PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2023, **25**, 6944

Solvation entropy, enthalpy and free energy prediction using a multi-task deep learning functional in 1D-RISM†

Daniel J. Fowles and David S. Palmer *

Simultaneous calculation of entropies, enthalpies and free energies has been a long-standing challenge in computational chemistry, partly because of the difficulty in obtaining estimates of all three properties from a single consistent simulation methodology. This has been particularly true for methods from the Integral Equation Theory of Molecular Liquids such as the Reference Interaction Site Model which have traditionally given large errors in solvation thermodynamics. Recently, we presented pyRISM-CNN, a combination of the 1 Dimensional Reference Interaction Site Model (1D-RISM) solver, pyRISM, with a deep learning based free energy functional, as a method of predicting solvation free energy (SFE). With this approach, a 40-fold improvement in prediction accuracy was delivered for a multi-solvent, multi-temperature dataset when compared to the standard 1D-RISM theory [Fowles *et al., Digital Discovery*, 2023, **2**, 177–188]. Here, we report three further developments to the pyRISM-CNN methodology. Firstly, solvation free energies have been introduced for organic molecular ions in methanol or water solvent systems at 298 K, with errors below 4 kcal mol^{-1} obtained without the need for corrections or additional descriptors. Secondly, the number of solvents in the training data has been expanded from carbon tetrachloride, water and chloroform to now also include methanol. For neutral solutes, prediction errors nearing or below 1 kcal mol^{-1} are obtained for each organic solvent system at 298 K and water solvent systems at 273–373 K. Lastly, pyRISM-CNN was successfully applied to the simultaneous prediction of solvation enthalpy, entropy and free energy through a multi-task learning approach, with errors of 1.04, 0.98 and 0.47 kcal mol^{-1}, respectively, for water solvent systems at 298 K.

Received 13th January 2023,
Accepted 13th February 2023

DOI: 10.1039/d3cp00199g

rsc.li/pccp

1 Introduction

State-of-the-art computational methodologies are capable of making accurate predictions of solvation free energy for aqueous systems. They are commonly used in the calculation of pK_a,[1–3] protein–ligand binding affinities[4] and aqueous solubility.[5,6] However, the development of computational approaches for determining solvation free energy often focus on aqueous systems, with a lack of progress for organic solvents or non-ambient temperatures. Further, although many methods exist for the prediction of solvation free energy, far fewer exist for the routine prediction of solvation entropy or enthalpy.

Methods of simulating the solvated environment for a given system can generally be separated into one of two categories, implicit or explicit solvent models. The most common implicit models treat bulk solvent as a uniform polarisable medium defined by a dielectric constant, and have found extensive use through models such as the solvation model based on solute electron density (SMD)[7] and the polarisable continuum model (PCM).[8,9] However, implicit models rely on incomplete representations of important molecular level details such as short-ranged solute–solvent interactions. Typically, implicit models only predict for solvation free energy. Several such methods do exist for the routine prediction of other important thermophysical properties however, such as COSMOTherm.[10] Fogolari *et al.* have discussed recent advances in the prediction of solvation thermodynamics,[11] and Karplus *et al.* have proposed a method of estimating the configurational entropy difference between two states,[12,13] both of which involve molecular dynamics (MD) simulations and implicit solvent. Explicit solvent models, such as those commonly used with MD, offer a viable alternative to implicit continuum based approaches,[14] with which a variety of studies report methods for predicting solvation enthalpy or entropy. Lin *et al.* proposed a two-phase thermodynamic model for calculating the entropy of molecular fluids from the trajectory of MD simulations,[15] and the inhomogeneous

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

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3cp00199g

solvation theory (IST) has seen increased use for predicting solvation thermodynamics.[16] However, the use of explicit solvent models and molecular dynamics simulations come at a far greater cost than their implicit counterparts, and often require time consuming and expensive simulations to model even a modest number of systems.

The reference interaction site model (RISM) is a third approach, capable of calculating solvation dependent thermodynamic parameters at a lower computational cost than explicit models, whilst modelling specific solute–solvent interactions. The RISM theory uses a simplified form of the high-dimensional molecular Ornstein–Zernike (MOZ) equations to model solvent density distribution around a solute molecule through a set of correlation functions, from which two distinct methods have been developed. The most commonly used of these is 3D-RISM, which approximates the MOZ equations by a set of three-dimensional integral equations. With the recent development of several semi-empirical[17,18] and theoretical free energy functionals,[19,20] 3D-RISM has found frequent use as a method to predict SFE.[21–25] Solvation enthalpies and entropies can also be obtained through 3D-RISM with the decomposition of solvation free energy into the entropic contribution using temperature derivatives.[26] With this method, solvation enthalpies and entropies were reported to within 2.12 and 1.93 kcal mol^{-1} of experiment, respectively. However, as solvation entropies were extrapolated from calculated free energies, for which state-of-the-art semi-empirical or theoretical free energy functionals are necessary to obtain reasonable agreement to experiment, any errors found within free energy calculations can also be found in the associated solvation entropy and enthalpy. By contrast, the 1D-RISM theory, in which the MOZ equations are approximated as a set of one-dimensional integral equations, is rarely used for quantitative calculations of solvation thermodynamics because it is considered to be too inaccurate in its common form.

Within the RISM framework, solvation free energy predictions are made analytically using one of several available free energy functionals. In 1D-RISM many of these functionals fail to accurately predict the energetic parameters of the chemical system under investigation. These functionals, such as the Hyper-Netted Chain (1D-RISM/HNC) model,[27] are too inaccurate for routine use and typically achieve absolute prediction errors above 20 kcal mol^{-1}. Much effort has been put into improving the predictive capabilities of 1D-RISM based functionals for SFE calculations. Some of these improved models, such as the Gaussian Fluctuations (1D-RISM/GF) and Partial Wave models (1D-RISM/PW), can more accurately predict SFE than previous methods.[28,29] Although reasonable qualitative agreement with experimental data has been reported, large predictive errors are still commonly observed for many chemical systems.

In previous work, we proposed a method of accurately predicting solvation free energy.[30] This method, pyRISM-CNN, combined our in-house 1D-RISM solver, pyRISM,[31] with a deep learning based free energy functional and was shown to accurately predict SFE for organic molecules in aqueous solvent at 273–373 K, as well as carbon tetrachloride or chloroform solvent systems at 298 K. Compared to the standard 1D-RISM theory, the pyRISM-CNN functional reduced the predictive error by up to 40-fold, obtaining a prediction accuracy below 1 kcal mol^{-1} of experiment across each tested solvent. Moving from 1D-RISM calculation to pyRISM-CNN prediction requires minimal additional computational expense as the solvation free energy density (SFED) functions that are used as input to the CNN model can be generated as part of the typical 1D-RISM workflow.

Here, we report three further developments to the pyRISM-CNN methodology. Solvation free energy data at 298 K has been introduced for methanol solvent systems, for a total of four solvents alongside carbon tetrachloride, chloroform and water. Organic molecular ions have also been introduced for water and methanol solvent systems, allowing pyRISM-CNN to predict SFE for both neutral and ionised solutes, either as a combined input or independently. Accurate predictions of SFE can be obtained for both neutral molecules and molecular ions without the need for additional descriptors or corrections. The pyRISM-CNN functional has also been successfully applied to the simultaneous and accurate prediction of solvation enthalpy, entropy and free energy through a multi-task learning approach. By expanding the range of chemical systems for which SFE predictions can be made, as well as enabling the accurate prediction of solvation enthalpy, entropy and free energy, pyRISM-CNN has expanded the potential application for the RISM theory.

2 Theory

2.1 1D-RISM

The details of the general RISM theory have been discussed in depth elsewhere,[32] and so only the 1D-RISM theory will be explained here. 1D-RISM uses an approximated one-dimensional form of the molecular Ornstein–Zernike equation with spherically symmetric site–site correlation functions for the modelling of molecular solutions. Both solute and solvent molecules are modelled as sets of spherically symmetric sites, with one site per atom. There are three types of site–site correlation functions that are considered in RISM, each of which varies with site–site separation only: intramolecular correlation functions, total correlation functions and direct correlation functions. The intramolecular correlation functions describe the structure of a given molecule. For two sites within a molecule, s and s', the intramolecular correlation function is written as

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

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

Intermolecular solute–solvent correlations are defined for each pair of solute and solvent sites by the total correlation functions $h_{s\alpha}(r)$ and direct correlation functions $c_{s\alpha}(r)$. Here, s refers to a solute site and α to a solvent site. The total correlation functions are closely related to the radial distribution function (RDF) as

$$h_{s\alpha}(r) = g_{s\alpha}(r) - 1 \quad (2)$$

where $g_{s\alpha}(r)$ is the radial distribution function of solvent sites around a given solute site.

The total and direct correlation functions are related *via* a set of RISM equations

$$h_{s\alpha}(r) = \sum_{s'=1}^{M}\sum_{\xi=1}^{N}\int_{R^3}\int_{R^3}\omega_{ss'}(|r_1 - r'|) \times c_{s'\xi}(|r' - r''|)\chi_{\xi\alpha}(|r'' - r_2|)\mathrm{d}r'\,\mathrm{d}r'' \quad (3)$$

where $r = |r_1 - r_2|$, $\chi_{\xi\alpha}(r)$ are the bulk solvent susceptibility functions, and M and N are the number of solute and solvent sites, respectively. Any mutual correlations between bulk solvent sites are described by the solvent susceptibility functions $\chi^{\mathrm{solv}}_{\xi\alpha}(r)$, which are determined from solvent–solvent site total correlation functions $h^{\mathrm{solv}}_{\xi\alpha}(r)$, intramolecular correlation function $\omega^{\mathrm{solv}}_{\xi\alpha}(r)$ and the solvent bulk number density ρ.

$$\chi_{\xi\alpha}(r) = \omega^{\mathrm{solv}}_{\xi\alpha}(r) + \rho h^{\mathrm{solv}}_{\xi\alpha}(r) \quad (4)$$

The solvent–solvent site $h^{\mathrm{solv}}_{\xi\alpha}(r)$ and $\omega^{\mathrm{solv}}_{\xi\alpha}(r)$ are obtained from preliminary solvent–solvent 1D-RISM calculations and molecular structure. To complete the set of RISM equations, closure relations must be introduced

$$h_{s\alpha}(r) = \exp(-\beta u_{s\alpha}(r) + \gamma_{s\alpha}(r) + B_{s\alpha}(r)) - 1 \quad (5)$$

where $u_{s\alpha}(r)$ is the atom–atom potential, $B_{s\alpha}(r)$ is a bridge function, $\beta = 1/k_BT$ and $\gamma_{s\alpha}$ is the indirect correlation function ($\gamma_{s\alpha}(r) = h_{s\alpha}(r) - c_{s\alpha}(r)$).

The exact bridge functions are typically unknown and so an approximation is needed to solve for the total correlation functions and direct correlation functions. A commonly used closure is the Kovalenko and Hirata (KH) closure[33]

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

where $\Xi_{s\alpha}(r) = -\beta u_{s\alpha}(r) + \gamma_{s\alpha}(r)$. A threshold constant, C, with a value of 0 was introduced with the KH closure to linearize the exponent when its argument grew larger than C. If the value of C is changed from zero to infinity, the KH closure becomes the HNC closure.

There are multiple expressions available within RISM for determining solvation free energy once the total and direct correlation functions have been solved. The functional is usually selected to be consistent with the closure used within the 1D-RISM calculations. The Gaussian fluctuations approximation (GF),[34] KH[35] and hypernetted chain (HNC)[27] expressions are shown below.

$$\Delta G_{\mathrm{GF}} = 2\pi\rho kT\sum_{s\alpha}\int_0^\infty(-2c_{s\alpha}(r) - h_{s\alpha}(r)c_{s\alpha}(r))r^2\mathrm{d}r \quad (7)$$

$$\Delta G_{\mathrm{KH}} = 2\pi\rho kT\sum_{s\alpha}\int_0^\infty[-2c_{s\alpha}(r) - h_{s\alpha}(r)(c_{s\alpha}(r) - \Theta(-h_{s\alpha}(r)))]r^2\,\mathrm{d}r \quad (8)$$

$$\Delta G_{\mathrm{HNC}} = 2\pi\rho kT\sum_{s\alpha}\int_0^\infty[-2c_{s\alpha}(r) - h_{s\alpha}(r)(c_{s\alpha}(r) - h_{s\alpha}(r))]r^2\,\mathrm{d}r \quad (9)$$

2.2 pyRISM

The pyRISM program[31] includes a general method of obtaining variables which contain solvation and desolvation relevant descriptors from the standard 1D-RISM free energy functionals. Previously this method was applied within the RISM-MOL framework and has been described in detail elsewhere,[36,37] so only a short summary of the process will be described here. Each of the free energy functionals described in eqn (7)–(9) can be condensed into a generalised form:

$$\Delta G_{\mathrm{RISM}} = \int_0^\infty w(r)\,\mathrm{d}r \quad (10)$$

where the integrand functional $w(r)$ combines the prefactor ($2\pi\rho kT$), and the total and direct correlation functions of a single solute into an individual function of r which is referred to as the solvation free energy density (SFED). By then omitting the integration over r, this functional can be used to obtain variables that quantify the response of solvent molecules to the solute at chosen distances r from the solute site. The SFED functions derived from the GF, KH and HNC SFE functionals are given below:

$$gf\_w(r) = 2\pi\rho kT\sum_{s\alpha}(-2c_{s\alpha}(r) - h_{s\alpha}(r)c_{s\alpha}(r)) \quad (11)$$

$$kh\_w(r) = 2\pi\rho kT\sum_{s\alpha}[-2c_{s\alpha}(r) - h_{s\alpha}(r)(c_{s\alpha}(r) - \Theta(-h_{s\alpha}(r)))] \quad (12)$$

$$hnc\_w(r) = 2\pi\rho kT\sum_{s\alpha}[-2c_{s\alpha}(r) - h_{s\alpha}(r)(c_{s\alpha}(r) - h_{s\alpha}(r))] \quad (13)$$

When the 1D-RISM equations are solved, the total and direct correlation functions are represented on a fine grid. The values of the SFED functions at selected grid points provide variables that are denoted as *m_w_n*, where *m* is the 1D-RISM free energy functional from which the variable is based and *n* is the grid point at which the variable is evaluated. Machine learning algorithms are then trained on these variables and the subsequent model can be used for solvation free energy prediction. pyRISM is made freely available as open-source software.

3 Methods

3.1 Dataset preparation

Experimental solvation free energies of small organic molecules were obtained from three different sources: the Minnesota Solvation Database (MSD),[38] those developed by Chamberlin *et al.*[39,40] and Zanith *et al.*[41] The MSD contains experimental solvation free energies in water and several organic solvents at 298 K. Experimental data obtained from Chamberlin *et al.* includes hydration free energies in a 273–373 K temperature

range, and data obtained from Zanith *et al.* includes solvation free energies for small organic molecules in methanol at 298 K.

A multi-solvent, multi-temperature dataset of neutral and ionised compounds was prepared from the available experimental solvation free energies. Ionised compounds were obtained exclusively from the MSD and consisted of ionised organic solutes. A total of four solvents made up the neutral portion of the dataset: water, methanol, chloroform and carbon tetrachloride, with 659, 25, 109 and 79 solutes respectively. Of the 659 water solutes, 272 were taken from Chamberlin *et al.*, and included hydration free energy data in a 273–373 K temperature range. By using free energies over a range of temperatures, a total of 3053 datapoints were available. The remaining 387 solutes were taken from the MSD at 298 K, for a total of 3440 datapoints. The ionised portion of the dataset consisted of anions and cations in water and methanol solvents. By solvent, 48 anions and 29 cations were present in methanol, and 56 anions and 47 cations in water.

Experimental solvation enthalpies, entropies and free energies of small organic molecules were taken from several different sources. This second dataset contained a full set of experimental solvation thermodynamic parameters for solute molecules in water at 298 K. Every molecule present in this second dataset can also be found in the solvation free energy dataset. Solvation enthalpies for solutes in water were obtained from the Acree dataset[42] and Abraham *et al.*[43] Solvation entropies were taken from Garza.[44] Experimental solvation free energies from the MSD were assigned to each molecule. If an experimental solvation enthalpy, entropy and free energy were not all available for a given molecule then a pseudo-experimental value would be calculated from the other two available terms using $\Delta G = \Delta H - T\Delta S$. In total, solvation data was obtained for 139 solutes in water. Where necessary, experimental values were converted to the Ben Naim standard state.[45,46] Table 1 provides a breakdown of the available experimental data for each solvent. A spreadsheet detailing the available experimental data by source can be found as part of the ESI.†

Fig. 1 provides violin plots which show the distribution of experimental solvation free energy and molecular weight by solvent across the neutral and ionised datasets. Additional violin plots of $\log P$ and the number of rotatable bonds per solute are available in Section S1 of the ESI.† Tables containing the mean and standard deviation (SD) of each experimental property across the neutral and ionised datasets, as well as the mean and SD of experimental solvation enthalpy, entropy and free energy values, can also be found in Section S1 of the ESI.†

3.2 Solute structure generation

The InChi[47] function within Open Babel[48] was used to generate a unique InChi descriptor for each solute within the MSD, Chamberlin *et al.* and Zanith *et al.* datasets. Any duplicate molecules were then removed using the ''unique'' function within Open Babel. Dataset sizes shown in Table 1 refer to solutes present after duplicate molecules were removed.

Solute coordinate files were taken from the MSD. The corresponding coordinate files were not available for 10 solute molecules taken from Zanith *et al.*, and so were obtained from the PubChem chemical database[49] as a 2D coordinate file. These files were converted to 3D structures and the lowest energy conformer found through a conformational search carried out using Open Babel. The conformational search involved a systematic rotor search of each solute with the GAFF forcefield.[50]

3.3 1D-RISM calculations

1D-RISM calculations were carried out with pyRISM using the KH closure within a system of 16 384 grid points over 20.48 Å from the solute. Aqueous solvent calculations used the dielectrically consistent reference interaction site model (DRISM),[51,52] while organic solvent calculations applied the extended reference interaction site model (XRISM).[53] Organic solvent calculations were found to converge more consistently with XRISM than with DRISM. Solvation free energy calculations with the KH, HNC and GF free energy functionals were performed for both aqueous and organic solvent systems. For all calculations it was assumed that solute molecules were embedded within an infinitely dilute aqueous solution. A convergence tolerance of 10^{-12} was set for all calculations, with a minimum tolerance of 10^{-5} if the initial calculation failed to converge. The impact from lowering the minimum convergence tolerance to 10^{-5}, as well as the choice of model for 1D-RISM calculations (DRISM or XRISM) on the quality of SFED generated has previously been found to be negligible.[30]

Table 1 Breakdown of descriptors generated from each solvent. Two datasets were compiled from the available experimental data. The first dataset was created from neutral and ionised experimental SFE at 273–373 K, and the second from experimental solvation enthalpies, entropies and free energies at 298 K. The number of datapoints listed for each solvent represents that solvents contribution to the total number of SFED in each dataset

Dataset	Solvent	Temperature range	SFE functional	Datapoints
$\Delta G_{solv}^{exp,neutral}$	Carbon tetrachloride	298 K	KH/HNC/GF	79
	Chloroform	298 K		109
	Methanol	298 K		25
	Water	273–373 K		3440
$\Delta G_{solv}^{exp,ionised}$	Methanol	298 K	KH/HNC/GF	77
	Water	298 K		103
ΔH_{solv}^{exp}, $T\Delta S_{solv}^{exp}$, $\Delta G_{solv}^{exp,neutral}$	Water	298 K	KH/HNC/GF	139

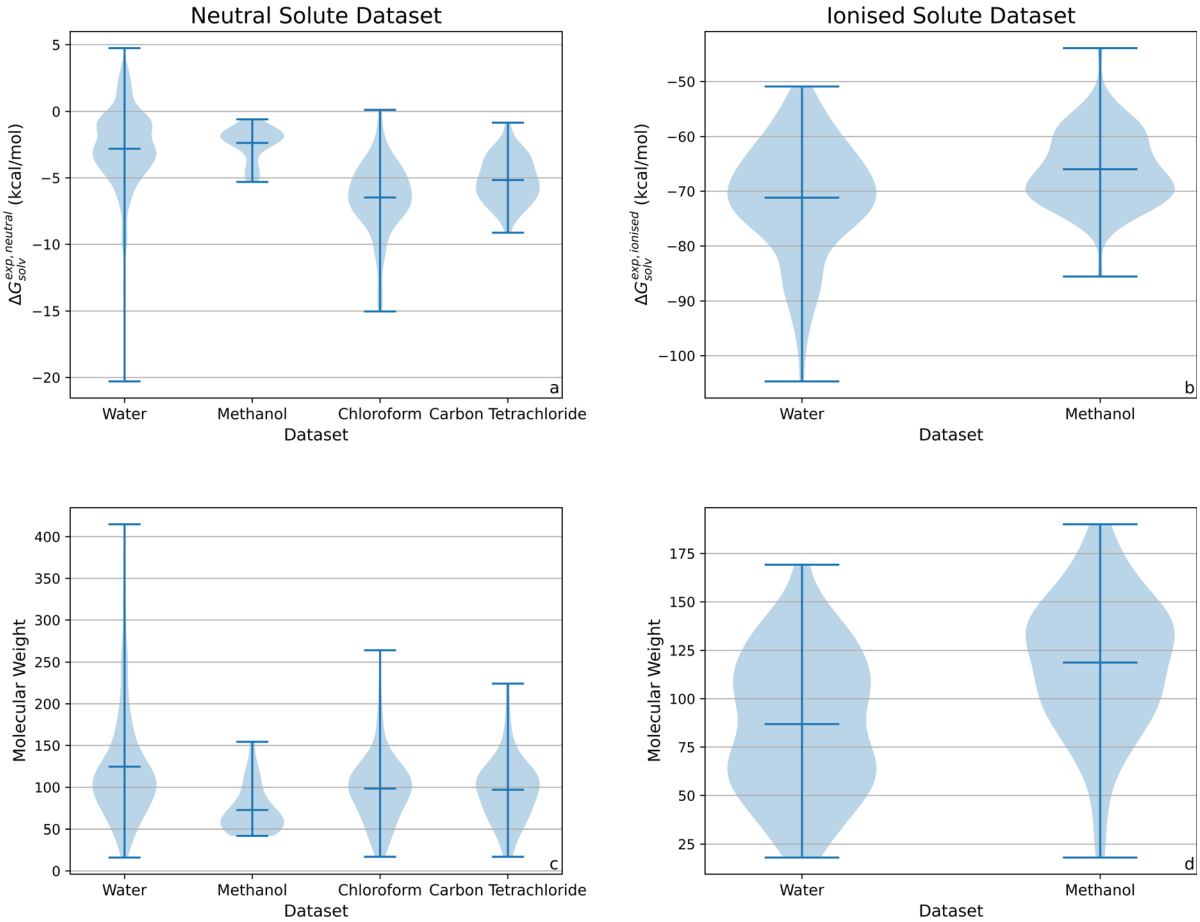

Fig. 1 Violin plots showing solute molecule data by solvent for the neutral and ionised datasets. a = experimental solvation free energies for the neutral solute dataset, b = experimental solvation free energies for the ionised solute dataset, c = molecular weights for the neutral solute dataset, d = molecular weights for the ionised solute dataset.

3.3.1 Solvent & solute parameters. The Lue and Blankschtein version of the SPC/E water model (MSPC/E)[54] was used for modelling aqueous solvent. This altered version differs from the original model with the inclusion of modified Lennard-Jones (LJ) potential energy parameters for water based hydrogen, which were adjusted to prevent any possible divergence of the algorithm.[53,55,56] Organic solvent models were modelled using the general Amber forcefield (GAFF) non-bonded parameters, which were assigned using the Antechamber and tLEaP programs within Amber18.[57] The Lorentz–Berthelot mixing rules[58] were used to generate solute–solvent LJ parameters *i.e.*, $\sigma_{s\alpha} = (\sigma_s + \sigma_\alpha)/2$ and $\varepsilon_{s\alpha} = \sqrt{\varepsilon_s \varepsilon_\alpha}$. GAFF[50] parameters were assigned to solute molecules using the Antechamber and tLEaP programs within Amber18.

3.4 Obtaining RISM solvation free energy density functions

Solute specific SFED functions were obtained as a 1D-RISM calculation output using the pyRISM program. A SFED function was generated for each free energy functional used, totalling three sets per solute. As the grid used to represent the 1D-RISM total and direct correlation functions was very fine, leading to multiple correlated variables, a coarser grid-spacing was used to obtain a representation of the SFED that was suitable for machine learning. To minimise the inclusion of redundant data and to exclude data at long solute–solvent separations in the region where SFEDs approach zero, only every 40th grid point from r = 0 Å to r = 8 Å was considered. This approach produced 160 SFED descriptors per SFE functional for each solute 1D-RISM calculation. These SFED were then used as an input to machine learning models to predict experimental solvation free energy, enthalpy and entropy.

3.5 Convolutional neural network models

Convolutional neural network (CNN) models were trained on all three SFED variations (KH, HNC and GF) for each dataset. Models were validated by nested cross-validation (CV), with hyper parameters tuned by an inner 5-fold CV loop. Final tuned model performance was evaluated by an outer 50-fold Monte Carlo cross-validation loop with a 70% train/30% test split. A stratified sampling approach was taken for multi-temperature and multi-solvent data to ensure datapoints were separated by molecule before splitting into training, validation and test sets. Each variable was centered by subtracting its mean value in the

training data, and scaled by dividing by the standard deviation of its values in the training data.

Single and multi-output convolutional neural networks were built using the 'sequential' and 'functional' model packages in Tensorflow[59] respectively, and accessed using Keras[60] with a Python implementation. Single output CNN were trained on SFED generated from the solvation free energy dataset, and multi-output CNN were trained on SFED generated from the solvation enthalpy, entropy and free energy dataset. The multi-task algorithm considered solvation enthalpy, entropy and free energy with an equal weighting. The CNN training datasets contained enthalpy, entropy and free energy values for each solute without any missing data. Final CNN architecture consisted of three blocks of Conv1D-MaxPooling1D-BatchNormalisation with a subsequent Flatten layer, and was based on the refined CNN architecture applied in our previous work.[30] Single output models contained a single Dense output layer while multi-output models had three separate Dense output layers connected to the Flatten layer. Convolutional layers were created using the 'Conv1D' layer package in Keras with 32 output filters, a kernal size of 3 and stride length of 2. No padding was included and the rectified linear activation function (ReLu)[61] was used. Each of the subsequent layers were also taken from Keras, with the max pool size within MaxPooling1D layers set to 2. Default parameters were used for BatchNormalisation and Flatten layers. The loss function and metric was set to 'mse' (mean squared error), with the 'Adam' optimiser.[62] Each model could run for a maximum of 60 epochs with a patience of 20 epochs included through the Keras 'EarlyStopping' callback.

3.6 Statistical analysis

Solvation thermodynamics predictions were evaluated against experimental values of solvation enthalpy, entropy or free energy using the coefficient of determination (R^2) and root mean squared deviation (RMSD).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y^i - y^i_{\text{exp}}\right)^2}{\sum_{i=1}^{N}\left(y^i - M\left(y^i_{\text{exp}}\right)\right)^2} \tag{14}$$

$$\text{RMSD}\left(y, y_{\text{exp}}\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y^i - y^i_{\text{exp}}\right)^2} \tag{15}$$

where index i goes through a set of N molecules, and y^i and y^i_{exp} are the predicted and experimental values for molecule i respectively. The coefficient of determination represents a statistical measure of how well the regression predictions fit the experimental data, and so negative values below 1 are possible for models which fit the data worse than the mean of the experimental data. The total deviation can be separated into two parts: bias (or mean displacement, M) and standard deviation (or SDEP, σ).

$$\text{bias} = M\left(y - y_{\text{exp}}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - y^i_{\text{exp}}\right) \tag{16}$$

$$\sigma\left(y - y_{\text{exp}}\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y^i - y^i_{\text{exp}} - M\left(y - y_{\text{exp}}\right)\right)^2} \tag{17}$$

The bias provides the systematic error, while the standard deviation gives the random error that is not explained by the model. The bias and standard deviation are connected to the RMSD by:

$$\text{RMSD}(y,y_{\text{exp}})^2 = M(y - y_{\text{exp}})^2 + \sigma(y - y_{\text{exp}})^2 \tag{18}$$

A model which reports an RMSD greater than the standard deviation of the experimental data provides less accurate predictions than the null model provided by the mean of the experimental data.

Statistical analyses were performed in a Python environment using the 'sklearn.metrics' module available in scikit-learn.[63]

4 Results and discussion

4.1 Solvation free energy – neutral & ionised datasets

Convolutional neural network models were trained on datasets of neutral or ionised solutes, separately. For both the neutral and ionised datasets, three variations were tested: KH, HNC and GF generated SFED, for a total of six models. Only models trained on GF based SFED will be discussed here, as it has been previously shown that the choice of free energy functional only marginally effects SFE prediction accuracy.[30] Models associated with KH and HNC generated SFED can be found in Section S2 of the ESI.† Table 2 provides a breakdown of the test set based performance for both models trained on GF generated SFED.

In our previous study of pyRISM-CNN, a CNN model trained on the GF generated multi-solvent, multi-temperature SFED dataset achieved an RMSD of 0.97 kcal mol^{-1} and R^2 of 0.94. This dataset consisted of the carbon tetrachloride, chloroform and water based solute data presented in the neutral SFE section of Table 2 under 'Solvent', 'Temperature' and 'Datapoints'. Here, we have expanded that benchmark dataset to include experimental solvation free energies for 25 solute molecules in methanol. As can be seen in Table 2, including this additional solvent does not impact upon the overall model performance, with an RMSD of 0.99 kcal mol^{-1} and R^2 of 0.93. Individually, SFE predictions made for methanol based solutes are no less accurate than for other solvents, with an RMSD of 0.73 kcal mol^{-1} and R^2 of 0.42. The relatively low R^2 is likely due to the small number of datapoints available for methanol, which represent a smaller spread of experimental SFE values than are available for the other solvents. The performance of this CNN model, trained on the expanded neutral solute dataset, further shows the generalisability and capabilities of this approach.

Similarly to the neutral solute dataset, a CNN trained on SFED generated from ionised organic solutes is capable of making accurate solvation free energy predictions. From the ionised SFE section of Table 2, the CNN trained on ionised solute data can be seen to accurately predict SFE to within 3.96 kcal mol^{-1} of experiment. This accuracy is not limited to a

Table 2 Breakdown of solvation free energy predictions for neutral and ionised solutes. Two separate CNN were trained on either neutral or ionised solute SFED, which were generated using the GF free energy functional. Statistics are given for both datasets, with model performance separated by solvent as well as across the full datasets of neutral or ionised solutes. Errors are given in kcal mol^{-1}

$\Delta G_{solv}^{exp,neutral}$ dataset										
Neutral solvation free energy			By solvent				Full dataset			
Solvent	Temperature	Datapoints	R^2	RMSD	Bias	SDEP	R^2	RMSD	Bias	SDEP
Carbon Tetrachloride	298 K	79	0.63	1.00	0.69	0.68	0.93	0.99	0.28	0.93
Chloroform	298 K	109	0.81	1.12	0.65	0.87				
Methanol	298 K	25	0.42	0.73	0.16	0.64				
Water	273–373 K	3440	0.95	0.96	0.11	0.93				
$\Delta G_{solv}^{exp,ionised}$ dataset										
Ionised solvation free energy			By solvent				Full dataset			
Solvent	Temperature	Datapoints	R^2	RMSD	Bias	SDEP	R^2	RMSD	Bias	SDEP
Methanol	298 K	77	0.61	3.60	0.51	3.06	0.74	3.96	0.35	3.62
Water	298 K	103	0.73	4.10	0.21	3.71				

single solvent as RMSD of 3.60 and 4.10 kcal mol^{-1} are achieved for predictions of solutes in methanol and water solvents, respectively. Fig. 2 provides the correlation plots of experimental SFE against predicted values for the neutral and ionised solute datasets. Ionised solute predictions are colour coded as anions and cations.

Solvation free energy predictions made with pyRISM-CNN are of comparable accuracy to the current state-of-the-art of the more computationally expensive 3D-RISM based methods. The semi empirical universal correction (UC) free energy functional paired with 3D-RISM has been shown to accurately predict hydration free energies for neutral and ionised solutes. Labute *et al.* calculated HFE values for a dataset of 504 neutral organic molecules, obtaining an RMSD of 1.18 kcal mol^{-1}.[17] A similar approach has been proposed by Casillas *et al.*,[64] in which an RMSD of 1.44 kcal mol^{-1} was achieved for 642 molecules from the Freesolv database. Sumi *et al.* performed hydration free energy calculations on an earlier version of the Freesolv database using the reference-modified density functional formulation, with which an RMSD of 1.46 kcal mol^{-1} was reached.[65] The authors also compared their approach against 3D-RISM/NgB and molecular DFT calculations performed on the same dataset, which managed RMSD of 1.29 and 1.80 kcal mol^{-1}, respectively. Tielker *et al.* performed HFE calculations on neutral and ionised organic solutes obtained from the MSD using the embedded cluster RISM theory (EC-RISM).[66] By including multiple solvent-specific empirical parameters, 3D-RISM calculated hydration free energies with an RMSD of 1.52, 4.48 and 2.91 kcal mol^{-1} were obtained for neutral, anionic and cationic solutes, respectively. Fewer studies have reported hydration free energy predictions for

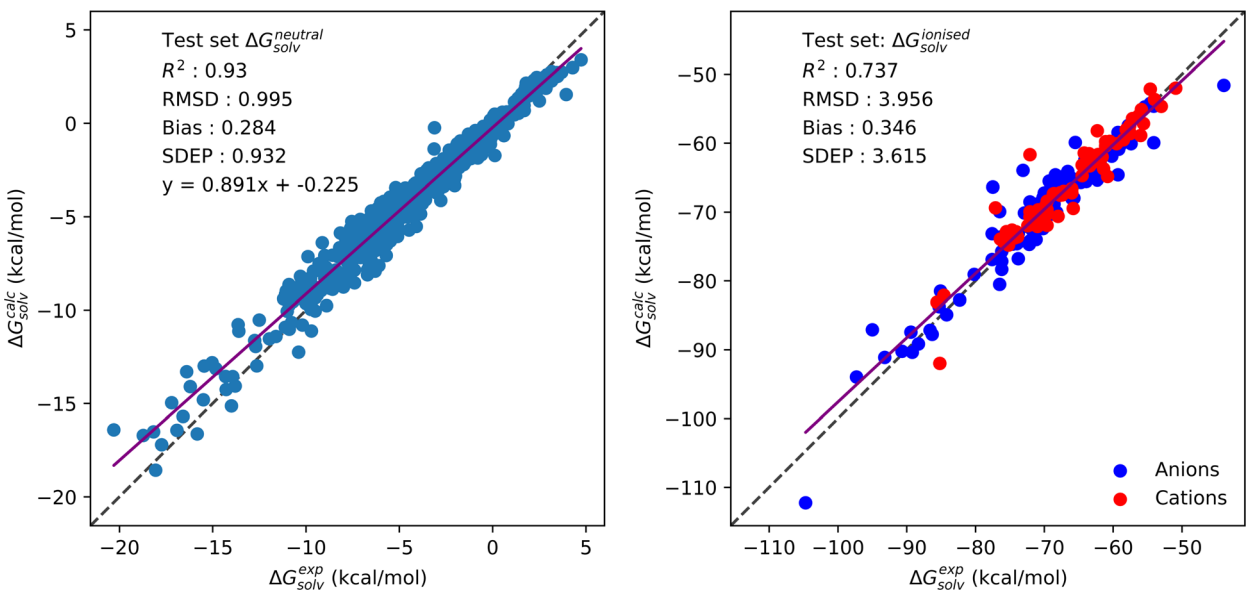

Fig. 2 Correlation plots showing solvation free energy predictions from CNN trained on the individual neutral and ionised datasets. SFED datasets were generated using the GF functional. The ionised dataset correlation plot is further separated into cation and anion based predictions. Errors are given in kcal mol^{-1}.

ionised datasets. Misin *et al.* reported HFE calculations for molecular ions involving 3D-RISM/PC+ and a correction for the Galvani potential, with an RMSD of 4.84 kcal mol^{-1}.[23] Johnson *et al.* treated 48 molecular ions in water with a variety of free energy functionals in 3D-RISM without a Galvani correction, achieving an RMSD of 6.51 kcal mol^{-1} with 3D-RISM/UC.[26]

4.2 Solvation free energy – combined dataset

Convolutional neural network models were trained on a combined dataset consisting of neutral and ionised solute SFED. In total, three separate models were tested with SFED generated from each free energy functional: KH, HNC and GF. Only models trained on GF based SFED will be discussed here. Models associated with KH and HNC generated SFED can be found in Section S3 of the ESI.†

Compared to the individual neutral and ionised datasets, use of a combined dataset results in only a small reduction in SFE prediction accuracy across solvents and neutral/ionised solutes. A breakdown of solvation free energy predictions for the combined dataset can be found in Table 3. By solvent, neutral solute RMSD increases by 0.23, 0.22 and 0.51 kcal mol^{-1} for chloroform, methanol and water, respectively, when compared against CNN trained on the individual neutral dataset. For carbon tetrachloride, RMSD decreases by 0.02 kcal mol^{-1}. Comparing ionised predictions, a CNN trained on the individual ionised dataset performed better than their combined dataset counterpart, with prediction errors increasing by 0.67 and 1.70 kcal mol^{-1} for methanol and water, respectively. A drop in SFE prediction accuracy is not unexpected for several reasons: experimental solvation free energies for ionised organic solutes are typically ten times greater than their neutral counterparts, and neutral solute data outnumbers ionised solute data by 20 : 1. Fig. 3 shows the correlation plot of experimental SFE against predicted values for the combined neutral and ionised solute dataset.

4.3 Solvation enthalpy, entropy & free energy

Multi-task CNN models were trained on the solvation thermodynamics dataset consisting of neutral organic solutes with experimental solvation enthalpy, entropy and free energy values. In total, three separate models were tested with SFED generated from each free energy functional: KH, HNC and GF.

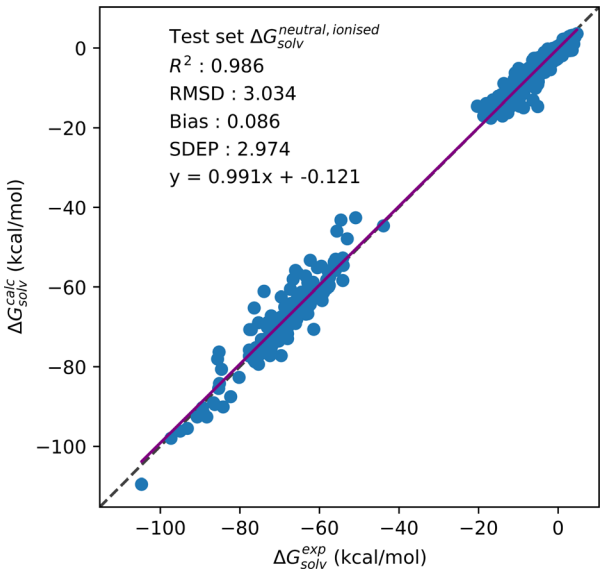

Fig. 3 Correlation plot showing solvation free energy predictions made with the combined neutral and ionised dataset. SFED were generated using the GF functional.

Single task CNN models were also trained on individual solvation enthalpy, entropy and free energy datasets across the KH, HNC and GF functionals, for a total of 9 single task models. Each single task model performed comparably to the multi-task CNN across each solvation property, with all of the statistics available in Section S5 of the ESI.† Correlation plots comparing pseudo-calculated values for solvation enthalpy, entropy and free energy parameters from single and multi-output models are available in Section S6 of the ESI.† Pseudo-calculated values are those determined using $\Delta G = \Delta H - T\Delta S$ and the predicted values for the two corresponding parameters. From these plots, it can be noted that multi-output CNN better learn the correlation between each solvation parameter than single output models.

Solvation enthalpy, entropy and free energy predictions from multi-output CNN trained on SFED generated from the KH, HNC or GF free energy functional can be found in Table 4. Across each free energy functional, solvation enthalpy, entropy

Table 3 Breakdown of solvation free energy predictions made using a CNN trained on a combined neutral and ionised dataset. SFED were generated using the GF free energy functional. Statistics are given across the full dataset, as well as separated by solvent and neutral/ionised data. Errors are given in kcal mol^{-1}

$\Delta G_{solv}^{exp,neutral}$, $\Delta G_{solv}^{exp,ionised}$ dataset			By solvent				Full dataset			
Solvent	Temperature	Datapoints	R^2	RMSD	Bias	SDEP	R^2	RMSD	Bias	SDEP
Neutral solvation free energy										
Carbon Tetrachloride	298 K	79	0.68	0.98	0.31	0.93				
Chloroform	298 K	109	0.74	1.35	0.29	1.32				
Methanol	298 K	25	0.40	0.95	−0.27	0.91				
Water	273–373 K	3440	0.88	1.47	−0.20	1.45	0.99	3.03	0.09	2.97
Ionised solvation free energy										
Methanol	298 K	77	0.52	4.27	0.40	3.83				
Water	298 K	103	0.59	5.80	0.48	5.55				

Table 4 Breakdown of solvation enthalpy, entropy and free energy predictions made using multi-output CNN trained on SFED generated using the KH, HNC or GF free energy functionals. Errors are given in kcal mol^{-1}

Multi-output CNN – ΔH_{solv}^{exp}, $T\Delta S_{solv}^{exp}$, $\Delta G_{solv}^{exp,neutral}$ dataset					
Functional	Parameter	R^2	RMSD	Bias	SDEP
KH	Solvation enthalpy	0.90	1.06	0.14	1.00
	Solvation entropy	0.77	1.00	0.10	0.95
	Solvation free energy	0.94	0.53	0.07	0.50
HNC	Solvation enthalpy	0.90	1.04	0.06	1.01
	Solvation entropy	0.79	0.98	0.03	0.95
	Solvation free energy	0.95	0.47	0.04	0.45
GF	Solvation enthalpy	0.89	1.14	0.03	1.08
	Solvation entropy	0.73	1.07	−0.03	1.01
	Solvation free energy	0.95	0.53	0.05	0.50

and free energy predictions are made with accuracies nearing or below 1 kcal mol^{-1}. By functional, RMSD of 1.06, 1.04 and 1.14 kcal mol^{-1} are obtained for solvation enthalpy predictions with KH, HNC and GF respectively. Similar values of 1.00, 0.98 and 1.07 kcal mol^{-1} are obtained for solvation entropy. Solvation free energy predictions remain the most accurate and closely resemble accuracies obtained with single task CNN, with RMSD of 0.53, 0.47 and 0.53 kcal mol^{-1}. These results suggest that use of a multi-task algorithm allows for the accurate prediction of all three properties simultaneously, as can be seen by comparing Tables 2–4. Fig. 4 provides the correlation plots of experimental solvation enthalpy, entropy and free energy against predicted values for the multi-task CNN trained on GF generated SFED.

Several methods have been reported for the prediction of solvation enthalpy and entropy, although few report predictions for both alongside solvation free energy. MD based free energy perturbation (FEP) calculations have been reported for a dataset of 239 neutral small molecules in water, achieving an average unsigned error (AUE) of 1.10 kcal mol^{-1}.[67] Comparisons can also be made against the SMD, which has been tested extensively against both aqueous and organic solvents at 298 K with which to calculate SFE for neutral and ionised solutes.[7] By solvent, AUE of 0.52, 0.84 and 0.59 kcal mol^{-1} were reported for neutral solutes in carbon tetrachloride, chloroform and water respectively, and AUE of 2.47 and 4.40 kcal mol^{-1} were also reported for ionised solutes in methanol and water. Although not directly comparable, here with GF based pyRISM-CNN models, RMSD values of 1.00, 1.12 and 0.96 kcal mol^{-1} were made for neutral solutes in carbon tetrachloride, chloroform and water, and 3.60 and 4.10 kcal mol^{-1} for ionised solutes in methanol and water. Jaquis *et al.* reported solvation enthalpy predictions for ethanol solvent systems using a deep learning feedforward neural network and chemistry development kit (CDK) descriptors, with a test set RMSD of 1.58 kcal mol^{-1}.[68] Similarly, Chung *et al.* developed a deep learning neural network model for the prediction of solvation enthalpy and free energy, reporting an RMSD of 0.75 and 0.71 kcal mol^{-1} respectively.[69] Irudayam *et al.* reported an MD based method of calculating the hydration entropy from individual entropic components, alongside solvation free energy.[70] With this method, hydration free energies are calculated with a mean unsigned error of 2.5 kJ mol^{-1}. The authors reported that solvation entropies, and enthalpies by extension using $\Delta G = \Delta H - T\Delta S$, are typically underestimated, and conclude that forcefield based methods are unable to account for the temperature dependence of solvation. Hydration free energies and hydration enthalpies/entropies were reported by Johnson *et al.* for a dataset of 1123 and subset of 74 molecules, respectively, calculated with the 3D-RISM theory and PC+ free energy functional.[26] Across solvation free energy, enthalpy and entropy, RMSD of 1.43, 2.12 and 1.93 kcal mol^{-1} were obtained respectively. However, empirical corrections to the standard 3D-RISM theory were necessary to obtain values with reasonable agreement to experiment.

This multi-task approach can be readily extended to organic solvent systems alongside the aqueous systems tested here. There are, however, limited samples available with which to train machine learning models. Increasing the availability of high quality experimental data in the literature would help to resolve this problem.

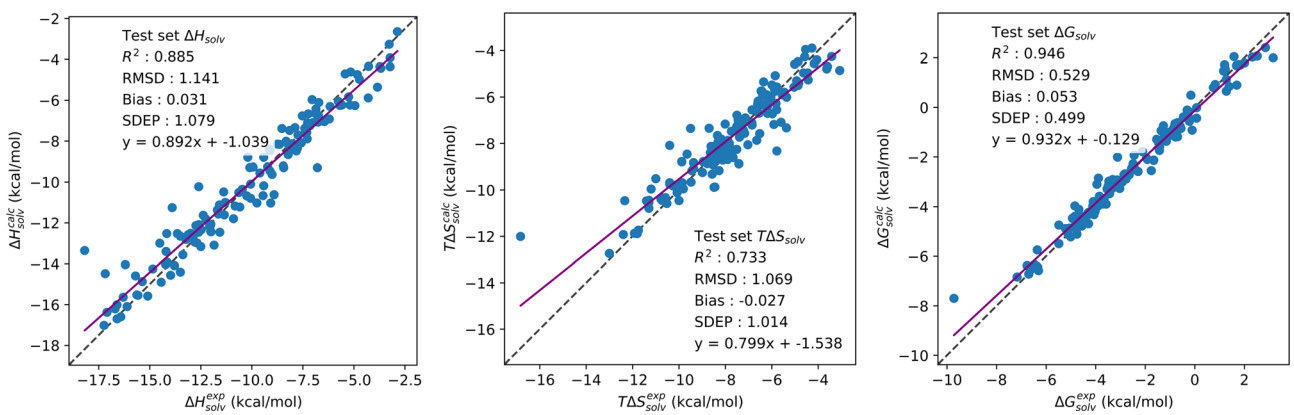

Fig. 4 Correlation plots of experimental solvation enthalpy, entropy and free energy values against predicted values for the solvation thermodynamics dataset. SFED were generated using the GF free energy functional and used to train a single multi-task CNN.

5 Conclusions

Previously, we presented a new method for accurately predicting solvation free energy, pyRISM-CNN, by combining the 1D-RISM solver, pyRISM, with a deep learning based free energy functional. With this deep learning approach, solvation free energy predictions could be made to within 1 kcal mol^{-1} of experiment across several different solvents and at temperatures beyond 298 K. These accuracies marked a 40-fold improvement in prediction accuracy when compared to the standard 1D-RISM theory. Here, we have reported several developments to the pyRISM-CNN methodology: enabling the prediction of SFE for organic molecular ions, to within 4 kcal mol^{-1} of experiment; introduced methanol as a solvent with which to train the pyRISM-CNN functional to predict solvation free energy; and successfully expanded pyRISM-CNN to a multi-task algorithm with the accurate and simultaneous prediction of solvation enthalpy, entropy and free energy for water solvent systems at 298 K. With this multi-task learning approach, three thermodynamic parameters fundamental to solvation processes can be accurately obtained without the need for extensive sampling, and can be determined as part of a standard 1D-RISM calculation with minimal additional computational expenditure.

Author contributions

Conceptualization: DJF, DSP. Methodology: DJF, DSP. Software: DJF. Validation/verification: DJF, DSP. Formal analysis: DJF. Investigation: DJF. Resources: DSP. Data curation: DJF. Writing – original draft: DJF. Writing – review & editing: DJF, DSP. Visualization: DJF. Supervision: DSP. Project administration: DSP. Funding acquisition: DSP.

Data availability

The pyRISM v0.1.1 code for solving 1D RISM equations and computing solvation free energy density functions can be found as freely available and open-source software at **https://zenodo.org/record/7107645**. Future versions of this software will be released at **https://github.com/2AUK/pyRISM**. Software and datasets used to develop the pyRISM-CNN models have been uploaded as part of the ESI.†

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

D. J. F. and D. S. P. thank the EPSRC for funding *via* Prosperity Partnership EP/S035990/1. The authors thank the ARCHIE-WeSt High-Performance Computing Centre (**www.archie-west.ac.uk**) for computational resources.

References

1 L. Xu and M. L. Coote, *J. Phys. Chem. A*, 2019, **123**, 7430–7438.
2 M. S. Bodnarchuk, D. M. Heyes, D. Dini, S. Chahine and S. Edwards, *J. Chem. Theory Comput.*, 2014, **10**, 2537–2545.
3 F. R. Dutra, C. de Souza Silva and R. Custodio, *J. Phys. Chem. A*, 2021, **125**, 65–73.
4 S. Genheden, T. Luchko, S. Gusarov, A. Kovalenko and U. Ryde, *J. Phys. Chem. B*, 2010, **114**, 8505–8516.
5 D. S. Palmer, J. L. McDonagh, J. B. O. Mitchell, T. van Mourik and M. V. Fedorov, *J. Chem. Theory Comput.*, 2012, **8**, 3322–3337.
6 D. J. Fowles, D. S. Palmer, R. Guo, S. L. Price and J. B. O. Mitchell, *J. Chem. Theory Comput.*, 2021, **17**, 3700–3709.
7 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
8 J. Tomasi, B. Mennucci and E. Cances, *J. Mol. Struct.*, 1999, **464**, 211–226.
9 S.-T. Lin and C.-M. Hsieh, *J. Chem. Phys.*, 2005, **125**, 124103.
10 J. E. Bara, J. D. Moon, K. R. Reclusado and J. W. Whitley, *Ind. Eng. Chem. Res.*, 2013, **52**, 5498–5506.
11 F. Fogolari, A. Corazza and G. Esposito, *Front. Mol. Biosci.*, 2018, **5**, 11.
12 M. Karplus, T. Ichiye and B. M. Pettitt, *Biophys. J.*, 1987, **52**, 1083–1085.
13 V. Ovchinnikov, M. Cecchini and M. Karplus, *J. Phys. Chem. B*, 2013, **117**, 750–762.
14 D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts and K. A. Dill, *J. Chem. Theory Comput.*, 2009, **5**, 350–358.
15 S.-T. Lin, P. K. Maiti and W. A. G. III, *J. Phys. Chem. B*, 2010, **114**, 8191–8198.
16 F. Waibl, J. Kraml, M. L. Fernández-Quintero, J. R. Loefer and K. R. Liedl, *J. Comput.-Aided Mol. Des.*, 2022, **36**, 101–116.
17 J.-F. Truchon, B. M. Pettitt and P. Labute, *J. Chem. Theory Comput.*, 2014, **10**, 934–941.
18 D. S. Palmer, A. I. Frolov, E. L. Ratkova and M. V. Fedorov, *J. Phys.: Condens. Matter*, 2010, **22**, 492101.
19 V. Sergiievskyi, G. Jeanmairet, M. Levesque and D. Borgis, *J. Chem. Phys.*, 2015, **143**, 184116.
20 M. Misin, M. V. Fedorov and D. S. Palmer, *J. Chem. Phys.*, 2015, **142**, 091105.
21 S. Tanimoto, N. Yoshida, T. Yamaguchi, S. L. Ten-no and H. Nakano, *J. Chem. Inf. Model.*, 2019, **59**, 3770–3781.
22 D. Roy and A. Kovalenko, *J. Phys. Chem. A*, 2019, **123**, 4087–4093.
23 M. Misin, M. V. Fedorov and D. S. Palmer, *J. Phys. Chem. B*, 2016, **120**, 975–983.
24 M. Misin, P. A. Vainikka, M. V. Fedorov and D. S. Palmer, *J. Phys. Chem.*, 2016, **145**, 194501.
25 M. Misin, M. V. Fedorov and D. S. Palmer, *J. Phys. Chem. B*, 2016, **120**, 975–983.
26 J. Johnson, D. A. Case, T. Yamazaki, S. Gusarov, A. Kovalenko and T. Luchko, *J. Phys.: Condens. Matter*, 2016, **28**, 344002.

27 S. J. Singer and D. Chandler, *Mol. Phys.*, 1985, **55**, 621–625.
28 K. Sato, H. Chuman and S. Ten-no, *J. Phys. Chem. B*, 2005, **109**, 17290–17295.
29 S. Ten-no, J. Jung, H. Chuman and Y. Kawashima, *Mol. Phys.*, 2010, **108**, 327–336.
30 D. J. Fowles, R. G. McHardy, A. Ahmad and D. S. Palmer, *Digital Discovery*, 2023, **2**, 177–188.
31 A. Ahmad, *2AUK/pyRISM: v0.1.1*, 2022, DOI: **10.5281/zenodo.7107645**.
32 E. L. Ratkova, D. S. Palmer and M. V. Fedorov, *Chem. Rev.*, 2015, **115**, 6312–6356.
33 A. Kovalenko and F. Hirata, *J. Phys. Chem. B*, 1999, **103**, 7942–7957.
34 S. Ten-no, *J. Chem. Phys.*, 2001, **115**, 3724–3731.
35 F. Hirata, *Molecular Theory of Solvation*, Springer Dordrecht, Dordrecht, The Netherlands, 1st edn, 2003.
36 V. P. Sergiievskyi, W. Hackbusch and M. V. Fedorov, *J. Comput. Chem.*, 2011, **32**, 1982–1992.
37 D. S. Palmer, M. Misin, M. V. Fedorov and A. Llinas, *Mol. Pharmaceutics*, 2015, **12**, 3420–3432.
38 A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer and D. G. Truhlar, *Minnesota Solvation Database – version 2012*, University of Minnesota, Minneapolis, 2012.
39 A. C. Chamberlin, C. J. Cramer and D. G. Truhlar, *J. Chem. Phys. B*, 2006, **110**, 5665–5675.
40 A. C. Chamberlin, C. J. Cramer and D. G. Truhlar, *J. Chem. Phys. B*, 2008, **112**, 3024–3039.
41 C. C. Zanith and J. R. P. Jr., *J. Comput.-Aided Mol. Des.*, 2015, **29**, 217–224.
42 A. Lang, *Acree Enthalpy of Solvation Dataset*, 2015, DOI: **10.6084/m9.figshare.1572326.v1**, Figshare.
43 C. Mintz, M. Clark, W. E. A. Jr. and M. H. Abraham, *J. Chem. Inf. Model.*, 2007, **47**, 115–121.
44 A. J. Garza, *J. Chem. Thoery Comput.*, 2019, **15**, 3204–3214.
45 A. Ben-Naim, *J. Phys. Chem.*, 1978, **82**, 792–803.
46 A. Ben-Naim and Y. Marcus, *J. Chem. Phys.*, 1984, **81**, 2016–2027.
47 S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi, *J. Cheminform*, 2015, **7**, 23.
48 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
49 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2020, **49**, D1388–D1395.
50 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.
51 J. Perkyns and B. M. Pettitt, *J. Chem. Phys.*, 1992, **97**, 7656–7666.
52 J. Perkyns and B. M. Pettitt, *Chem. Phys. Lett.*, 1992, **190**, 626–630.
53 P. H. Lee and G. Maggiora, *J. Phys. Chem.*, 1993, **97**, 10175–10185.
54 L. Lue and D. Blankschtein, *J. Phys. Chem.*, 1992, **96**, 8582–8594.
55 F. Hirata, P. J. Rossky and B. M. Pettitt, *J. Chem. Phys.*, 1983, **78**, 4133.
56 G. N. Chuev, M. V. Fedorov and J. Crain, *Chem. Phys. Lett.*, 2007, **448**, 198–202.
57 H. A. D. A. Case, K. Belfon, I. Ben-Shalom, J. Berryman, S. Brozell, D. Cerutti, T. Cheatham, G. C. III, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, M. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, J. Wang, H. Wei, R. Wolf, X. Wu, Y. Xiong, Y. Xue, D. York, S. Zhao and P. Kollman, *Amber 2018*, 2018, **http://ambermd.org/**.
58 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1989.
59 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015, **https://www.tensorflow.org/**, Software available from tensorflow.org.
60 F. Chollet *et al.*, Keras, 2015, **https://github.com/fchollet/keras**.
61 A. F. Agarap, *Deep learning using rectified linear units (relu)*, 2018, DOI: **10.48550/arXiv.1803.08375**, arXiv.
62 D. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2014, DOI: **10.48550/arXiv.1412.6980**, arXiv.
63 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *JMLR*, 2011, **12**, 2825–2830.
64 L. Casillas, V. M. Grigorian and T. Luchko, *Molecules*, 2023, **28**, 925.
65 T. Sumi, A. Mitsutake and Y. Maruyama, *J. Comput. Chem.*, 2015, **36**, 1359–1369.
66 N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Gussregen, K. F. Schmidt and S. M. Kast, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 1035–1044.
67 D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley and W. Sherman, *J. Chem. Theory Comput.*, 2010, **6**, 1509–1519.
68 B. J. Jaquis, A. Li, N. D. Monnier, R. G. Sisk, W. E. A. Jr. and A. S. I. D. Lang, *J. Solut. Chem.*, 2019, **48**, 564–573.
69 Y. Chung, F. H. Vermeire, H. Wu, P. J. Walker, M. H. Abraham and W. H. Green, *J. Chem. Inf. Model.*, 2022, **62**, 433–446.
70 S. J. Irudayam, R. D. Plumb and R. H. Henchman, *Faraday Discuss.*, 2009, **145**, 467–485.

