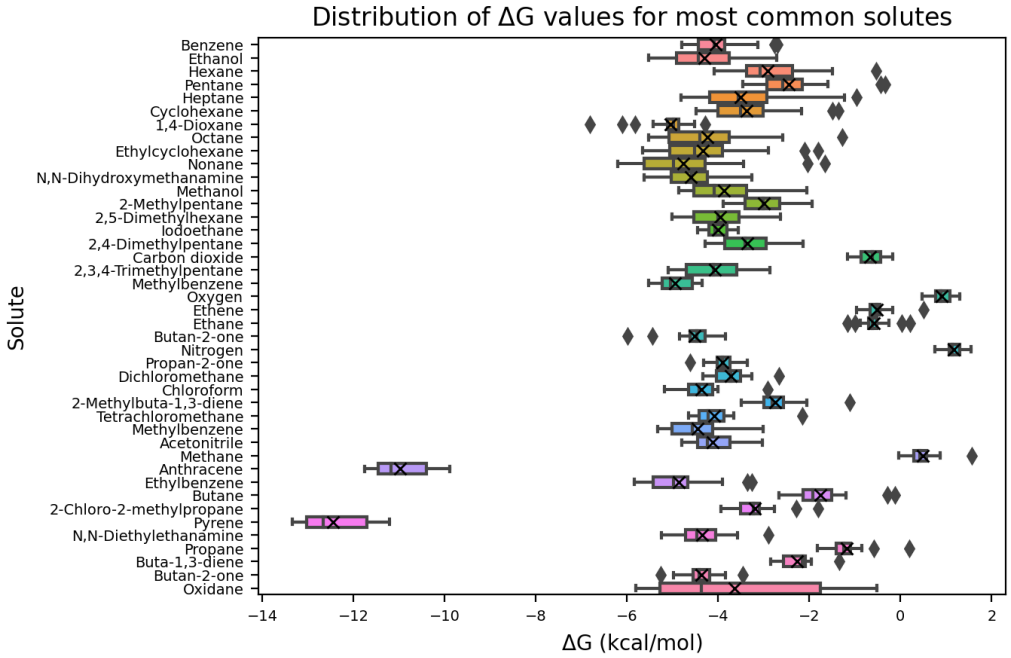

Figure 4. Boxplots showing the SFE distribution for the top 10% of solutes with respect to how frequently they appear in the dataset. The mean SFE value for each solute is shown by a cross.

3.6. CNN Training

A Convolutional Neural Network (CNN) was developed using TensorFlow V2.12.0 for the prediction of SFE based on 1D-RISM calculations. The model was trained and validated by nested cross-validation (CV). An outer cross-validation consisted of 50 resamples, each with a different random 70%/30% train/test split. Hyperparameter optimisation was performed to minimise the mean square error (MSE) estimated from 5-fold CV against the 70% training set, using the GridSearchCV function in SciKit-Learn and the SciKeras package [45]. Only the batch size was optimised. Validation metrics were computed for each of the (30%) test sets and then averaged over the 50 resamples.

The CNN was reoptimised from the initial work [39]. The reoptimised architecture was built using the Keras Sequential API and consists of three convolution blocks, a flatten layer, three densely connected hidden layers, and an output layer. The hidden layers consisted of 64, 32, and 16 nodes, based on the architecture used in previous work [46]. Each convolution block consisted of a Conv1D, MaxPooling1D, and BatchNormalisation layer, in that order. The Conv1D layer had 32 output filters, a kernal size of 3, a stride length of 2, and did not make use of padding, while the MaxPooling1D layers had a max pool size of 2. The ReLU activation function was used in each convolution block and hidden layer. The SciKit-Learn StandardScaler was used to autoscale the feature values such that the mean and standard deviation of the training data were used to scale the training, testing, and validation sets. The Adam optimiser was used throughout the network and had a set learning rate of 0.001. The mean signed error was used for the loss function. Each model was trained for up to 200 epochs, with the Keras callback EarlyStopping used to cease training when the validation loss stopped decreasing with a patience of 20.

4. Results

Model Validation

The CNN model was trained and validated by nested cross-validation in which the inner loop was used to optimise model hyperparameters and the outer loop was used to estimate the model performance. The R^2, RMSE, SDEP, and bias are reported for both the inner and outer loops in Table 1. The metrics from the inner loop are for 5-fold CV against the 70% training set reported as means (with associated standard deviations)

over 50 resamples. The metrics from the outer loop are for prediction of the 30% testing set reported as means (and associated standard deviations) over 50 resamples. In all cases, the performance of the model on the outer loop is consistent with that on the inner loop, with each metric the same to within 1 standard deviation, which shows that the model is performing consistently without significant under- or overfitting. The bias is small compared to the RMSE, showing that the majority of the error is random rather than systematic.

The effect of reoptimising the model hyperparameters against the inner loop is evident in Figure 5, which shows that, as expected, the new model is more accurate and more consistent than the model taken from Ref. [38] without reoptimisation of the hyperparameters. This is not surprising given that the model from Ref. [38] was trained on data from only four solvents, whereas the dataset considered here has over 100 different solvents covering a wider array of solvation chemistries.

Table 1. Table showing mean statistics for the inner and outer loop of the nested cross-validation. The metrics from the inner loop are for 5-fold CV against the 70% training set reported as means (with associated standard deviations) over 50 resamples. The metrics from the outer loop are for prediction of the 30% testing set reported as means (and associated standard deviations) over 50 resamples.

	R^2	RMSE (kcal/mol)	Bias (kcal/mol)	SDEP (kcal/mol)
Inner Loop (5-fold CV)	0.87 (0.02)	1.50 (0.14)	0.04 (0.13)	1.49 (0.14)
Outer Loop (Testing Set)	0.89 (0.02)	1.41 (0.11)	−0.02 (0.14)	1.41 (0.11)

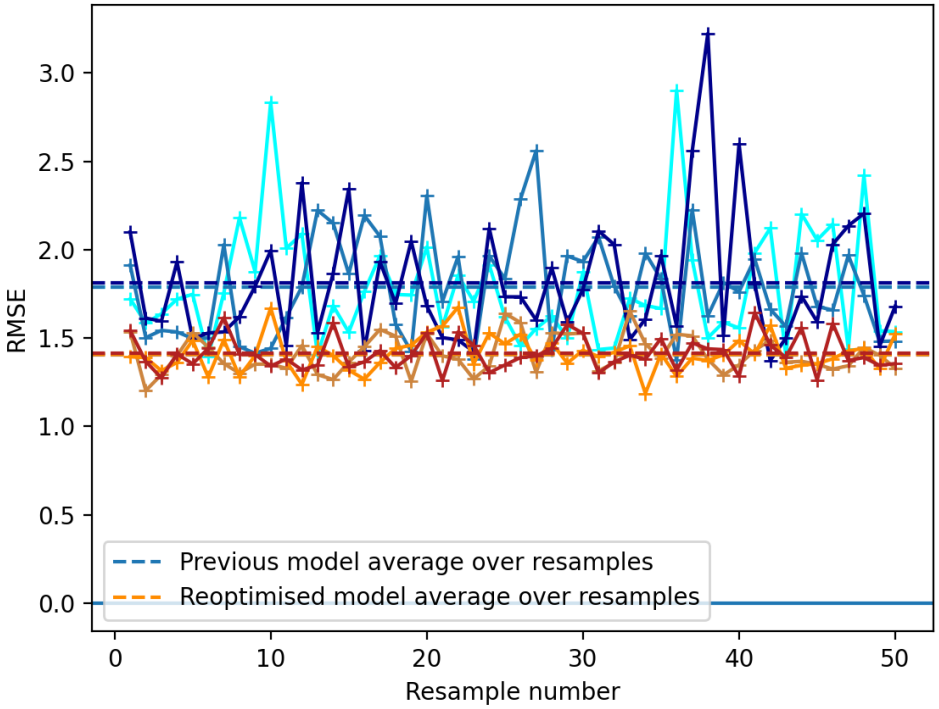

Figure 5. Plots showing the predictive performance of both the previous model (blue) and reoptimised model (orange). Each model was run in triplicate and the RMSE at each resample is plotted. Also plotted is the mean RMSE for each run (dotted lines).

The correlation between the experimental and predicted values for each sample throughout the full set of 50 resamples can be seen in Figure 6, which shows the predictive correlation coloured by (a) solvent and (b) solute, respectively. The overall agreement between the experimental and predicted values is high with few outliers, showing the model has performed well. The line of best fit is very close to the $x = y$ line, and when considered

alongside the R^2 value of 0.89, illustrates that the model is performing satisfactorily. The largest outlier is the solute–solvent system 1,1-difluoroethane-1,4-dimethylbenzene, with an average error of 12 kcal/mol. The least consistently predicted solute–solvent system is niflumic acid-formamide, with a standard deviation of 5.02 kcal/mol.

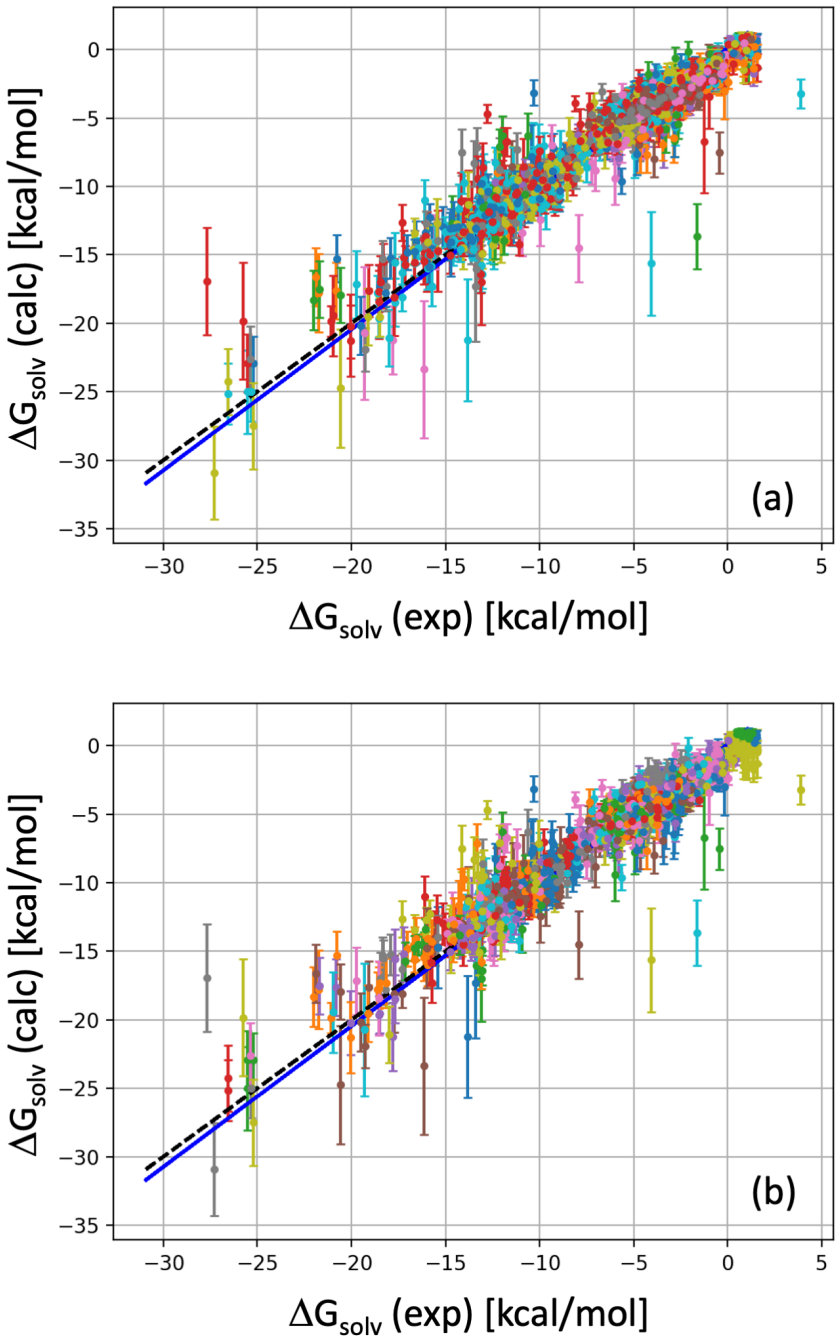

Figure 6. Plots showing the mean prediction vs. the experimentally measured SFE with standard deviations included on each data point, coloured according to (**a**) solvent used and (**b**) solute used. The $x = y$ line is included as a black dotted line, while the trendline is shown as a solid blue line.

Given the sparsity of the experimental data, in which not all solutes have been analysed in all solvents, it is interesting to consider how the availability of training data influences the prediction accuracy. Figure 7 shows that the ML SFE functional is more accurate for solvents that are well represented in the training data. The solvents with 50 or more data points have significantly higher correlations between experimental and predicted

values ($0.76 < R^2 < 1$) than those solvents that are under-represented in the dataset. Indeed, all solvents with 80 or more data points are modelled accurately ($R^2 > 0.9$), as is evident in Table 2. These solvents are generally small and contain some polar heteroatoms (Figure 8), but these characteristics are not directly related to the predictive accuracy since several apolar solutes with fewer than 80 data points are also modelled accurately (e.g., 1,2-demethylbenzene, $R^2 = 0.78$). The performance of the CNN model validates the main hypothesis of this work—that a single ML SFE functional can accurately model different solvent systems—but it also suggests that expanding the dataset size to ensure an even representation of all solvents would improve the model. Unfortunately, the availability of experimental data in the published literature is a limitation. Despite the challenges associated with measuring SFE, there is a clear need for new experimental data.

Table 2. The most common solvents within the dataset with the number of solutes for which each solvent has a measured SFE. The R^2, RMSE, Bias, and SDEP are given for each solvent, calculated across all predictions made on samples containing these solvents.

Solvent	Solutes	R^2	RMSE	Bias	SDEP
Methanol	150	0.96	1.07	−0.19	1.05
N,N-Dimethylformamide	143	0.95	1.06	0.08	1.06
Propan-1-ol	138	0.96	1.00	0.08	1.00
Ethanol	133	0.96	1.10	−0.20	1.08
Propan-2-ol	124	0.96	1.00	0.07	1.00
Butan-1-ol	123	0.97	0.90	−0.15	0.88
Tetrachloromethane	118	0.95	0.70	0.05	0.70
Propan-2-one	96	0.93	1.18	−0.28	1.15
Ethyl acetate	83	0.96	0.97	−0.06	0.97
Methyl acetate	78	0.94	1.27	−0.41	1.20
N,N-Dimethylacetamide	77	0.93	0.88	−0.03	0.88
Butan-2-one	72	0.92	0.79	−0.23	0.76
Formamide	71	0.89	1.21	0.45	1.13
1,2-Dichloroethane	70	0.85	1.29	−0.34	1.24
Acetonitrile	64	0.78	0.69	−0.16	0.67
1,4-Dioxane	61	0.82	1.16	0.36	1.10

Given the overall high performance of the most common solvents, the model was retrained on a filtered variation of the dataset such that only samples corresponding to a solvent which appeared 50 or more times were retained. This reduced the dataset from 2698 samples to 1705. The predictive performance is shown in Table 3.

Table 3. Table showing the mean and standard deviation of 30% testing split prediction statistics calculated over 50 resamples.

	R^2	RMSE (kcal/mol)	Bias (kcal/mol)	SDEP (kcal/mol)
Mean	0.91	1.35	0.01	1.34
Standard Deviation	0.02	0.13	0.18	0.13

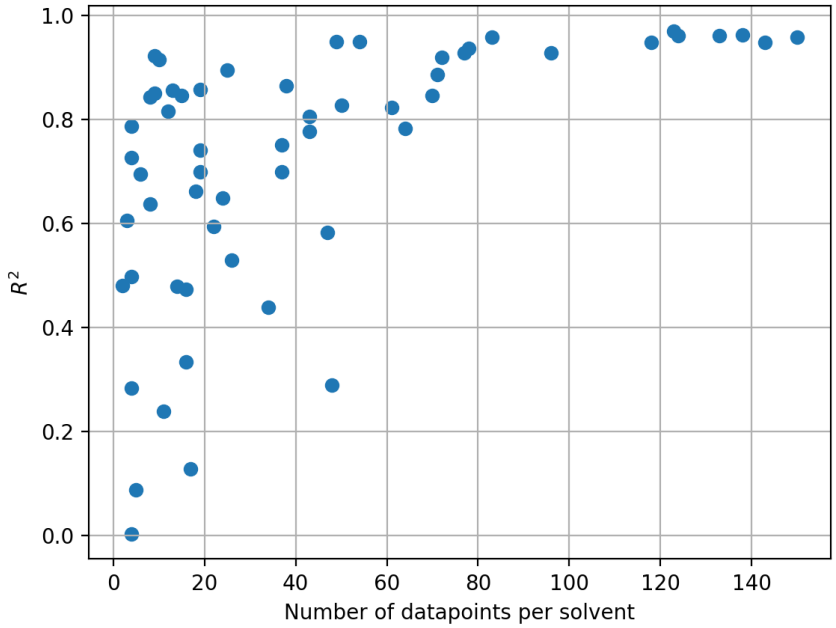

Figure 7. Plot showing the R^2 between the experimental and mean prediction ΔG values vs. the number of data points per solvent.

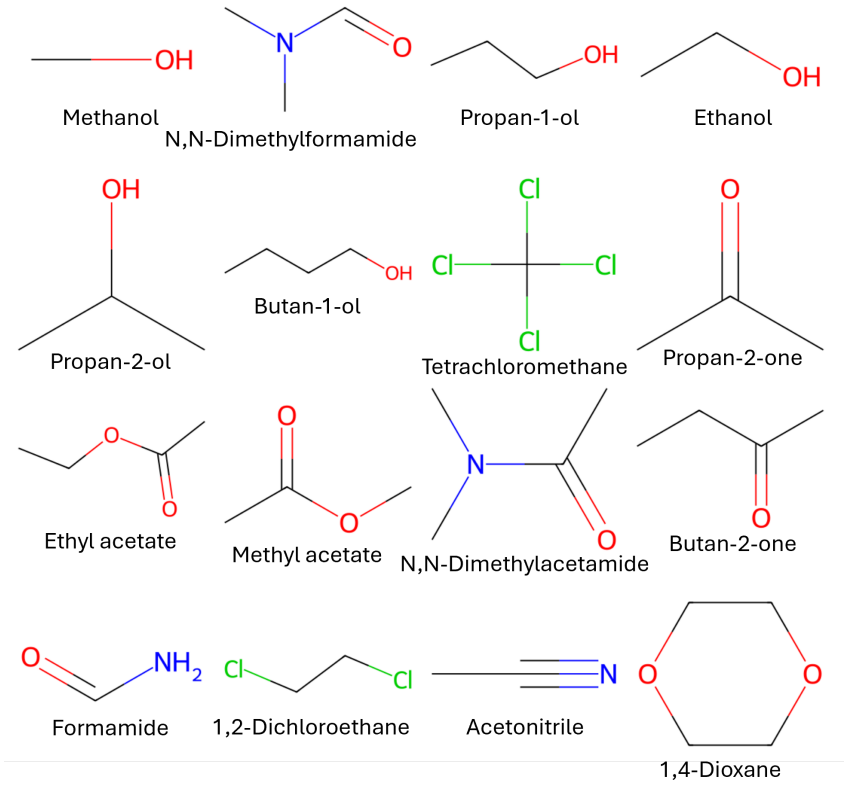

Figure 8. The most common solvents which appear in the dataset used for training.

All the mean values have improved slightly, but only the R^2 has had a significant increase, as the other statistics are within a standard deviation of the earlier results. Interestingly, the standard deviations for RMSE, bias, and SDEP have all slightly increased. Additionally, the sign on the bias has switched, meaning that this model is slightly overpredicting, when it was previously slightly underpredicting. Figure 9 shows the experimental vs. mean predicted SFE values for the common solvents dataset, coloured by (a) solvent and

(b) solute. The same general trends can be drawn from this plot as with the full dataset. The overall agreement between the experimental and mean predictions is satisfactory. There are fewer outliers with the common solvents, which is unsurprising as some scarce data have been removed. The region up to −15 kcal/mol experimental SFE and lower is still less accurately predicted, with larger standard deviations than the region above −15 kcal/mol. It is interesting to note that the highest standard deviation is once again observed on the niflumic acid-formamide data point.

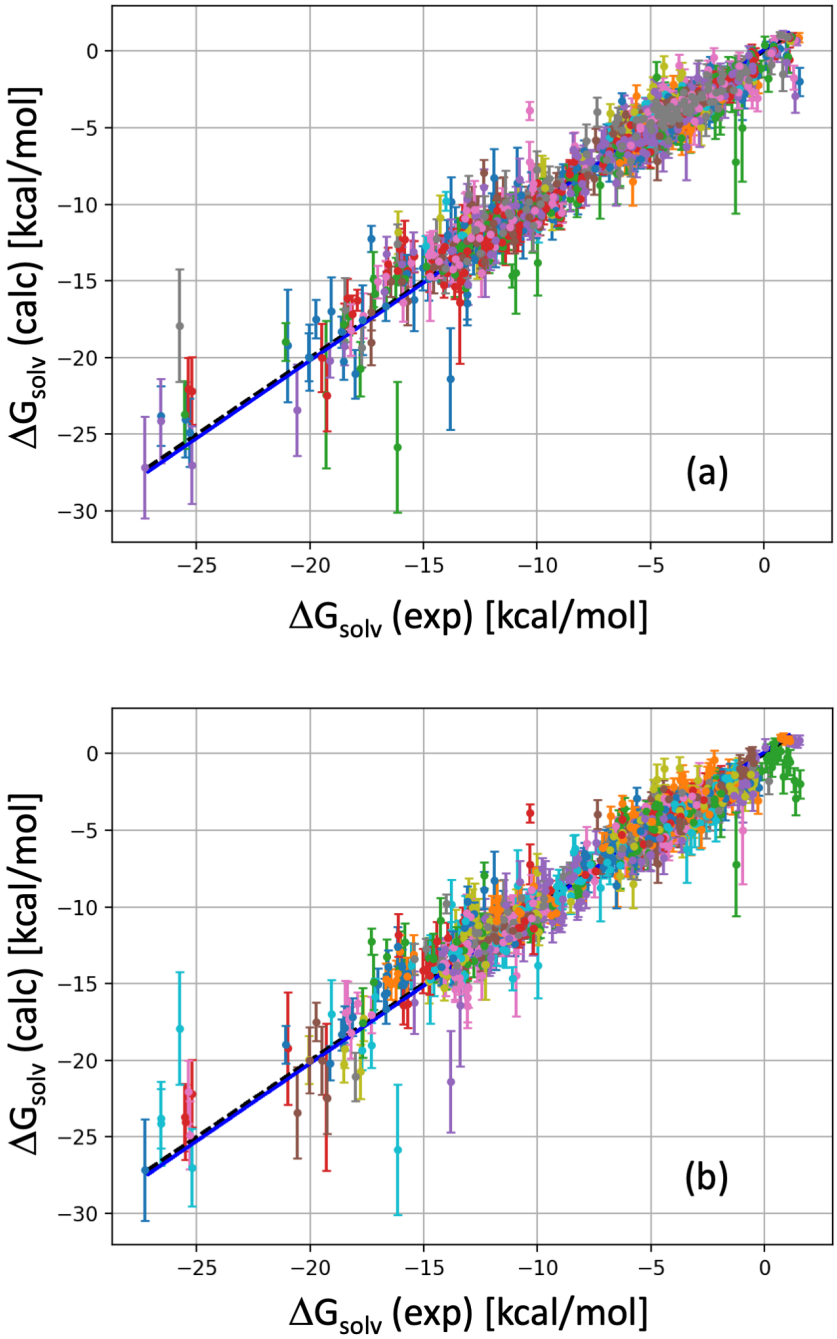

Figure 9. Plots showing the mean prediction vs. the experimentally measured SFE with standard deviations included on each data point, coloured according to (**a**) solvent used and (**b**) solute used for the common solvents. The $x = y$ line is included as a black dotted line, while the trendline is shown as a solid blue line.

Although only common solvents are retained, there has been no direct filtering of the solute data, meaning there are still solutes in the dataset which have a low amount of data. This is a limitation of the dataset and will likely also limit the predictive accuracy of the model. In order to improve the model performance, there would need to be a sufficiently large amount of data for each solute as well as for each solvent. Unfortunately, filtering the dataset by common solvents and solutes reduces the overall sample size to only 14 when setting a threshold of 50 data points on each, and 378 when setting a threshold of 50 for solvents and 20 for solutes. This is not an appropriately large enough volume of data to train a neural network and so more data would be required to verify this.

5. Conclusions

We have developed an ML SFE functional for 1D-RISM that gives accurate predictions for multiple different solvents. A CNN model was trained on SFED functions obtained from 1D-RISM calculations. We achieved accurate predictions, with an RMSE of 1.41 kcal/mol and R^2 of 0.89, using a dataset comprised of 2698 solute/solvent pairs. The CNN was retrained on a reduced version of this dataset, which included only solvents with a large volume of data present in the original dataset. A slight improvement to the model was observed, with an RMSE of 1.35 kcal/mol and R^2 of 0.91. Analysing these results has shown that the predictions are generally more accurate and less varied for those solvents which have larger volumes of data. We have also successfully treated a large number of solvent systems in this way, with a total of 108 solvents being considered in this dataset. The code and datasets used in this work are available at https://github.com/PalmerChem/Conn_Liquids_SI (accessed on 3 November 2024) [47]. The pyRISM software can be accessed at https://github.com/2AUK/pyRISM (accessed on 3 November 2024) [48].

Author Contributions: J.G.M.C. developed the workflow for performing the RISM calculations and developed and trained the CNN used for predictions, as well as curated the datasets used for training. A.A. developed the pyRISM software which was used to perform the RISM calculations. J.G.M.C. and D.S.P. conceptualised the work, methodological approach, and validation processes, as well as analysed the results produced in this work. D.S.P. supervised all work discussed in this manuscript. The manuscript was primarily written by J.G.M.C. with input from co-authors. All authors have read and agreed to the published version of the manuscript.

Funding: J.G.M.C. and D.S.P. would like to thank the EPSRC and IBM for funding via an i-Case Ph.D. studentship.

Data Availability Statement: The datasets used in this work, the CNN python script, and the conda environment yaml file are available at https://github.com/PalmerChem/Conn_Liquids_SI (accessed on 3 November 2024). The pyRISM software can be accessed via the GitHub page of A.A.: https://github.com/2AUK/pyRISM (accessed on 3 November 2024)

Acknowledgments: J.G.M.C., A.A., and D.S.P. would like to thank the ARCHIE-WeSt High-Performance Computing Centre (www.archie-west.ac.uk) for computational resources. J.G.M.C. would like to thank D.J.F. for support in automating RISM calculations.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMBER	Assisted Model Building with Energy Refinement
CEV	Cumulative Explained Variance
CSE	CombiSolv-Exp
CV	Cross-Validation
EVR	Explained Variance Ratio
GAFF	General AMBER Forcefield
GB	Generalised-Born
GF	Gaussian Fluctuations

HFE	Hydration Free Energy
HNC	Hypernetted-Chain
IET	Integral Equation Theory
KH	Kovalenko–Hirata
LJ	Lennard-Jones
MD	Molecular Dynamics
MDIIS	Modified Direct Inversion of the Iterative Subspace
MOZ	Molecular Ornstein–Zernike
OZ	Orstein–Zernike
PB	Poisson–Boltzmann
PC	Pure Pressure Correction
PCM	Polarisable Continuum Model
PC+	Expanded Pure Pressure Correction
RDF	Radial Distribution Function
ReLU	Rectified Linear Unit
RISM	Reference Interaction Site Model
RMSE	Root Mean Squared Error
SASA	Surface Accessible Surface Area
SDEP	Standard Deviation of the Error of Prediction
SFE	Solvation Free Energy
SFED	Solvation Free Energy Density
SMILES	Simplified Molecular-Input Line-Entry System
XRISM	Extended Reference Interaction Site Model

References

1. Ben-Naim, A. *A Molecular Theory of Solutions*; Oxford University Press: New York, NY, USA, 2006.
2. Abel, R.; Wang, L.; Harder, E.D.; Berne, B.J.; Friesner, R.A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res.* **2017**, *50*, 1625–1632. [CrossRef] [PubMed]
3. Ganguly, A.; Tsai, H.; Fernández-Pendás, M.; Lee, T.; Giese, T.J.; York, D.M. AMBER Drug Discovery Boost Tools: Automated Workflow for Production Free-Energy Simulation Setup Analysis (ProFESSA). *J. Chem. Inf. Model.* **2022**, *62*, 6069–6083. [CrossRef] [PubMed]
4. Skyner, R.E.; McDonagh, J.L.; Groom, C.R.; Mourick, T.V.; Mitchell, J.B.O. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174–6191. [CrossRef] [PubMed]
5. Li, W.; Ding, G.; Gao, H.; Zhuang, Y.; Gu, X.; Peijnenburg, W.J.G.M. Prediction of octanol-air partition coefficients for PCBs at different ambient temperatures based on the solvation free energy and the dimer ratio. *Chemosphere* **2020**, *242*, 125246. [CrossRef] [PubMed]
6. Ding, W.; Chen, Y.; Ge, Z.; Cao, W.; Jin, H. A molecular simulation study on solvation free energy and structural properties of polycyclic aromatic hydrocarbons in supercritical water environment. *J. Mol. Liq.* **2020**, *318*, 114274. [CrossRef]
7. Marenich, A.V.; Cramer, C.J.; Truhlar, D.G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B.* **2009**, *113*, 6378–6396. [CrossRef]
8. Marenich, A.V.; Olson, R.M.; Kelly, C.P.; Cramer, C.J.; Truhlar, D.G. Self-Consistent Reaction Field Model for Aqueous and Nonaqueous Solutions Based on Accurate Polarized Partial Charges. *J. Chem. Theory Comput.* **2007**, *3*, 2011–2033. [CrossRef]
9. Cramer, C.J.; Truhlar, D.G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760–768. [CrossRef]
10. Marenich, A.V.; Cramer, C.J.; Truhlar, D.G. Generalized Born Solvation Model SM12. *J. Chem. Theory Comput.* **2013**, *9*, 609–620. [CrossRef]
11. Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117–129. [CrossRef]
12. Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **2003**, *24*, 669–681. [CrossRef] [PubMed]
13. Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: theoterical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032–3041. [CrossRef]
14. Cancès, E.; Mennucci, B. New applications of integral equations methods for solvation continuum models: ionic solutions and liquid crystals. *J. Math. Chem.* **1998**, *23*, 309–326. [CrossRef]
15. Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc.* **1993**, *2*, 799–805. [CrossRef]
16. 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. [CrossRef]

17. Leung, K.; Rempe, S.B.; von Lilienfeld, O.A. Ab Initio molecular dynamics calculations of ion hydration free energies. *J. Chem. Phys.* **2009**, *130*, 204507. [CrossRef]
18. Geballe, M.T.; Skillman, A.G.; Nicholls, A.; Guthrie, J.P.; Taylor, P.J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 259–279. [CrossRef]
19. Geballe, M.T.; Guthrie, J.P. The SAMPL3 blind prediction challenge: transfer energy overview. *J. Comput. Aided Mol. Des.* **2012**, *26*, 489–496. [CrossRef]
20. Mobley, D.L.; Wymer, K.L.; Lim, N.M.; Guthrie, J.P. Blind prediction of solvation free energies from the SAMPL4 challenge. *J. Comput. Aided Mol. Des.* **2014**, *28*, 135–150. [CrossRef]
21. Varilly, P.; Patel, A.J.; Chandler, D. An improved coarse-grained model of solvation and the hydrophobic effect. *J. Chem. Phys.* **2011**, *134*, 074109. [CrossRef]
22. Genheden, S. Solvation free energies and partition coefficients with the coarse-grained and hybrid all-atom/coarse-grained MARTINI models. *J. Comput. Aided Mol. Des.* **2017**, *31*, 867–876. [CrossRef] [PubMed]
23. Shivakumar, D.; Deng, Y.; Roux, B. Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. *J. Chem. Theory Comput.* **2009**, *5*, 919–930. [CrossRef] [PubMed]
24. Vyboishchikov, S.F. Predicting Solvation Free Energies Using Electronegativity-Equalization Atomic Charges and a Dense Neural Network: A Generalized-Born Approach. *J. Chem. Theory Comput.* **2023**, *19*, 8340–8350. [CrossRef] [PubMed]
25. Steinmann, S.; Sautet, P.; Michel, C. Solvation free energies for periodic surfaces: comparison of implicit and explicit solvation models. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31850. [CrossRef] [PubMed]
26. Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *J. Chem. Theory Comput.* **2017**, *13*, 1034–1043. [CrossRef]
27. Chandler, D.; Anderson, H.C. Optimized cluster expansions for classical fluids. 2. Theory of molecular liquids. *J. Chem. Phys.* **1972**, *57*, 1930–1937. [CrossRef]
28. Palmer, D.S.; Frolov, A.I.; Ratkova, E.L.; Federov, M.V. Towards a universal method for calculating hydration free energies: 3D refernce interaction site model with partial molar volume correction. *J. Phys. Condens. Matter* **2010**, *22*, 492101. [CrossRef]
29. 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. [CrossRef]
30. 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. [CrossRef]
31. Ratkova, E.L.; Chuev, G.N.; Sergiievskyi, V.P.; Federov, 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. [CrossRef]
32. Truchon, J.F.; Pettitt, B.M.; Labute, P. A Cavity Corrected 3D-RISM Functional for Accurate Solvation Free Energies. *J. Chem. Theory Comput.* **2014**, *10*, 934–941. [CrossRef] [PubMed]
33. Sergiievskyi, V.P.; Jeanmairet, G.; Levensque, M.; Borgis, D. Solvation free-energy pressure corrections in the three dimensional reference interaction site model. *J. Chem. Phys.* **2015**, *143*, 184116. [CrossRef] [PubMed]
34. Misin, M.; Federov, M.V.; Palmer, D.S. Hydration Free Energies of Molcular Ions from Theory and Simulation. *J. Phys. Chem. B* **2016**, *120*, 975–983. [CrossRef] [PubMed]
35. Misin, M.; Palmer, D.S.; Federov, M.V. Predicting Solvation Free Energies Using Parameter-Free Solvent Models. *J. Phys. Chem B* **2016**, *120*, 5724–5731. [CrossRef]
36. Misin, M.; Vainikka, P.A.; Federov, M.V.; Palmer, D.S. Salting-out effects by pressure-corrected 3D-RISM. *J. Chem. Phys.* **2016**, *145*, 194501. [CrossRef]
37. 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. [CrossRef]
38. Fowles, D.J.; McHardy, R.G.; Ahmad, A.; Palmer, D.S. Accurately predicting solvation free energy in aqueous and organic solvents beyond 298 K by combining deep learning and the 1D reference interaction site model. *Digit. Discov.* **2022**, *2*, 177–188. [CrossRef]
39. Fowles, D.J.; Palmer, D.S. Solvation entropy, enthalpy, and free energy prediction using a multi-task deep learning functional in 1D-RISM. *Phys. Chem. Chem. Phys.* **2023**, *25*, 6944–6954. [CrossRef]
40. Kovalenko, A.; Hirata, F. Hydration free energy of hydrophobic solutes studied by a reference interaction site model with repulsive bridge correction and a thermodynamic perturbation method. *J. Chem. Phys.* **2000**, *113*, 2793–2805. [CrossRef]
41. Vermeire, F.H.; Chung, Y.; Green, W.H. Predictinf Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022**, *144*, 10785–10797. [CrossRef]
42. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and Testing of General AMBER Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174. [CrossRef] [PubMed]
43. The Open Babel Package, Version 3.1.1. Available online: http://openbabel.org (accessed on 1 November 2020)
44. Ahmad, A. *2AUK/pyRISM: pyRISM 0.3.0*; University of Strathclyde: Glasgow, UK, 2023. [CrossRef]
45. Badaracco, A.G. Adriangb/Scikeras: Scikeras 0.11.0. Available online: https://pypi.org/project/scikeras/ (accessed on 3 November 2024).

46. Conn, J.G.M.; Carter, J.W.; Conn, J.J.A.; Subramanian, V.; Baxter, A.; Engkvist, O.; Llinas, A.; Ratkova, E.L.; Pickett, S.D.; McDonagh, J.L.; et al. Blinded Predictions and Post Hoc Analysis of the Second Solubility Challenge Data: Exploring Training Data and Feature Set Selection for Machine and Deep Learning Models. *J. Chem. Inf. Model.* **2023**, *63*, 1099–1113. [CrossRef] [PubMed]
47. Conn, J.G.M. *PalmerChem/Conn_Liquids_SI, 2024*; University of Strathclyde: Glasgow, UK. Available online: https://github.com/PalmerChem/Conn_Liquids_SI (accessed on 3 November 2024)
48. Ahmad, A. *2AUK/pyRISM, 2024*; University of Strathclyde: Glasgow, UK. Available online: https://github.com/2AUK/pyRISM (accessed on 3 November 2024)