Fluid Phase Equilibria 592 (2025) 114335

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.journals.elsevier.com/fluid-phase-equilibria

Self-solvation energies: Extended open database and GNN-based prediction

Hugo Marques [a], Simon Müller [b,*]

[a] Centro de Química Estrutural and Institute of Molecular Sciences, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, 1049-001 Lisboa, Portugal
[b] Hamburg University of Technology, TUHH, Institute of Thermal Separation Processes, Eißendorfer Straße 38 (O), 21073, Hamburg, Germany

ARTICLE INFO

Keywords:
Free energy
Solvation
Artificial neural networks
Machine learning models
Property Prediction

ABSTRACT

Solvation energies play a crucial role in various chemical processes, ranging from chemical synthesis to separation techniques. To optimize these processes, it is essential to accurately predict solvation energies across different temperatures and solvents. However, most existing studies primarily focus on the standard temperature of 298.15 K. In this work, we address this limitation by creating an extensive database, which combines the DIPPR and Yaws databases. Our comprehensive dataset includes 5420 pure compounds, resulting in 71,656 data points spanning a wide range of temperatures. Additionally, besides the development of this novel database, another key contribution of this work is the coupling of the well-known Graph Convolutional Neural Network Chemprop, with our database with the aim of predicting self-solvation energies across diverse temperatures for the first time. The results presented here demonstrate the overall effectiveness of the model, evidenced by a low Mean Absolute Error (MAE) of 0.09 kcal mol^{-1} and a high Determination Coefficient (R^2) of 0.992. Additionally, the Average Relative Deviation (ARD) of the data is 2.2 %, further confirming the accuracy of the model. In fact, the model demonstrates robust predictive performance across data of varying quality, including a significant fraction of pseudo-experimental values derived from predictive schemes. However, it is worth noting that some groups of compounds, such as small sized compounds and low-numbered ring structures, exhibited somewhat larger deviations than expected. This suggests areas for further refinement and indicates that while the model is robust, there is still room for improvement in specific cases. This approach represents an overall improvement over previous models and offers enhanced versatility for practical applications in chemical synthesis and separation processes.

1. Introduction

Solvation is a fundamental physico-chemical process wherein a small quantity of a solid or gaseous substance, called the solute, disperses in large quantities of a liquid compound or mixture known as the solvent [1]. Various definitions of solvation exist in scientific literature, often emphasizing different aspects of the physical interactions involved, such as electrostatic forces, van der Waals forces, and hydrogen bond formation [2]. According to the International Union of Pure and Applied Chemistry (IUPAC), solvation energy is defined as the change in Gibbs energy when an ion or molecule transfers from a vacuum (or gas phase) to a solvent [3]. This definition, also adopted in this work, highlights the thermodynamic nature of solvation and its dependency on local molecular interactions [4].

The quantification of solvation effects is crucial for various practical and theoretical applications, from chemical synthesis and purification processes to drug design and environmental management [5–7]. For instance, solvation energies are used to measure and compare the affinities of solutes for solvents, which is essential in selecting appropriate solvents for specific tasks [8]. In pharmacology, solvation properties influence the behavior of active molecules in different environments, affecting their transport in biological systems and their interaction with target molecules [9]. Moreover, solvation quantities are also vital in studying protein structure stability, drug design, and the interaction of nanoparticles with biological systems [10,11]. This is due to solvation Gibbs free energy (Δg_{solv}) of a solute in a solvent being directly related to the partition coefficient of the solute between the gas and solvent phases [8]. Despite these values being typically reported at room temperature, this property is crucial for predicting the liquid-liquid partition coefficient and solid solubility of the solute in organic solvents [12].

More specifically, the self-solvation energy of compounds is crucial in several areas, including thermodynamics, molecular design and engineering, and environmental impact assessment [13–15]. In terms of molecular stability and conformation, self-solvation energy reflects the

* Corresponding author.
E-mail address: simon.mueller@tuhh.de (S. Müller).

https://doi.org/10.1016/j.fluid.2025.114335
Received 18 November 2024; Received in revised form 5 January 2025; Accepted 7 January 2025
Available online 8 January 2025

intramolecular interactions of the solute with itself. In predictive modeling, accounting for self-solvation energy is essential. Models that incorporate these interactions can more accurately predict solvation energies across different solvents, thereby improving the reliability of simulations and calculations used in various chemical engineering applications [14,15]. Therefore, accurate Δg_{solv} values, whether for pure substances or mixtures, are necessary for optimizing processes under different operating conditions.

The wide range of applications for this property has fueled the growing interest in solvation free energies. This has led to the development of numerous predictive methods, ranging from molecular dynamics and quantum chemistry methods to empirical and data-driven approaches [8,16]. Quantum chemistry methods often utilize the implicit polarizable continuum model for solvent representation, such as the SMx methods by Cramer et al. and Marenich and co-works, [17–19] and the COSMO-RS models by Klamt et al. [20–23]. Afterwards, Müller et al. [24] have published a recent advancement on this field with the development of openCOSMO-RS [25], an improved open-source COSMO-RS model capable of predicting solvation free energies alongside other liquid-phase properties. This model demonstrated significant accuracy for the prediction of solvation free energies of binary systems at 298.15 K. Another field where this open-source model excelled was in the prediction of halocarbon properties [26]. Additionally, Kröger et al. [27] also used COSMO-RS-based models to study ion solvation free energy, where they observed a high level of accuracy in their predictions for these ionic systems which usually present unique challenges.

A different alternative is the LSER model, developed by Abraham et al. [28,29], relating the free energy of solvation to molecular descriptors through a linear equation. Several methods have also been proposed to estimate some or all the solute parameters required for the LSER model from its molecular structure [28,30–39]. Some of those include contribution (GC) approaches. For instance, Platts et al. [34,35] were the first to devise a GC scheme that can predict all solute parameters, where they developed 81 functional group fragments for prediction. Afterwards, other works [40] adopted the Platts-type fragments and further optimized the fragments for their module. This approach was shown to give reasonably good estimates for many compounds, but it had large prediction errors for certain classes of molecular structure, such as highly halogenated compounds, triazoles, and bridged ring structures. Additionally, Brown et al. [41,42] developed a different set of fragments or substructures using the iterative fragment selection approach, in which the fragments are selected by using k-fold cross-validation. In their work, solute parameter data of around 3700 compounds was used to build an open-access GC model that is available through the UFZ-LSER database. Their model showed good predictive performance for some parameters, while the performance on the other solute parameters was not reported.

On the topic of self-solvation energies, these are typically associated with vapor pressure. One example of this is the work by Moine et al. [1]. In their work, they developed the CompSol database, which contains extensive experimental solvation data for both pure substances and mixtures. For their self-solvation energies, these were calculated using physical properties such as liquid density and vapor pressure. Moreover, other works, such as those by Wang et al. [43] and Tsai and Lin [44], utilized the PR + COSMO-SAC Equation of State (EoS), which includes a self-solvation term. This EoS integrates quantum mechanical solvation calculations to determine the energy and molecular volume parameters in the PR EoS, highlighting the importance of self-solvation calculations for property prediction, particularly in predicting vapor pressure.

This leads us to conclude that predictive methods continue to evolve, incorporating empirical data and advanced computational techniques. For example, group contribution methods and state-of-the-art graph-convolutional message passing neural networks are increasingly used to enhance predictions of Δg_{solv} and Δh_{solv} at 298.15 K [8]. However, despite the extensive use and importance of solvation data, the existing studies and databases often fall short in several significant ways. One major limitation is the narrow focus on standard temperature conditions, typically at 298.15 K. This constraint severely limits the practical applicability of the data, as real-world chemical processes often occur over a range of temperatures. The standard temperature data do not provide insights into the temperature-dependent behavior of solvation, which is critical for understanding and optimizing processes under varying thermal conditions.

Moreover, despite the availability of these predictive methods, the scarcity and quality of experimental data remain a significant bottleneck. To address this, large databases like the Minnesota Solvation (MNSol) [45] database and the FreeSolv [46] database have been developed. The MNSol database, updated in 2012, contains experimental values of molar Gibbs energy of solvation for over 3000 binary systems [1]. Hutchinson and Kobayashi [47] first used this database to incorporate different solvents in a neural network using functional class fingerprints. The FreeSolv databse, introduced in 2014, provides hydration Gibbs energy values for >600 solutes at 298.15 K [1]. The experimental values in the latter were derived from various sources, while the calculated ones come from alchemical free energy calculations using molecular dynamics simulations [48]. Additionally, it is extensively used both as an input for new calculations and as a benchmark for comparing different computational methods [48,49], particularly in the SAMPL blind prediction challenge, where it serves as a critical reference. Furthermore, there were also some works being published following a larger database Solv@TUM [50]. Two works that are based on this database were published by Pathak et al. [51] and Lim and Jung [52]. The former proposed a chemically interpretable graph interaction network model with message passing, interaction, and prediction phases using 6239 unique solvent-solute combinations from Solv@TUM and FreeSolv data sets, while the latter developed MLSolvA, a Machine-Learning (ML) architecture that computes pairwise atomic interactions from solvent and solute atomistic feature vectors and makes predictions by summing these interactions.

With this being said, our work aims to create a wider comprehensive self-solvation database that combines data from the DIPPR [53] and Yaws [54] databanks, resulting in a larger dataset of 5420 pure compounds and 71,656 data points spanning a wide range of temperatures. This extensive dataset not only addresses the issue of temperature variability but also provides a robust foundation for developing predictive models. In fact, we intend to apply this database and consequent results to other thermodynamic models, like openCOSMO-RS, to improve the description of temperature dependence and, afterwards, vapor pressure prediction, as demonstrated in other works with similar objective of enhancing model applicability [55–60]. Furthermore, we integrate a Machine-Learning model, specifically a Graph Neural Network (GNN), with this new database, using it for training and testing its predictions. This model is designed to predict self-solvation energies across diverse temperatures, providing a versatile tool that extends beyond the constraints of standard temperature conditions.

2. Methodology

2.1. Theoretical framework and database curation

Solvation energy, as defined by Ben-Naim, refers to the process in which a molecule transitions from an ideal gas mixture to a real-fluid mixture, typically assumed to be a liquid, under constant temperature and pressure conditions [1]. This transition is quantified by the partial molar Gibbs energy of solvation, also known as the chemical potential of solvation. This quantity indicates the change in chemical potential that occurs during the solvation process [4]. Assuming that the gas-phase internal partition function remains unaffected by the solvent, a general expression for the chemical potential of solvation can be derived [1, 4], as represented by Eq. (1).

$$\Delta_{solv}\bar{g}_i(T,P,x) = RT\ln\left[\frac{P\,\varphi_i^{liq}(T,P,x)}{RT\,\rho_i^{liq}(T,P,x)}\right] \quad (1)$$

with $\Delta_{solv}\bar{g}_i$ being the chemical potential of solvation of mixed i, T the temperature, P the pressure, x the molar fractions x_i of all components i of a system, R the gas constant, φ_i^{liq} the fugacity coefficient of i in the liquid, and ρ_i^{liq} the molar density of i in the liquid.

It is noteworthy that this equation is derived directly from the definition of the solvation process without requiring any additional assumptions, thus making it applicable to mixtures containing an arbitrary number of components. For the purposes of this study, which focuses exclusively on self-solvation energies, the system will contain only a single compound. Under these conditions, Eq. (1) simplifies to Eq. (2).

$$\Delta_{solv}g_i^{pure}(T,P) = RT\ln\left[\frac{P\,\varphi_i^{pure\ liq}(T,P)}{RT\,\rho_i^{pure\ liq}(T,P)}\right] \quad (2)$$

where the $\Delta_{solv}g_i^{pure}$, $\varphi_i^{pure\ liq}$, and $\rho_i^{pure\ liq}$ are the chemical potential of solvation of pure i, the fugacity coefficient of pure liquid i, and the density of pure liquid i, respectively.

However, the molar density is typically reported for a larger dataset under saturated conditions, specifically for a pure liquid phase in equilibrium with its vapor [1]. This substitution is applicable when the pure component i is in vapor-liquid equilibrium (VLE) at temperature T, leading to the formulation of Eq. (3).

$$\Delta_{solv}g_i^{sat}(T) = RT\ln\left[\frac{P_i^{sat}(T)\varphi_i^{sat}(T)}{RT\,\rho_i^{sat}(T)}\right] \quad (3)$$

where $\Delta_{solv}g_i^{sat}$ is the saturation chemical potential of solvation of pure i, φ_i^{sat} is the fugacity coefficient of pure i in VLE that is common to the saturated-liquid and saturated-vapor phases, ρ_i^{sat} is the saturated-liquid density of pure i, and P_i^{sat} is the vapor pressure of i.

In Eq. (3), vapor pressure is a well-documented property, alongside density measurements. Still, fugacity coefficients cannot be readily estimated under all conditions from experimentally accessible properties, justifying the simplification from Eqs. (2) to (3). This is significant because fugacity coefficients are not directly measurable and are usually absent from common databases. Nevertheless, they can be derived from models, such as the law of corresponding states [61] or the truncated virial equation of state (EoS) [1], which involves the second virial coefficient - a property frequently tabulated in pure-component databases.

To estimate the self-solvation energies of compounds listed in various databases, experimental data of pure-component thermodynamic properties were utilized. These properties span extensive ranges of temperature and pressure, and the previously mentioned Eq. (3) was applied according to its respective assumptions. Under the framework established by the above equation, estimating solvation energies for pure compounds at a specific temperature T_k requires the saturation pressure P_k^{sat}, saturated-liquid density ρ_k^{sat}, and saturation fugacity coefficient φ_k^{sat}.

In this study, the DIPPR (Design Institute for Physical Properties) [53] database was used to generate pseudo-experimental data. The Yaws [54] database served as a supplementary source for compounds lacking physical property information in the DIPPR database. These databases provide direct data for P^{sat} and ρ_{liq}^{sat}, as temperature-dependent correlations which in most cases are fitted to experimental data for each pure component.

Following a similar methodology to that proposed by Moine et al. [1], which also used the DIPPR database, the process starts with filtering which compounds present critical temperatures T_c below 950 K, as species with higher critical temperatures than that are usually pure metals, which are beyond the scope of this study. Afterwards, the applicable temperature range was defined based on the lower and upper bounds of each temperature-dependent correlation for liquid density and vapor pressure to ensure both correlations could be applied. Specifically, the minimum temperature T_{min} for each compound was determined by the largest value between $T_{min}^{\rho_{liq}^{sat}}$ and $T_{min}^{P^{sat}}$, where $T_{min}^{\rho_{liq}^{sat}}$ and $T_{min}^{P^{sat}}$ are the lower bounds for the liquid density and vapor pressure correlations, respectively, while, in a similar manner, the maximum temperature T_{max} was given by the smallest value between $T_{max}^{\rho_{liq}^{sat}}$and $T_{max}^{P^{sat}}$, where $T_{max}^{\rho_{liq}^{sat}}$ and $T_{max}^{P^{sat}}$ are the upper bounds for the respective correlations. If no correlation is available for either the liquid density or the vapor pressure, the pure component is excluded from consideration. Subsequently, the values were evaluated at different intervals based on the amplitude of the temperature range, where Moine et al. go into further details regarding the procedure.

Based on these temperature points T_k, it is possible to obtain values for the saturated liquid density and vapor pressure, ρ_k^{sat} and P_k^{sat}, at those temperatures, respectively. Regarding the fugacity coefficients φ_k^{sat}, these were estimated using the corresponding-states law as proposed by Hougen, Watson, and Ragatz (HWR) [61]. This method provides correlations for calculating fugacity coefficients, although it is only applicable to molecules with a critical compressibility factor between 0.25 and 0.29. Nevertheless, this 3-parameter HWR corresponding-states law is known for its accuracy and applicability from low to critical pressures. For other molecules, the truncated virial EoS [1] mentioned above can be used under certain conditions. This is usually suited to model weakly non-ideal gases. The procedure for each of the possible situations regarding this calculation is as follows:

(1) When the reduced pressure $P_{r_k}^{sat}$ is lower than 0.05, the pure gaseous component in VLE is considered to exhibit ideal gas behavior, allowing the saturation fugacity coefficient to be approximated as 1.
(2) Afterwards, for water or other pure compounds with critical compressibility factor z_c within the range $0.25 \leq z_c \leq 0.29$, the HWR corresponding-states law was applied. Further information regarding this method can be found in [61].
(3) Next, in the cases outside of the conditions stated above, the truncated virial EoS was used provided the reduced saturated pressure $P_{r_k}^{sat}$ does not exceed 0.5. This model expresses the saturation fugacity coefficient as shown in Eq. (4). Here, the second virial coefficient, B, is obtained from DIPPR correlations, assuming temperature T falls within the second virial coefficient's application range. Since the Yaws databank does not present tabulated values for this property, data that stems from this databank that does not meet the criteria in steps (1) or (2) were not considered further.

$$\varphi^{sat}(T) = \exp\left[\frac{B(T)P^{sat}(T)}{RT}\right] \quad (4)$$

(4) If none of the above cases are valid, the compound temperature point is removed from the database.

With the combination of the saturated liquid density, vapor pressure, and saturation fugacity coefficient, it is possible to obtain the self-solvation energy $\Delta_{solv}g_k^{sat}$ for each temperature value. A schematic of the described procedure is presented in Fig. 1.

Overall, by combining the DIPPR and Yaws databanks, a large dataset of 5420 pure compounds, corresponding to 71,656 data points, was obtained. These data points were subsequently used as inputs for a Machine Learning (ML) model, with the objective of developing a model capable of making temperature-dependent predictions of self-solvation energies.

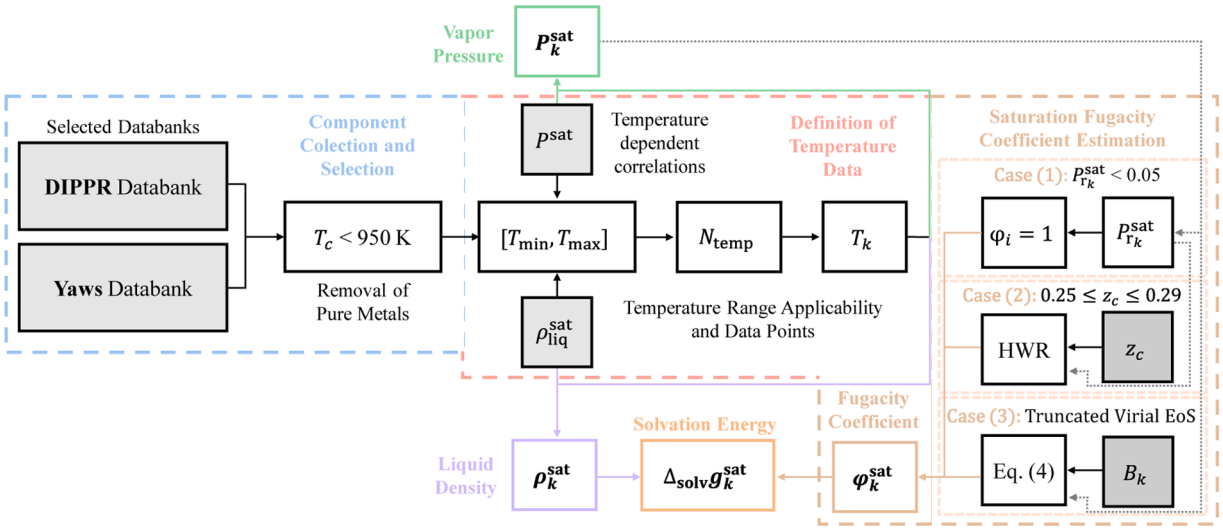

Fig. 1. Flowchart for the procedure to generate self-solvation energy data ($\Delta_{solv}g_k^{sat}$) from the DIPPR and Yaws databanks. The process involves selecting data for compounds that are not pure metals ($T_c < 950\ K$) and determining their temperature-dependent properties, such as vapor pressure (P_k^{sat}) and saturated liquid density (ρ_k^{sat}). The saturation fugacity coefficient (φ_k^{sat}) is estimated based on specific conditions. HWR refers to the Hougen, Watson, and Ragatz corresponding-states law, while Eq. (4) represents the Truncated Virial EoS:

$$\varphi^{sat}(T) = \exp\left[\frac{B(T)P^{sat}(T)}{RT}\right].$$

2.2. Machine-learning model framework

In this work, a GNN which excels at property prediction [8], namely Chemprop [49,62] was used to create the first predictive model for temperature dependent self-solvation energies. This tool has already been used to predict standard solvation energies at the temperature of 298.15 K [8], establishing itself as great candidate for the target property of this work. For a more thorough description of the features and applications of Chemprop, the reader is suggested to explore in-depth the works of its developers found in [49,62].

For the case of standard self-solvation energy at a specific temperature, the only inputs required would be the SMILES of the pure chemical compounds and its corresponding solvation energy. However, since the scope of this work is dealing with temperature dependencies, these require to be added as additional features to the input file. This combination of SMILES, corresponding temperature and self-solvation energy for each point can be found in the **Supporting Information**. To provide a clearer understanding of the behavior of the data, a graphical representation of self-solvation energy as a function of temperature for each compound has been included in the **Supporting Information**. Additionally, histograms showing the distribution of compounds by critical temperature, molar mass, and number of rings - key properties discussed further in this work - have also been added. These visual aids offer deeper insights into the characteristics and diversity of the dataset utilized in this study.

In this regression task, the Chemprop model was trained on the SMILES and solvation energy data, together with the temperature as an additional feature. The model employs a three-layer Message Passing Neural Network with the Rectified Linear Unit (ReLU) activation function, operating on molecular graph representations. Atom and bond descriptor scaling was enabled, and the mean aggregation method was used with a normalization factor of 100. Training was executed over 30 epochs with a batch size of 50. This value of epochs proved to be enough, as no substantial improvement seemed to occur near the last few epochs. An initial learning rate of 0.0001, and a maximum learning rate of 0.001 was used. The network was optimized using mean squared error (MSE) as the loss function, with the dataset split into training, validation, and test sets using the standard splitting of 80, 10, and 10 %, respectively. Further information regarding the specifics of Chemprop can be found in the appropriate sources [49,62].

After the network was trained and predictions obtained, these were evaluated using different metrics. These comprised of the Mean Absolute Error (MAE), the Determination Coefficient (R^2), and the Average Relative Deviation (ARD). These can be computed through the following equations:

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \tag{5}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{6}$$

$$ARD = \frac{1}{n}\sum \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{7}$$

where n is the number of samples (data points), y_i is the target value for sample i, $\hat{y}_i$ is the predicted output value for sample i, and $\bar{y}$ is the average of the target samples values.

3. Results

To provide a baseline comparison, it was initially calculated the standard self-solvation energies at 298.15 K for 5387 compounds in our database. Although this single-temperature model achieved a relatively low MAE, the R^2 value indicated poor correlation between predictions and experimental data, possibly due to the limitation of using a small dataset for Chemprop. A detailed summary of these results is provided in the **Supporting Information**. However, given the suboptimal performance of the single-temperature model, our focus was shifted to a temperature-dependent model, which seemed to offer superior predictive accuracy and generalization ability.

Examining the temperature-dependent results, data for all 5420 compounds was used from our proposed extended dataset. Fig. 2 illustrates the performance of the model using a parity plot between the true and the predicted values. A full numerical description of these results can be found in the **Supporting Information**, together with relevant statistical metrics.

The results for the temperature-dependent data show significantly better performance compared to the single-temperature data, which

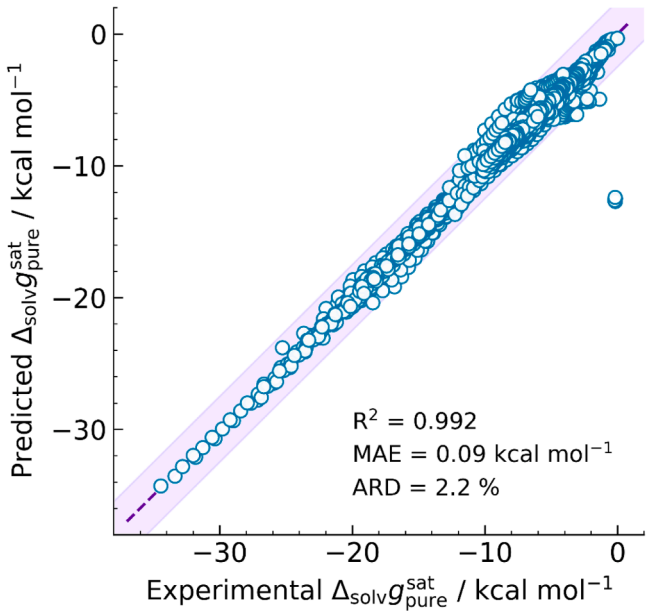

Fig. 2. Fitness plot for temperature-dependent $\Delta_{solv}g^{sat}_{pure}$ prediction using Chemprop model. The purple band corresponds to a deviation of 2.5 kcal mol^{-1}. Statistical metrics for this model are also displayed in the figure.

corroborates our proposed hypothesis of the increase in data size resulting in a better generalization ability of the model. A comparison between metrics of the different models can be seen in Table 1. There is a higher degree of correlation between the target values and the corresponding predictions, as indicated by the increase in R^2. This metric is now comparable to the previously mentioned results from other studies on single-temperature data. Additionally, the ARD has decreased significantly, demonstrating the closeness of the predictions to the actual data samples. This improvement in performance is also evident in the reduction of MAE, which is now lower than 0.10 kcal mol^{-1}. In fact, most of the results fall within the defined band in the figure, corresponding to a deviation of 2.5 kcal mol^{-1}. Given that the self-solvation energies of our compounds can reach up to around 35 kcal mol^{-1} for some compounds, this indicates that the deviations in our predictions are relatively small in absolute terms. This suggests that our model predictions are generally close to the actual values for most of the studied compounds.

Moreover, a more in-depth assessment of the training, validation, and testing statistical metrics was conducted to further evaluate the generalization capabilities of the model and check for potential overfitting. When analyzing only the training and validation data, the MAE values were found to be 0.09 and 0.10 kcal mol^{-1}, respectively, while R^2 values were 0.995 and 0.992, respectively. These results were expected, as this data was used to train the network, meaning that predictions from these sets would naturally exhibit high accuracy. However, the most crucial metric came from the testing subset, where the MAE was 0.10 kcal mol^{-1} and R^2 was 0.993 - both closely aligning with the overall results presented in Fig. 2. This consistency suggests that the model does not suffer from overfitting and, instead, demonstrates strong generalization ability. Consequently, given the similarity between the testing and overall results, the subsequent statistical metrics reported for the rest of this work report the entire extended database. A summary of these supporting metrics can be found in Table 2. For the sake of reproducibility, it is also presented in the **Supporting Information** a full description of which data is contained in the training, validation, and testing subsets of the overall data. Additionally, relative deviation plots for this model are provided in the **Supporting Information**, offering a clear visualization that the model does not exhibit systematic bias toward overestimation or underestimation, as the residuals are symmetrically distributed around zero and show no discernible pattern.

Additionally, the temperature-dependent model was evaluated at a fixed temperature of 298.15 K to facilitate a direct comparison with the single-temperature model. The predicted values, along with the associated metrics, are provided in the **Supporting Information**. The results indicate an improvement in both MAE and ARD, which decreased to 0.75 kcal mol^{-1} and 7.9 %, respectively. However, R^2 exhibited a decrease, with a reduced value of 0.768, suggesting a decline in the model's ability to capture the variance in the data. This indicates that this refined model offers more precise predictions when compared to the previous, albeit doing it at the cost of slightly reduced explanatory power, characterized by a lower R^2. This trade-off indicates that while the predictions of the model are closer to the true values on average, its ability to generalize and capture the overall trend in the data has diminished somewhat. This makes it evident that both models exhibit limitations when applied to single-temperature data, which is not the main subject of this study. Nonetheless, to further validate our results, both the experimental values from our database and the corresponding predictions generated by the single-temperature and temperature-dependent models were compared with data reported by Borhani et al. [14], which contained self-solvation energies at 298.15 K. For compounds common to both datasets, the comparison of target values yielded a MAE of 0.16 kcal mol^{-1} and an ARD of 2.1 %, demonstrating a high level of agreement between their data with our extended database, as can be seen in Fig. 3. However, a specific point seemed to present high deviation, which was further identified as triethanolamine. Additional investigation of its temperature-dependent properties revealed that both liquid density and vapor pressure were in good agreement with the external NIST database values, suggesting that the source of this discrepancy is unlikely to be related with the database provided in this work.

Regarding predictions, the temperature-dependent model outperformed the single-temperature model, with lower MAE (0.18 vs 0.50 kcal mol^{-1}) and ARD (2.6 % vs 8.4 %), which help further underscore the superior performance and enhanced flexibility of the temperature-dependent model, which is the focus of this work. Since all the following results of this work only pertain to temperature-dependent data, a summarized version of all the results at 298.15 K is presented in Table 3.

Additionally, a sensitivity analysis of +/- 5 % on the temperature-dependent properties in this temperature-dependent model revealed that liquid density and vapor pressure were the solvation properties with the most influence. Moreover, it was observed that the uncertainty of these predictions increased as the system approached the critical point, as shown in the **Supporting Information**. To facilitate reproducibility, the critical temperatures of all compounds included in the database are also provided in the **Supporting Information** for Readers interested in generating the plots themselves.

Considering this, it was proposed evaluating the quality of the

Table 1
Summary of key statistical metrics for Model Type performance.

Model Type	R^2	MAE (kcal mol^{-1})	ARD (%)
Single-Temperature	0.812	0.87	10.1
Temperature-Dependent	0.992	0.09	2.2

Table 2
Statistical metrics for the different Datasets of the temperature-dependent model.

Dataset	R^2	MAE (kcal mol^{-1})
Training	0.995	0.09
Validation	0.992	0.10
Testing	0.993	0.10
Overall	0.992	0.09

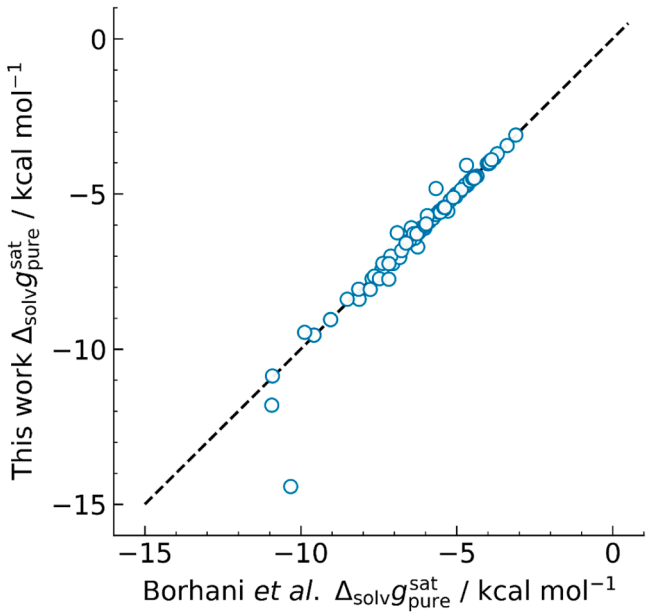

Fig. 3. Fitness plot for $\Delta_{solv}g^{sat}_{pure}$ experimental data at 298.15 K between Borhani et al. [14] and this work.

temperature-dependent data by ranking the accuracy and type of data for liquid density and vapor pressure, as these were identified as the most critical properties. Since the DIPPR and Yaws databases classify their data using different methods, a unified ranking system was applied to both. This system involved assigning a grade from 0 to 2 to each property and then summing these grades to obtain an overall data rank. The lower the grade, the better the overall quality of the data. For the DIPPR database, accuracy is specified by a percentage, so, data was categorized as follows: if the accuracy is below 2 %, the rank is 0, if it is between 2 % and 5 %, the rank is 1, and, if it exceeds 5 %, the rank is 2. For the Yaws database, data is typically classified as 1, 2, or 3, corresponding to correlations based on experimental data, predictions, and rough predictions, respectively. To align with the DIPPR ranking framework, these values were adjusted by subtracting one unit, resulting in a range from 0 to 2. The global ranking values obtained for each compound can be found in the **Supporting Information**. As mentioned above, the predictions obtained using this model generally showed good agreement with the actual values in the database, regardless of whether the data had the lowest or highest rank, indicating that the model's predictive capability extends beyond a mere re-parametrization of pre-existing calculation schemes. This underscores its potential for application even in datasets with varying levels of experimental validation. However, as shown in Fig. 2, there are some data points with higher deviations, which occur independently of whether the data used for calculation was based only experimental or on predicted data. Additionally, these points were analyzed based on their functional groups, as detailed in the **Supporting Information**. The analysis shows that the average deviation values for different categories of compounds are relatively consistent, indicating that the model does not exhibit bias towards any specific chemical group. Moreover, it is also possible to observe that no correlation is present between the prediction errors and the number of data points within each chemical group. A more detailed discussion follows in the subsequent paragraphs regarding these prediction errors.

Table 3
Performance comparison of predictions at 298.15 K with different experimental data.

Model Type	R^2	MAE (kcal mol^{-1})	ARD (%)
Compared to this work self-solvation energy data			
Single-Temperature	0.812	0.87	10.1
Temperature-Dependent	0.768	0.75	7.9
Compared to the self-solvation energy data by Borhani et al. [14]			
Single-Temperature	0.828	0.50	8.4
Temperature-Dependent	0.941	0.18	2.6

In fact, it can be seen that there are some instances where points fall outside of the confidence band, indicating potential shortcomings of the model. These outliers can be categorized into two cases: 1) a small cluster of points around 0 kcal mol^{-1} for the experimental self-solvation energy, and 2) an aggregation of points near the outer edges of the band. This prompted us to investigate the nature of these points further, leading us to conclude that they corresponded to small compounds, such as hydrogen and deuterium, and low-numbered ring compounds, respectively. Details regarding both cases will be given in the following paragraphs.

For the first case, we evaluated the effect of size differences between compounds on their prediction relative deviations from the true values. The simplest and most straightforward way to correlate different compounds with their sizes is by using their molar mass M_M, which can be easily obtained. The results are shown in Fig. 4.

Looking at the figure (left panel), there seems to exist large deviations for very small compounds (molar mass up until 90 g mol^{-1}). However, when looking at the remnant of the graph, no correlation seems to exist between the performance of the network and the size of the compound. Following this, looking at the zoomed-in version of the first bar of the left panel (which corresponds to the right panel), it clearly illustrates that the significant errors are primarily due to hydrogen, helium, and their corresponding isotopes. These compounds exhibit deviations exceeding 4000 %, while other small compounds do not reach even a 10 % deviation. This stark contrast indicates that the problem lies specifically with hydrogen, helium, and their isotopes, rather than with the size of the compounds overall. Consequently, this suggests that the inadequacies of the model are not inherently due to the small size of these compounds but are more likely attributed to unique properties of these very specific compounds that the model fails to accurately capture. In fact, when removing these compounds, the ARD decreases from 2.2 to 1.8 % and the R^2 increases from 0.992 to 0.994, indicating a small increase of the model performance, which was to be expected since despite these points carrying out high deviations, they only comprise a very small fraction of the total database (<0.01 %).

Another key detail found in Fig. 4 is the existence of a higher deviation peak for compounds between 540 and 630 g mol^{-1} when compared to the rest of the database. Upon further inspection, this deviation is mostly due to one-ringed compounds, motivating a deeper study regarding the model performance toward the number of rings present in each compound. The results are presented in Fig. 5.

Most deviations from the bisector are primarily found in compounds with no rings or a single ring. In fact, compounds with two, three, four, seven, ten, and thirteen rings display an almost linear trend in their true versus predicted values, indicating good model performance for describing these compounds. For zero and one-ring compounds, some predictions are near the bisector, while others show larger deviations. Although these deviations are significant, they are still smaller than those observed for hydrogen and helium. Nonetheless, these reflect the underperformance of the model in these very specific cases. Notably, while both no-ring and one-ring compounds exhibit deviations, it is primarily the one-ring compounds that fall outside the confidence band of 2.5 kcal mol^{-1}. This suggests that the model faces more pronounced challenges with one-ring compounds. Interestingly enough, the success of the model in predicting no-ring and high-ring compounds with a high accuracy degree would imply that there is no inherent reason it should not predict single-ring compounds effectively. The deviations for single-ring compounds do not appear to stem from unique theoretical challenges, suggesting that these errors might be due to a specific aspect of the model or data.

To address the observed deviations, the number of rings was

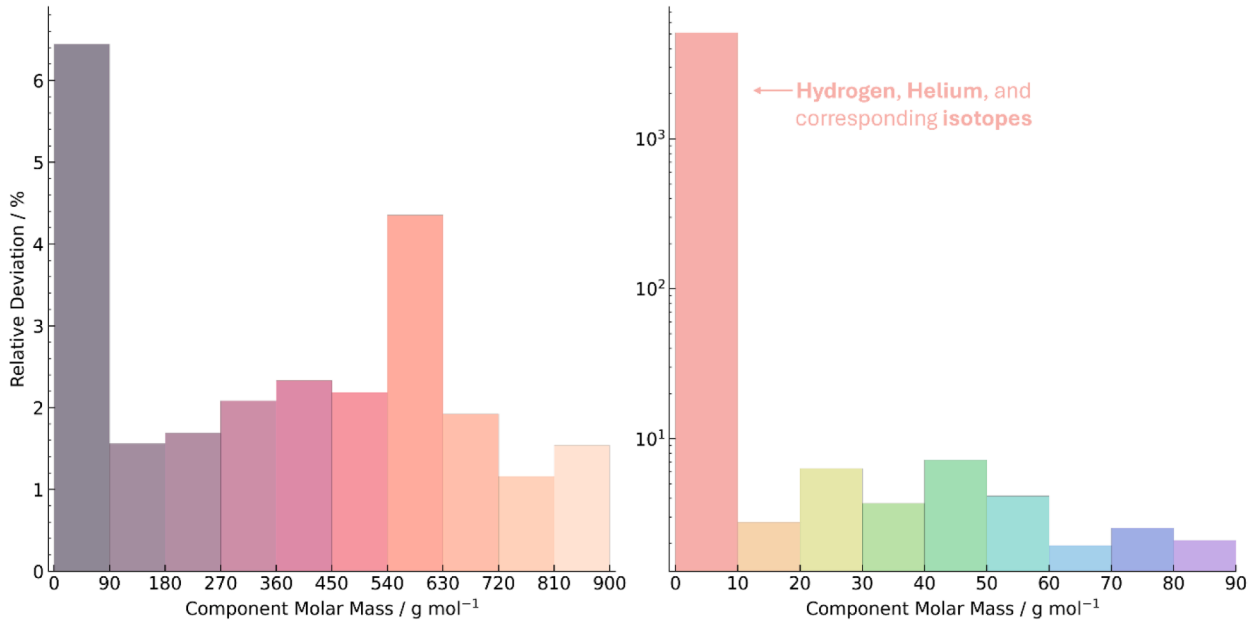

Fig. 4. Relative deviation plots of the Chemprop model regarding the molar mass of the compounds present in the database. On the left panel these deviations correspond to all database, while on the right panel these correspond to compounds up until 90 g mol^{-1}.

incorporated as an additional feature in the model. This approach aimed to better understand and improve the performance for one-ring compounds, while also benchmarking the capability of the model across different ring structures. Given that Chemprop already uses molecular descriptors likely correlated with ring structures, it was anticipated that only marginal improvements would occur.

As expected, the inclusion of the ring feature led to a slight improvement in model performance, with the R^2 increasing to 0.993 and the ARD reducing slightly to 2.1 %. However, the MAE remained unchanged at 0.09 kcal mol^{-1}, indicating that the enhancements were minimal. These results suggest that while the ring feature adds some value, its impact is limited due to existing descriptors already capturing relevant information. Full predictions and detailed statistics for this model can be found in the **Supporting Information**.

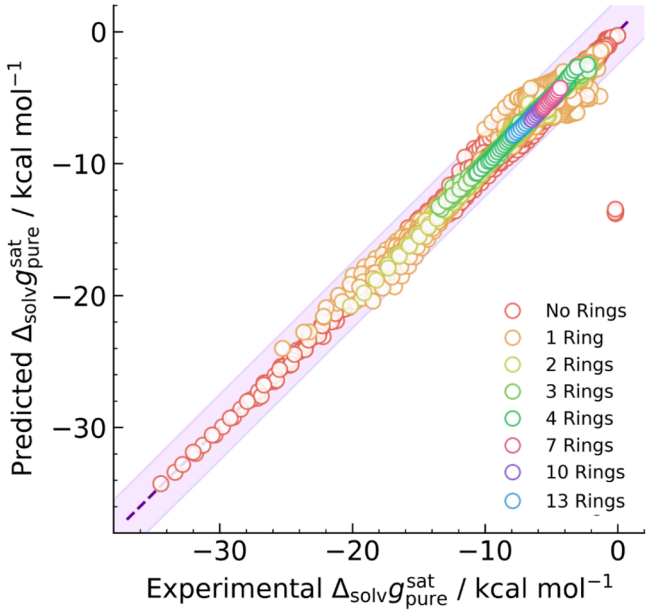

Fig. 5. Fitness plot for temperature-dependent $\Delta_{solv}g^{sat}_{pure}$ prediction with ring-number distribution using the Chemprop model. Number of rings that are not presented in the plot are not contained on the database. The purple band corresponds to a deviation of 2.5 kcal mol^{-1}.

Since adding ring features yielded only marginal improvements, the model was further tested by splitting the data into two subsets: one for compounds with no rings and another for compounds with rings, each trained with its own model. This step served mainly as an exercise to evaluate the flexibility of the model, as the initial one had already shown excellent overall performance. Predictions and statistical metrics for both models can be found in the **Supporting Information**.

The results show that the same issues observed in the main model (Fig. 2) also persist in these smaller models. For the no-ring compounds, deviations continue to be significant for small compounds like hydrogen and helium. Removing these compounds improves the ARD of the model, while increasing R^2. However, despite these changes, the ARD and MAE of this no-ring model are only slightly better than the main model, whereas these improvements come at the expense of higher deviations in predicted values, making this model more precise but less accurate. Moreover, some no-ring compounds still show high deviations near the outer confidence bands. On the other hand, the model for ring compounds exhibits a slightly improved ARD, though overall metrics worsen slightly, indicating higher accuracy but lower precision.

Further analysis comparing both smaller models to the main model reveals no major improvements. For no-ring compounds, the main model performs similarly to the no-ring model, and for ring compounds, the performance of the main model is slightly better than that of the new model. This highlights that using a larger, more diverse dataset enhances the generalization capabilities of the network across various molecular structures. Thus, the main model, which encompasses all compound types, offers better universality and flexibility compared to the two reduced models. Additionally, all statistical metrics pertaining to the temperature-dependent model and derived models based on ring structures are presented in Table 4 for clarity purposes.

As a final evaluation, the model was tested under two distinct conditions: one where it was expected to perform worse and another where it was expected to perform better. In the first case, non-scaled features were used, resulting in poorer performance across all metrics due to the difficulty of training with inputs of varying magnitudes. In the second case, a transformation was applied using the inverse of temperature and the logarithm of the self-solvation energy to test for an Arrhenius-like behavior. While this reduced the ARD slightly, the R^2 remained the

Table 4
Performance comparison of the temperature-dependent model and corresponding ring-derived models.

Model Type	R^2	MAE (kcal mol^{-1})	ARD (%)
Temperature-Dependent	0.992	0.09	2.2
Additional Ring Features	0.993	0.09	2.1
No-Ring Compounds	0.993	0.08	2.3
Only Ring Compounds	0.988	0.11	2.1

Table 5
Performance comparison of temperature-dependent models under different conditions.

Model Type	R^2	MAE (kcal mol^{-1})	ARD (%)
Temperature-Dependent	0.992	0.09	2.2
Non-scaled Features	0.985	0.15	3.2
Arrhenius-like Behavior	0.992	0.10	1.9

same, and the MAE increased slightly, indicating that the model captured relative trends better but introduced minor absolute errors. These results, along with full predictions, are available in the **Supporting Information**, with the statistical metrics summary being presented in Table 5. Regarding such results, it is evident that neither test provided enough reason to replace the standard model, which remains the most reliable approach.

In summary, this model, that takes advantage of the Chemprop tool, served as a baseline to evaluate various properties such as compound size and number of ring structures, enabling a deeper study into the addition of extra features and the use of smaller subsets of the original database. Moreover, it facilitated benchmarking in terms of model specifics and input data transformations. All of this provided valuable information to select the best model for predicting self-solvation energies across diverse molecular structures over a wide temperature range with uncertainties always within the expected thresholds observed in experimental data reported in literature.

4. Conclusions

For this study, the DIPPR and Yaws databases provided thermodynamic information necessary to obtain saturation properties for pure compounds that allowed us to obtained temperature-dependent self-solvation data. The resulting dataset, encompassing 5420 pure compounds and 71,656 data points, was utilized to train a Machine-Learning model using the Chemprop package. This graph convolutional neural network model incorporated these temperature dependencies as additional features, aiming to predict self-solvation energies over a much broader range of temperature than are generally used for a wide spectrum of different compounds.

Initially, self-solvation energies were calculated at a single temperature of 298.15 K for the compounds in our database to compare them with temperature-dependent results. These results allowed us to conclude that Chemprop is less effective with small datasets, but its performance improves with larger, temperature-dependent datasets. Afterwards, this study encompassed a much larger database that included all temperature-dependent data. This resulted in a model with ARD of 2.2 %, R^2 of 0.992, and MAE of 0.09 kcal mol^{-1}, which confirmed the hypothesis that a larger dataset enhances the generalization ability of Chemprop. Further, the consistency of the model's predictions across all data quality ranks - including pseudo-experimental values - highlights its potential as a reliable tool, even when applied to datasets with varying levels of experimental validation.

Most predictions fell within a 2.5 kcal mol^{-1} deviation band, indicating small absolute deviations relative to the self-solvation energy range. However, some outliers were noted, primarily small compounds, like hydrogen and helium, and single-numbered ring compounds. For the first case, removing these outliers improved the ARD and R^2 values slightly, indicating that these specific compounds posed unique challenges for the model, rather than it being a limitation of the model that is linked to the size of the compounds. For the second case, the number of rings were incorporated as an additional feature in the model. It was found that this addition yielded only marginal improvements in the overall statistical metrics, which aligned with expectations, given the use of molecular descriptors by Chemprop, which may already consider ring structures.

This prompted additional studies that involved dividing the data into two separate datasets - one for compounds with no ring structures and another for compounds with only ring structures – which provided additional information regarding the flexibility of the model. The results for these models reveal that the issues present in the main model persist in these smaller datasets. For no-ring compounds, deviations for small components remain significant. For ring compounds, overall metrics worsen, showing more accuracy but less precision. Further in-depth analysis reveals that the larger dataset enhances the generalization capability of the network across diverse molecular structures, allowing different compound types to contribute to explaining a broader set of data. Additionally, the main temperature-dependent model, which includes all types of data, functions more effectively as a universal model than the reduced models.

To further assess the robustness and flexibility of our Chemprop-based model, additional benchmarking was conducted. This included testing the model in two specific scenarios where performance was expected to vary: using non-scaled features and applying transformations to the temperature and self-solvation energy variables. These tests aimed to explore how the preprocessing of input features affects the predictive capability of the model. Regarding the feature scaling, it was confirmed that it is crucial for better results. Without scaling, the network struggled to learn effectively due to the varying magnitudes of the input variables. The larger prediction errors and reduced variance explanation underscore the importance of preprocessing steps like feature scaling in achieving high model performance. Regarding transformation of the input variables, these were shown to be able to improve relative prediction accuracy, potentially by aligning the data more closely with linear trends. However, this may come at the cost of slightly increased absolute prediction errors.

Overall, while our main model remained robust and effective, all the additional experiments and benchmarks provided insights into how different preprocessing techniques can influence model performance. These highlight the trade-offs between absolute and relative accuracy, and the necessity of tailored preprocessing to enhance model training and prediction capabilities, which are of extreme importance in self-solvation models, often used in the field of chemistry, biochemistry, material science, and environmental management.

CRediT authorship contribution statement

Hugo Marques: Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation. **Simon Müller:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the financial support from Fundação para a Ciência e Tecnologia, FCT/MCTES (Portugal) through the projects UIDB/00100/2020, UIDP/00100/2020, and IMS-LA/P/0056/2020,

through PhD grant 2022.10217.BD (H. M.).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fluid.2025.114335.

Data availability

All original data from YAWS and DIPPR are available from the cited sources. All derived data is available in the SI.

References

[1] E. Moine, R. Privat, B. Sirjean, J.-N. Jaubert, Estimation of solvation quantities from experimental thermodynamic data: development of the comprehensive CompSol databank for pure and mixed solutes, J. Phys. Chem. Ref. Data 46 (2017) 033102, https://doi.org/10.1063/1.5000910.

[2] P. Muller, Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994), Pure Appl. Chem. 66 (1994) 1077–1184, https://doi.org/10.1351/pac199466051077.

[3] V.I. Minkin, Glossary of terms used in theoretical organic chemistry, Pure Appl. Chem. 71 (1999) 1919–1981, https://doi.org/10.1351/pac199971101919.

[4] A. Ben-Naim, Standard thermodynamics of transfer. Uses and misuses, J. Phys. Chem. 82 (1978) 792–803, https://doi.org/10.1021/j100496a008.

[5] J.D. Thompson, C.J. Cramer, D.G. Truhlar, Predicting aqueous solubilities from aqueous free energies of solvation and experimental or calculated vapor pressures of pure substances, J. Chem. Phys. 119 (2003) 1661–1670, https://doi.org/10.1063/1.1579474.

[6] M. Sixt, I. Koudous, J. Strube, Process design for integration of extraction, purification and formulation with alternative solvent concepts, Comptes Rendus Chim. 19 (2016) 733–748, https://doi.org/10.1016/j.crci.2015.12.016.

[7] P. Ruelle, The n-octanol and n-hexane/water partition coefficient of environmentally relevant chemicals predicted from the mobile order and disorder (MOD) thermodynamics, Chemosphere 40 (2000) 457–512, https://doi.org/10.1016/S0045-6535(99)00268-4.

[8] Y. Chung, F.H. Vermeire, H. Wu, P.J. Walker, M.H. Abraham, W.H. Green, Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy, J. Chem. Inf. Model. 62 (2022) 433–446, https://doi.org/10.1021/acs.jcim.1c01103.

[9] L. Ferreira, R. Dos Santos, G. Oliva, A. Andricopulo, Molecular docking and structure-based drug design strategies, Molecules 20 (2015) 13384–13421, https://doi.org/10.3390/molecules200713384.

[10] N. Matubayasi, Solvation energetics of proteins and their aggregates analyzed by all-atom molecular dynamics simulations and the energy-representation theory of solvation, Chem. Commun. 57 (2021) 9968–9978, https://doi.org/10.1039/D1CC03395F.

[11] S. Grinter, X. Zou, Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design, Molecules 19 (2014) 10150–10176, https://doi.org/10.3390/molecules190710150.

[12] Y. Chung, R.J. Gillis, W.H. Green, Temperature-dependent vapor–liquid equilibria and solvation free energy estimation from minimal data, AIChE J. 66 (2020) e16976, https://doi.org/10.1002/aic.16976.

[13] A. Alibakhshi, B. Hartke, Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model, Nat. Commun. 12 (2021) 3584, https://doi.org/10.1038/s41467-021-23724-6.

[14] T.N. Borhani, S. García-Muñoz, C. Vanesa Luciani, A. Galindo, C.S. Adjiman, Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs, Phys. Chem. Chem. Phys. 21 (2019) 13706–13720, https://doi.org/10.1039/C8CP07562J.

[15] H. Choi, H. Kang, H. Park, New solvation free energy function comprising intermolecular solvation and intramolecular self-solvation terms, J. Cheminform. 5 (2013) 8, https://doi.org/10.1186/1758-2946-5-8.

[16] J.H. Hildebrand, A history of solution theory, Annu. Rev. Phys. Chem. 32 (1981) 1–24, https://doi.org/10.1146/annurev.pc.32.100181.000245.

[17] C.J. Cramer, D.G. Truhlar, Molecular orbital theory calculations of aqueous solvation effects on chemical equilibria, J. Am. Chem. Soc. 113 (1991) 8552–8554, https://doi.org/10.1021/ja00022a069.

[18] A.V. Marenich, R.M. Olson, C.P. Kelly, C.J. Cramer, D.G. Truhlar, Self-Consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges, J. Chem. Theory Comput. 3 (2007) 2011–2033, https://doi.org/10.1021/ct7001418.

[19] A.V. Marenich, C.J. Cramer, D.G. Truhlar, 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 113 (2009) 6378–6396, https://doi.org/10.1021/jp810292n.

[20] A. Klamt, Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena, J. Phys. Chem. 99 (1995) 2224–2235, https://doi.org/10.1021/j100007a062.

[21] A. Klamt, V. Jonas, T. Bürger, J.C.W. Lohrenz, Refinement and parametrization of COSMO-RS, J. Phys. Chem. A 102 (1998) 5074–5085, https://doi.org/10.1021/jp980017s.

[22] A. Klamt, F. Eckert, COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, Fluid Phase Equilib. 172 (2000) 43–72, https://doi.org/10.1016/S0378-3812(00)00357-5.

[23] A. Klamt, The COSMO and COSMO-RS solvation models, WIREs Comput. Mol. Sci. 1 (2011) 699–709, https://doi.org/10.1002/wcms.56.

[24] S. Müller, T. Nevolianis, M. Garcia-Ratés, C. Riplinger, K. Leonhard, I. Smirnova, Predicting solvation free energies for neutral molecules in any solvent with openCOSMO-RS, (2024). https://doi.org/10.48550/ARXIV.2407.03434.

[25] T. Gerlach, S. Müller, A.G. De Castilla, I. Smirnova, An open source COSMO-RS implementation and parameterization supporting the efficient implementation of multiple segment descriptors, Fluid Phase Equilib. 560 (2022) 113472, https://doi.org/10.1016/j.fluid.2022.113472.

[26] D. Grigorash, S. Müller, P. Paricaud, E.H. Stenby, I. Smirnova, W. Yan, A comprehensive approach to incorporating intermolecular dispersion into the openCOSMO-RS model. Part 1: Halocarbons, (2024). https://doi.org/10.48550/ARXIV.2406.05244.

[27] L.C. Kröger, S. Müller, I. Smirnova, K. Leonhard, Prediction of solvation free energies of ionic solutes in neutral solvents, J. Phys. Chem. A 124 (2020) 4171–4181, https://doi.org/10.1021/acs.jpca.0c01606.

[28] M.H. Abraham, W.E. Acree Jr., Correlation and prediction of partition coefficients between the gas phase and water, and the solvents dodecane and undecane, New J. Chem. 28 (2004) 1538, https://doi.org/10.1039/b411303a.

[29] M.H. Abraham, A. Ibrahim, A.M. Zissimos, Determination of sets of solute descriptors from chromatographic measurements, J. Chromatogr. A 1037 (2004) 29–47, https://doi.org/10.1016/j.chroma.2003.12.004.

[30] C. Mintz, M. Clark, W.E. Acree, M.H. Abraham, Enthalpy of solvation correlations for gaseous solutes dissolved in water and in 1-octanol based on the Abraham model, J. Chem. Inf. Model. 47 (2007) 115–121, https://doi.org/10.1021/ci600402n.

[31] A. Jalan, R.W. Ashcraft, R.H. West, W.H. Green, Predicting solvation energies for kinetic modeling, Annu. Rep. Prog. Chem., Sect. C: Phys. Chem. 106 (2010) 211, https://doi.org/10.1039/b811056p.

[32] P. Havelec, J.G.K. Ševčík, Extended additivity model of parameter log(*L* 16), J. Phys. Chem. Ref. Data 25 (1996) 1483–1493, https://doi.org/10.1063/1.555989.

[33] D. Svozil, J.G.K. Ševčík, V. Kvasnička, Neural network prediction of the solvatochromic polarity/polarizability parameter, J. Chem. Inf. Comput. Sci. 37 (1997) 338–342, https://doi.org/10.1021/ci960347e.

[34] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, Estimation of molecular linear free energy relation descriptors using a group contribution approach, J. Chem. Inf. Comput. Sci. 39 (1999) 835–845, https://doi.org/10.1021/ci980339t.

[35] J.A. Platts, M.H. Abraham, D. Butina, A. Hersey, Estimation of molecular linear free energy relationship descriptors by a group contribution approach. 2. Prediction of partition coefficients, J. Chem. Inf. Comput. Sci. 40 (2000) 71–80, https://doi.org/10.1021/ci990427t.

[36] T. Ghafourian, J.C. Dearden, The use of atomic charges and orbital energies as hydrogen-bonding-donor parameters for QSAR studies: comparison of MNDO, AM1 and PM3 methods, J. Pharm. Pharmacol. 52 (2010) 603–610, https://doi.org/10.1211/0022357001774435.

[37] A.M. Zissimos, M.H. Abraham, A. Klamt, F. Eckert, J. Wood, A Comparison between the two general sets of linear free energy descriptors of Abraham and Klamt, J. Chem. Inf. Comput. Sci. 42 (2002) 1320–1331, https://doi.org/10.1021/ci025530o.

[38] J.S. Arey, W.H. Green, P.M. Gschwend, The electrostatic origin of Abraham's solute polarity parameter, J. Phys. Chem. B 109 (2005) 7564–7573, https://doi.org/10.1021/jp044525f.

[39] Y. Liang, T.L. Torralba-Sanchez, D.M. Di Toro, Estimating system parameters for solvent–water and plant cuticle–water using quantum chemically estimated Abraham solute parameters, Environ. Sci. Processes Impacts 20 (2018) 813–821, https://doi.org/10.1039/C7EM00601B.

[40] A. Stenzel, K. Goss, S. Endo, Prediction of partition coefficients for complex environmental contaminants: validation of COSMOtherm, ABSOLV, and SPARC, Enviro. Toxic. Chem. 33 (2014) 1537–1543, https://doi.org/10.1002/etc.2587.

[41] T.N. Brown, J.A. Arnot, F. Wania, Iterative fragment selection: a group contribution approach to predicting fish biotransformation half-lives, Environ. Sci. Technol. 46 (2012) 8253–8260, https://doi.org/10.1021/es301182a.

[42] T.N. Brown, Predicting hexadecane–air equilibrium partition coefficients (*L*) using a group contribution approach constructed from high quality data, SAR QSAR Environ. Res. 25 (2014) 51–71, https://doi.org/10.1080/1062936X.2013.841286.

[43] L.-H. Wang, C.-M. Hsieh, S.-T. Lin, Improved prediction of vapor pressure for pure liquids and solids from the PR+COSMOSAC equation of state, Ind. Eng. Chem. Res. 54 (2015) 10115–10125, https://doi.org/10.1021/acs.iecr.5b01750.

[44] C. Tsai, S. Lin, Improved vapor pressure prediction from PR + COSMOSAC EOS using normal boiling temperature, AIChE J. 69 (2023) e17997, https://doi.org/10.1002/aic.17997.

[45] A.V. Marenich, C.P. Kelly, J.D. Thompson, G.D. Hawkins, C.C. Chambers, D.J. Giesen, P. Winget, C.J. Cramer, D.G. Truhlar, Minnesota Solvation Database (MNSOL) version 2012, (2020). https://doi.org/10.13020/3EKS-J059.

[46] D.L. Mobley, J.P. Guthrie, FreeSolv: a database of experimental and calculated hydration free energies, with input files, J. Comput. Aided Mol. Des. 28 (2014) 711–720, https://doi.org/10.1007/s10822-014-9747-x.

[47] S.T. Hutchinson, R. Kobayashi, Solvent-specific featurization for predicting free energies of solvation through machine learning, J. Chem. Inf. Model. 59 (2019) 1338–1346, https://doi.org/10.1021/acs.jcim.8b00901.

[48] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, (2017). https://doi.org/10.48550/ARXIV.1703.00564.
[49] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (2019) 3370–3388, https://doi.org/10.1021/acs.jcim.9b00237.
[50] C. Hille, S. Ringe, M. Deimel, C. Kunkel, W.E. Acree, K. Reuter, H. Oberhofer, Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme, J. Chem. Phys. 150 (2019) 041710, https://doi.org/10.1063/1.5050938.
[51] Y. Pathak, S. Mehta, U.D. Priyakumar, Learning atomic interactions through solvation free energy prediction using graph neural networks, J. Chem. Inf. Model. 61 (2021) 689–698, https://doi.org/10.1021/acs.jcim.0c01413.
[52] H. Lim, Y. Jung, MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning, J. Cheminform. 13 (2021) 56, https://doi.org/10.1186/s13321-021-00533-z.
[53] T.E. Daubert, H.M. Sibul, C.C. Stebbins, R.P. Danner, T.L. Marshall, R.L. Rowley, M.E. Adams, W.V. Wilding, Physical and Thermodynamic Properties of Pure Chemicals: DIPPR, Data Compilation: Core + Supplements 1-10, Taylor & Francis, 2000, p. 2000.
[54] C.L. Yaws, Yaws' Critical Property Data for Chemical Engineers and Chemists, Knovel, 2012.
[55] T. Gerlach, S. Müller, I. Smirnova, Development of a COSMO-RS based model for the calculation of phase equilibria in electrolyte systems, AIChE J. 64 (2018) 272–285, https://doi.org/10.1002/aic.15875.
[56] S. Müller, A. González De Castilla, C. Taeschler, A. Klein, I. Smirnova, Evaluation and refinement of the novel predictive electrolyte model COSMO-RS-ES based on solid-liquid equilibria of salts and Gibbs free energies of transfer of ions, Fluid Phase Equilib. 483 (2019) 165–174, https://doi.org/10.1016/j.fluid.2018.10.023.
[57] S. Müller, A. González De Castilla, C. Taeschler, A. Klein, I. Smirnova, Calculation of thermodynamic equilibria with the predictive electrolyte model COSMO-RS-ES: improvements for low permittivity systems, Fluid Phase Equilib. 506 (2020) 112368, https://doi.org/10.1016/j.fluid.2019.112368.
[58] A. González De Castilla, S. Müller, I. Smirnova, On the analogy between the restricted primitive model and capacitor circuits: semi-empirical alternatives for over- and underscreening in the calculation of mean ionic activity coefficients, J. Mol. Liq. 326 (2021) 115204, https://doi.org/10.1016/j.molliq.2020.115204.
[59] A. González De Castilla, S. Müller, I. Smirnova, On the analogy between the restricted primitive model and capacitor circuits. Part II: a generalized Gibbs-Duhem consistent extension of the Pitzer-Debye-Hückel term with corrections for low and variable relative permittivity, J. Mol. Liq. 360 (2022) 119398, https://doi.org/10.1016/j.molliq.2022.119398.
[60] M. Arrad, K. Thomsen, S. Müller, I. Smirnova, Thermodynamic modeling using Extended UNIQUAC and COSMO-RS-ES models: case study of the cesium nitrate - water system over a large range of temperatures, Fluid Phase Equilib. (2024) 114037, https://doi.org/10.1016/j.fluid.2024.114037.
[61] O.A. Hougen, K.M. Watson, R.A. Ragatz, Chemical Process Principles : Part II - Thermodynamics : Second Edition, John Wiley & Sons, 1964.
[62] E. Heid, K.P. Greenman, Y. Chung, S.-C. Li, D.E. Graff, F.H. Vermeire, H. Wu, W. H. Green, C.J. McGill, Chemprop: a machine learning package for chemical property prediction, J. Chem. Inf. Model. 64 (2024) 9–17, https://doi.org/10.1021/acs.jcim.3c01250.