Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures

Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures

Florence H. Vermeire, Yunsie Chung, and William H. Green*

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, United States

E-mail: whgreen@mit.edu

Abstract

The solubility of organic molecules is crucial in organic synthesis and industrial chemistry, it is important in the design of many phase separation and purification units, and it controls the migration of many species into the environment. To decide which solvents and temperatures can be used in the design of new processes, trial and error is often used, as the choice is restricted by unknown solid solubility limits. Here we present a fast and convenient computational method for estimating the solubility of solid neutral organic molecules in water and many organic solvents for a broad range of temperatures. The model is developed by combining fundamental thermodynamic equations with machine learning models for solvation free energy, solvation enthalpy, Abraham solute parameters, and aqueous solid solubility at 298K. We provide free open-source and online tools for the prediction of solid solubility limits and a curated data collection (SolProp) that includes more than 5,000 experimental solid solubility values for validation of the model. The model predictions are accurate for aqueous systems and for a huge range of organic solvents up to 550K or higher. Methods to further improve solid solubility predictions by providing experimental data on the

solute of interest in another solvent, or on the solute's sublimation enthalpy, are also presented.

Introduction

The solubility of solids in organic solvents plays a huge role in various chemical systems such as organic chemistry, environmental chemistry, organo(photo)catalysis, and (petro)chemical industry. Solid-solution equilibria are, for example, used in the design of flow batteries, purification units, extraction units, and liquid phase chemical processes. Specific to the pharmaceutical industry, the solubility of active pharmaceutical ingredients (API) in a variety of organic solvents is an important property in the development of new drugs. Many curated databases (*e.g.* the AqSol database[1]) are available for the aqueous solid solubility of API's as this property is of interest in the initial screening process of potential drugs. Further in the drug development chain during lab-scale synthesis, purification, crystallization, and scale-up from batch to continuous processes, information on the solubility of the API's in organic solvents other than water is required.

Many methods exist for the prediction of the solubility of solids in specific organic solvents. A recently published review by Kuentz et al.[2] summarizes recent advances in solubility predictions. The main limitation of existing methods is that they require physical properties of the solute, such as the fusion enthalpy and/or melting temperature (*e.g.* the well-known NRTL method[3]) or empirical solute parameters (*e.g.* in the relationships developed by Abraham et al.[4]). Other methods for the direct prediction of solid solubility are always limited to one or a few solvents. For example, the aqueous solid solubility is important in many applications, including the initial screening of potential API's, and thus there are enough data available to train deep neural networks for the prediction of this property. Some examples of recent successful data-driven predictors for the aqueous solid solubility are AquaSol developed by Lusci et al.,[5] the models of Cui et al.[6] and by Boobier et al.,[7] and SolTranNet

developed by Francoeur and Koes.[8] Despite of the tremendous progress made in machine learning for molecular and material science,[9] for the direct prediction of solid solubility in organic solvents, the application of data driven methods is limited by data scarcity.[2] Many have applied machine learning for the prediction of properties related to the solubility of gaseous solutes in organic solvents and water.[10–14] Recently, one attempt has been made by Boobier et al.[7] to also use machine learning for the direct prediction of solid solubility in organic solvents using experimental data and computational descriptors. Machine learning models were built for the prediction of the solid solubility in water, ethanol, benzene, and acetone. For each of those solvents, a separate dataset was built and a different machine learning model was trained. While a good performance was achieved, data availability limits the application of this method to a narrow range of solvents and temperatures. Because solid solubility data at elevated temperatures are even more scarce, thermodynamic models such as NRTL, UNIFAC, and UNIQUAC are typically used to predict the temperature dependence of solubility;[15,16] nonetheless, these models are limited by the availability of empirical parameters. The physics based computation of solubility has significant potential as demonstrated by, for example, Fowles et al.[17] for the aqueous solid solubility. However, the large computational cost is a limiting factor for the fast screening of the solid solubility in different solvents.

In this work, we provide a method, software package, and online tool for the prediction of the solid solubility for moderate temperatures up to 350K without the use of any empirical parameters. Additionally, a more accurate method to predict the temperature dependence of solubility for a broader temperature range even approaching the critical point of the considered solvent is available for $\sim$100 solvents whose critical temperature and critical density are known. To compensate for the lack of available data, we use machine learning for the fast prediction of several properties, and thermodynamic equations to relate those properties to the solid solubility. The main advantage of this method is that only the molecule identifiers (SMILES or InChI) of the solute and solvent are required to make the predictions.

With additional user information on the solid solubility of the solute in one organic solvent at room temperature, the accuracy of the method can be improved. The main limitation of the model is that the predictions are limited to neutral solvents in the liquid phase and neutral solutes in the solid phase. The proposed method could be extended to the solubility of ions in organic solvents, for example for applications in nonaquous redox flow batteries.[18]

Many correlations and machine learning models are combined to calculate the solid solubility in organic solvents at different temperatures. An overview of the different models, and the inputs and outputs to the method are given in Figure 1. We use thermodynamics to relate the solid solubility of solutes in organic solvents to properties for which plenty of experimental data are available, *i.e.* (i) the aqueous solid solubility $\log(S_{aq})$, (ii) the gas-solution solvation free energy ΔG_{solv},[19–21] and (iii) the gas-solution solvation enthalpy ΔH_{solv}. Besides the solid solubility, the model and online tool also provide values for other thermodynamic properties that are used in the calculation of the solid solubility, for example thermodynamic properties related to the solubility of gases (solvation free energy and enthalpy, gas-solution equilibrium coefficients), related to the partitioning of solutes between solutions (dry partitioning coefficient), and related to solid state phase transitioning (sublimation enthalpy and heat capacity at 298K). All datasets that are constructed as part of this work are made publicly available in the SolProp data collection.

The Gibbs free energy required to dissolve a solute from the ideal gas phase into a solution, both at a standard state of 1 mol/L, is the solvation free energy ΔG_{solv}. Eq. 1 relates the solvation free energy in the molar standard state to the dimensionless gas-solution partitioning coefficient K and the molar concentrations of the solute in the gas phase c_{gas} and in solution $c_{solution}$. If the gas-solution partitioning coefficient is known for multiple solvents (K_{X_1} and K_{X_2}), the dry partitioning coefficient for the solute between those solvents (P_{dry,X_1-X_2}) and the relative solid solubility in two different solvents ($\frac{S_{X_1}}{S_{X_2}}$) can be calculated from Eq. 2. The solvation free energy at 298K in a variety of organic solvents is a property that can be predicted with a deep neural network because of data availability.

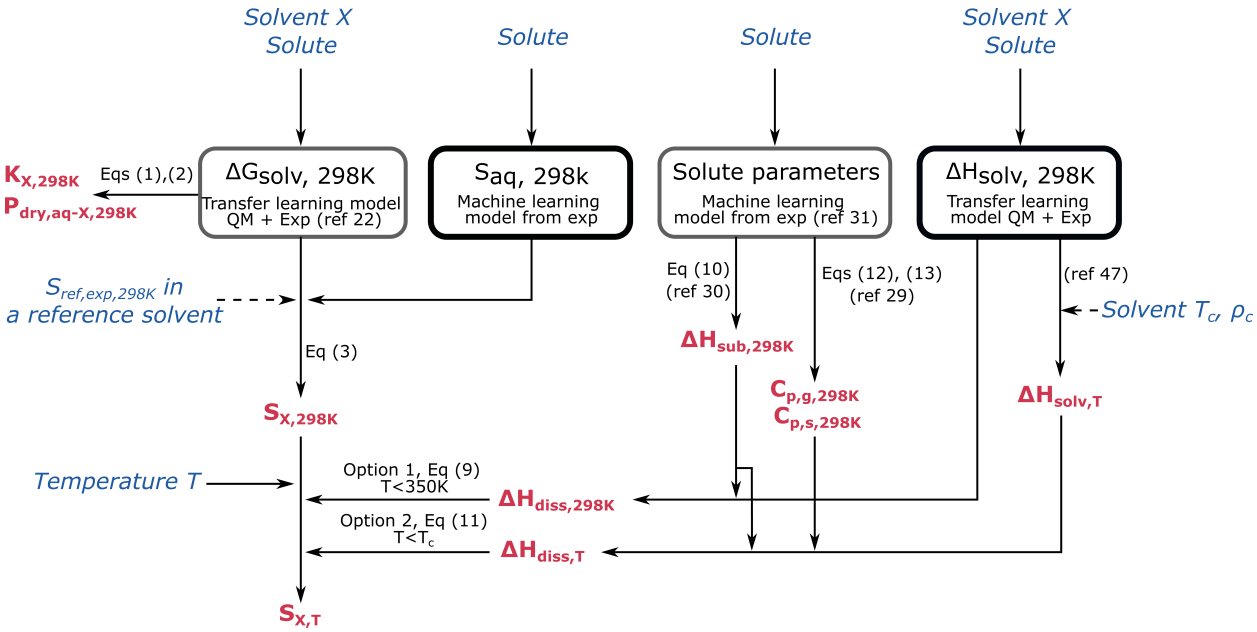

Figure 1: Overview of the models used to compute $S_{X,T}$, the solid solubility of a solute in solvent X at temperature T. Inputs are in blue italic text. Existing models are in grey boxes, new models are in black boxes. Outputs are in red bold text. If either the solid solubility of the solute in some other solvent at 298K or the critical density and temperature of the solvent are available, one can optionally use those inputs to compute $S_{X,T}$ more accurately.

Recently, we published two databases, CombiSolv-QM and CombiSolv-Exp, with 1 million quantum chemical (COSMO-RS) and 11 thousand experimental datapoints respectively for the solvation free energy in close to 300 solvents.[22] A transfer learning method was employed where the deep neural network was pre-trained on quantum chemical data and fine-tuned with experimental data. The final model can predict solvation free energies at 298K with a RMSE/MAE lower than 0.44/0.21 kcal/mol.[22] The transfer learning approach improves the performance on higher molar mass solutes compared to direct training of the deep neural network on experimental data. With the published transfer learning model, $\Delta G_{solv,298\mathrm{K}}$ (Eq. 1) can be predicted, and K_{X_1}, K_{X_2}, and P_{dry,X_1-X_2} (Eq. 2) can be calculated.

$$K = \exp\left(\frac{-\Delta G_{solv}}{RT}\right) = \frac{c_{solution}}{c_{gas}} \tag{1}$$

$$\log(\frac{K_{X_1}}{K_{X_2}}) = \log(P_{dry,X_1-X_2}) = \log(\frac{S_{X_1}}{S_{X_2}}) \tag{2}$$

To calculate the molar solid solubility of solutes, we relate the unknown solid solubility of the solute in one organic solvent to its known solid solubility in another solvent. The solid solubility in both solvents (S_{X_1} and S_{X_2}, in mol/L) are related to the dry partitioning coefficient of the solute between both solvents (P_{dry,X_1-X_2}) and hence to the gas-solution partitioning coefficients of the solute in both solvents (K_{X_1} and K_{X_2}) (see Eq. 2). With the transfer learning model, $\Delta G_{solv,298\text{K}}$ can be predicted. Thus, given the solubility in a reference solvent $\log(S_{ref,298K})$ at 298K, the solid solubility in a different solvent at 298K can be calculated using Eq. 3. In this work, we consider a reference solid solubility acquired from experimental measurements in organic solvents and from machine learning predictions of the aqueous solid solubility. Even though some models exist for the prediction of aqueous solid solubility, for consistency among the machine learning models in this work, a new and improved model for the prediction of aqueous solid solubility is constructed. The model is trained on the novel AqueousSolu-Exp database compiled in this work with 11804 unique components. This database is a curated collection of data from ALOGpS,[23] drugbank,[24] the DLS-100 dataset,[25] the AqSol dataset,[1] the PHYSPROP dataset,[26] and OChem.[27]

$$\begin{aligned}\log(S_{X,298K}) &= \log(S_{ref,298K}) + \log(K_{X,298K}) - \log(K_{ref,298K}) \\ &= \log(S_{ref,298K}) - \frac{(\Delta G_{X,solv,298K} - \Delta G_{ref,solv,298K})}{R \cdot 298K}\end{aligned} \tag{3}$$

Even though Eq. 2 is valid for a broad temperature range, the available machine learning models for the solvation free energies and aqueous solid solubility are restricted to 298K. To calculate the solid solubility at temperature T, the enthalpy required to dissolve a solute from the solid phase into the organic solvent at the specified temperature (*i.e.* the dissolution enthalpy, $\Delta H_{diss,T}$) is required. Those are related through the modified Van 't Hoff

equation[28] (Eq. 4), with R the ideal gas constant. Integration of Eq. 4 between 298K and the required temperature T yields Eq. 5.

$$\frac{d\ln(S_T)}{dT} = \frac{\Delta H_{diss,T}}{RT^2} \tag{4}$$

$$\ln(\frac{S_T}{S_{298\mathrm{K}}}) = \int_{298\mathrm{K}}^{T} \frac{\Delta H_{diss,T}}{RT^2}\, dT \tag{5}$$

With the temperature dependent dissolution enthalpy ($\Delta H_{diss,T}$) and the solid solubility at 298K ($S_{298\mathrm{K}}$), the solid solubility at another temperature (S_T) can be calculated. The determination of $\Delta H_{diss,T}$ is more challenging. Since no or limited experimental data are available for $\Delta H_{diss,T}$, we relate it to other properties through the thermodynamic cycle. Thermodynamics relate the enthalpy changes between the ideal gas, the solution, and the solid phase at 298K and temperature T as can be seen in Figure 2. Through this cycle, the dissolution enthalpy can be written as a summation (see Eq. 6) of (i) the enthalpy change associated with cooling the solid solute from temperature T to 298K ($\Delta H_{s,T\rightarrow 298\mathrm{K}}$), (ii) the sublimation enthalpy at 298K ($\Delta H_{sub,298\mathrm{K}}$), (iii) the enthalpy change associated with heating the gas phase solute from 298K to temperature T ($\Delta H_{g,298\mathrm{K}\rightarrow T}$), and (iv) the enthalpy required to dissolve the solute from the gas phase into the organic solvent (*i.e.* the solvation enthalpy) at temperature T ($\Delta H_{solv,T}$). Some of the enthalpy terms can further be replaced by the temperature difference and the heat capacity of the solute in the solid phase ($C_{p,s}$) and in the gas phase ($C_{p,g}$) (see Eq. 7). The temperature-dependence of the heat capacities is neglected for this work. We confirmed that this is a valid assumption, as explained further in the results and the supporting information (section S.13).

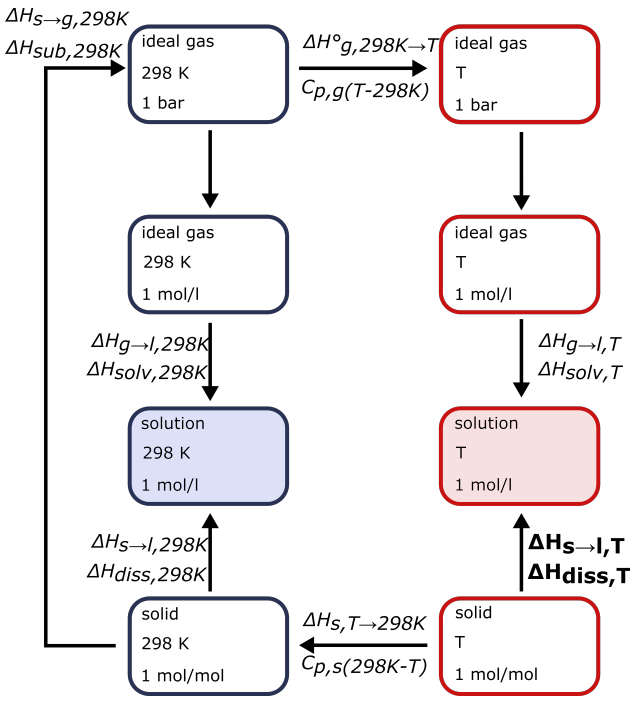

Figure 2: Thermodynamic cycle for enthalpy changes between the ideal gas phase, the solution, and the solid phase at room temperature (298 K, blue boxes) and a different temperature (T, red boxes), used to compute $\Delta H_{diss,T}$ (bold font). In the boxes, the phase, the temperature, and the reference state used to calculated the phase change are indicated. The shaded boxes indicate the solution, *i.e.* the presence of a solvent.

$$\begin{aligned}\Delta H_{diss,T} &= \Delta H_{s,T\to 298\mathrm{K}} + \Delta H_{sub,298\mathrm{K}} + \Delta H_{g,298\mathrm{K}\to T} + \Delta H_{solv,T} \\ &= \int_{T}^{298\mathrm{K}} C_{p,s}\, dT + \Delta H_{sub,298\mathrm{K}} + \int_{298\mathrm{K}}^{T} C_{p,g}\, dT + \Delta H_{solv,T}\end{aligned} \tag{6}$$

$$\Delta H_{diss,T} \approx C_{p,s}(298\mathrm{K} - T) + \Delta H_{sub,298\mathrm{K}} + C_{p,g}(T - 298\mathrm{K}) + \Delta H_{solv,T} \tag{7}$$

For the remainder of this work, we make a distinction between two different methods for the calculation of $\Delta H_{diss,T}$. Since the temperature dependence of $\Delta H_{solv,T}$ can currently only be calculated for some common solvents, our first method neglects the temperature dependence of the dissolution enthalpy. In the second method, the temperature dependence is accounted for through numerical integration. Other properties ($\Delta H_{sub,298\mathrm{K}}$, $C_{p,g}$, and $C_{p,s}$) can be calculated using recently published correlations by Abraham and Acree[29,30] combined

with machine learning predictions of solute parameters.[31]

Results and Discussion

The SolProp Data Collection

In order to assess the accuracy and robustness of the predictions of the new methods, and also to train some of the machine learning submodels we employed, extensive quantum chemical and experimental datasets were constructed or compiled in this work. They are all collected in the SolProp data collection (`https://zenodo.org/record/5970538`), and an overview of the different databases is given in Table 1. More details on the databases in SolProp that are developed as part of this work are given in the supporting information (sections S.2, S.3, and S.5), while some important details are given below.

For the CombiSolu-Exp database, solid solubility data in pure organic solvents are extracted from 105 different literature sources. The temperatures in this database are between 243 and 364K. To compare both methods for the calculation of $\Delta H_{diss,T}$ described before, experimental data at a broader temperature range are required. However, for drug-like solutes in organic solvents those experimental data are scarce. For this reason, additional datapoints, for example, for the solid solubility of polyaromatic hydrocarbons in water are collected at higher temperatures in the CombiSolu-HighT-Exp database. The data in the CombiSolu-HighT-Exp database are from 68 different sources for temperatures up to 593K.

The data in both databases are converted to have unique molecular identifiers (InChI) for solvent and solute. The solid solubilities are reported in the units of the original data source such as mole fraction, g/kg, and mol/kg. They are also converted to molar units required for model comparison $\log10(\mathrm{mol_{solute}}/\mathrm{L_{solution}})$ (further referred to as log(mol/L)). Ideally, the density used for conversion of the units should be the density of the solute and solvent mixture. However, those densities are in most cases not available, and thus the density of the pure solvent is used. The distribution of the $\log(S)$ values in the CombiSolu-Exp database

Table 1: Overview of the databases in the SolProp collection

Name	Description	Data entries	Reference
CombiSolv-QM	Quantum chemical database (COSMO-RS) for $\Delta G_{solv,298K}$	1000000 data 284 solvents 11029 solutes	[22]
CombiSolv-Exp	Experimental database for $\Delta G_{solv,298K}$	10145 data 291 solvents 1368 solutes	[22]
CombiSolvH-QM	Quantum chemical database (COSMO-RS) for $\Delta H_{solv,298K}$	800000 data 284 solvents 10891 solutes	This work
CombiSolvH-Exp	Experimental database for $\Delta H_{solv,298K}$	6322 data 1432 solvents 1665 solutes	[31]
SoluteDB	Empirical Abraham Solute Parameters (E, S, A, B, L)	8366 solutes	[31]
AqueousSolu-Exp	Experimental database for aqueous solubility	11804 data	This work
CombiSolu-Exp	Experimental database for solid solubility in organic solvents (243-364K)	4953 data 97 solvents 115 solutes	This work
CombiSolu-HighT-Exp	Experimental database for solid solubility in organic solvents, higher T up to 593K	1306 data 15 solvents 67 solutes	This work

is added to the supporting information (Figure S4) and compared to the distribution of the $\log(S)$ values in the AqueousSolu-Exp database. Duplicate entries are retained in both datasets to better analyze the effect of different experimental data sources and associated experimental uncertainties. For 12% of the duplicate data, the absolute deviation between $\log(S)$ data points is >0.2. The uncertainties in the experimental data are significant and affect the performance of the models, see supporting information (Figure S5).

Performance of the New Models for Aqueous Solid Solubility and Solvation Enthalpy

Our method for computing the solid solubility in organic solvents, Figure 1, requires new models for the aqueous solid solubility and the enthalpy of solvation, both at 298K. Below, the results of those models are shortly discussed. A more comprehensive discussion is provided in the supporting information (sections S.2 and S.3).

Aqueous Solid Solubility at 298K

The performance of the aqueous solid solubility model is tested against randomly selected experimental data and against a separate test set with lower experimental uncertainty (579 data points). To model the performance against the test set with lower experimental uncertainty, the solutes present in the test set are removed from the training and validation set. Note that those solutes are only removed to assess the model performance and they are included in the training set of the final model. A detailed description on the construction of this test set and details on the model performance are included in the supporting information (sections S.2.1 and S.2.3). In short, the model can predict the aqueous solid solubility at 298K for the random test set with a RMSE/MAE of 0.75/0.49 log(mol/L) and for data in the more accurate test set with a RMSE/MAE of 0.49/0.34 log(mol/L). These results show how the experimental uncertainty in the test set influences the model performance. The new model outperforms two publicly available models on the test set with lower experimental uncertainty: (i) the ALOGpS implementation from VCCLab,[23] and (ii) the recently published machine learning model SolTranNet.[8] Those predict $\log(S_{aq})$ of the same more accurate test set with a RMSE/MAE of 0.79/0.55 and 0.76/0.58 log(mol/L) respectively.

Solvation Enthalpy at 298K

The new transfer learning model for $\Delta H_{solv,298\mathrm{K}}$ is accurate to an RMSE/MAE of 0.81/0.49 kcal/mol on a randomly selected test set. The COSMO-RS method, as a comparison, can

compute the experimental data with a RMSE/MAE of 1.14/0.64 kcal/mol. The new model is comparable to the model published by Chung et al.[31] for small molecules; however it is expected to be more robust for solutes with a higher molar mass because of the employed transfer learning method.[22]

Performance of the New Models for Predicting Solid Solubility in Organic Solvents

Predicting Solid Solubility in Organic Solvents at 298K

The CombiSolu-Exp database has 1053 datapoints at 298K that are used for evaluation. The subset contains 98 unique solutes, 88 unique solvents and originates from 87 different sources. The solubility model is used to calculate $\log(S_{X,298\mathrm{K}})$ with Eq. 3 using (i) the transfer learning model for $\Delta G_{solv,298\mathrm{K}}$,[22] and (ii) the machine learning model for $\log(S_{aq,298\mathrm{K}})$ developed as part of this work as the reference solid solubility. The standard deviation of the predictions of the different models in the model ensembles are used to estimate the model uncertainties for the predictions of $\Delta G_{solv,298\mathrm{K}}$ and $\log(S_{aq,298\mathrm{K}})$. Those model uncertainties are further used to calculate the uncertainty on $\log(S_{X,298\mathrm{K}})$. In case the uncertainty is too large, the calculated solid solubility is not used for evaluation. This is the case for 2 solvent/solute combinations.

The performance of the model for solid solubility in organic solvents when using the model predictions for the aqueous solubility as the reference solid solubility is given in Figure 3 (left, blue). The solid solubility of the 1051 datapoints at 298K can be predicted with a RMSE/MAE of 0.89/0.62 log(mol/L). When the results for individual solutes in a range of organic solvents are studied in detail, trends are observed in the deviations from the experimental data. This can be seen in Figure 4 for 1-chloroanthraquinone, phenothiazine, benzoin, and 4-nitrobenzoic acid. The experimental data for those solutes originate from the publications by Dai et al.,[32] Flanagan et al.,[33] Grubbs et al.,[34] Hoover et al.,[35,36] Saifullah et

al.,[37] Stephens et al.,[38] Strickland et al.,[39] Yang et al.,[40] and Zhu et al.[41] In the supporting information (Figure S6), the same analysis is made for the 40 solutes that are most common in the database at 298K. As can be seen in Figure 4 (blue), the solid solubility of some solutes is predicted very well, whilst a constant deviation from the parity line is observed for others. In the latter case, the absolute value of the predicted solid solubility is off, but the relative solid solubility between various organic solvents is well captured by the model. This is the case, for example, for benzoin and 4-nitrobenzoic acid in Figure 4 (blue). The main reason for the constant deviation of $\log(S_{X,298\mathrm{K}})$ in different solvents from the parity line is the inaccurate prediction of the aqueous solid solubility of that solute.

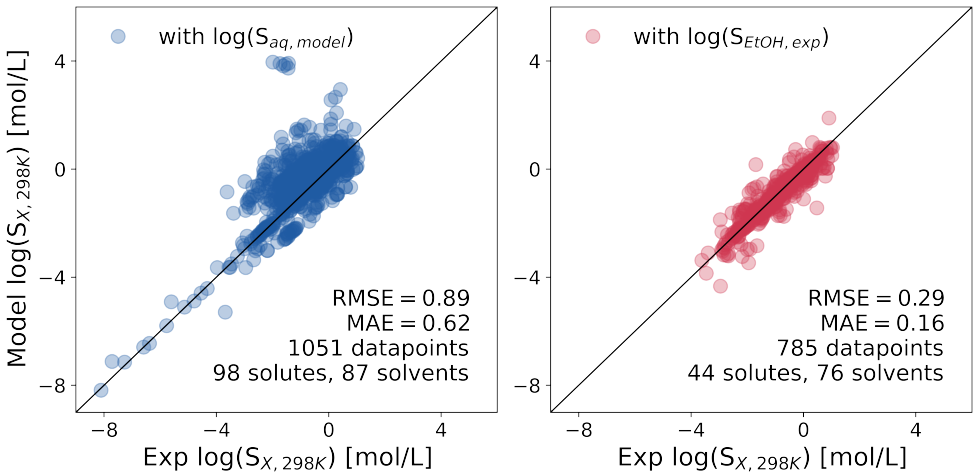

Figure 3: Parity plot for the performance of the solid solubility model at 298K using a model prediction for the aqueous solid solubility $\log(S_{aq,model})$ (left, blue), and using experimental data for the solid solubility in ethanol at 298K $\log(S_{EtOH,exp})$ when available (right, red).

Predictions of $\log(S_{X,298\mathrm{K}})$ can be improved by using experimental data for the solid solubility of that solute in a different solvent as the reference solid solubility in Eq. 3. This is demonstrated by using all experimental data in the CombiSolu-Exp database that are measured in ethanol at 298K as a reference for the calculations of $\log(S_{X,298\mathrm{K}})$ in other solvents. Experimental data in ethanol at 298K ($\log(S_{EtOH,exp})$) are available for 44 solutes. These data are used in Eq. 3, together with the model predictions of $\Delta G_{solv,298\mathrm{K}}$, to calculate the solid solubility of those solutes in 76 different solvents for a total of 785 datapoints.

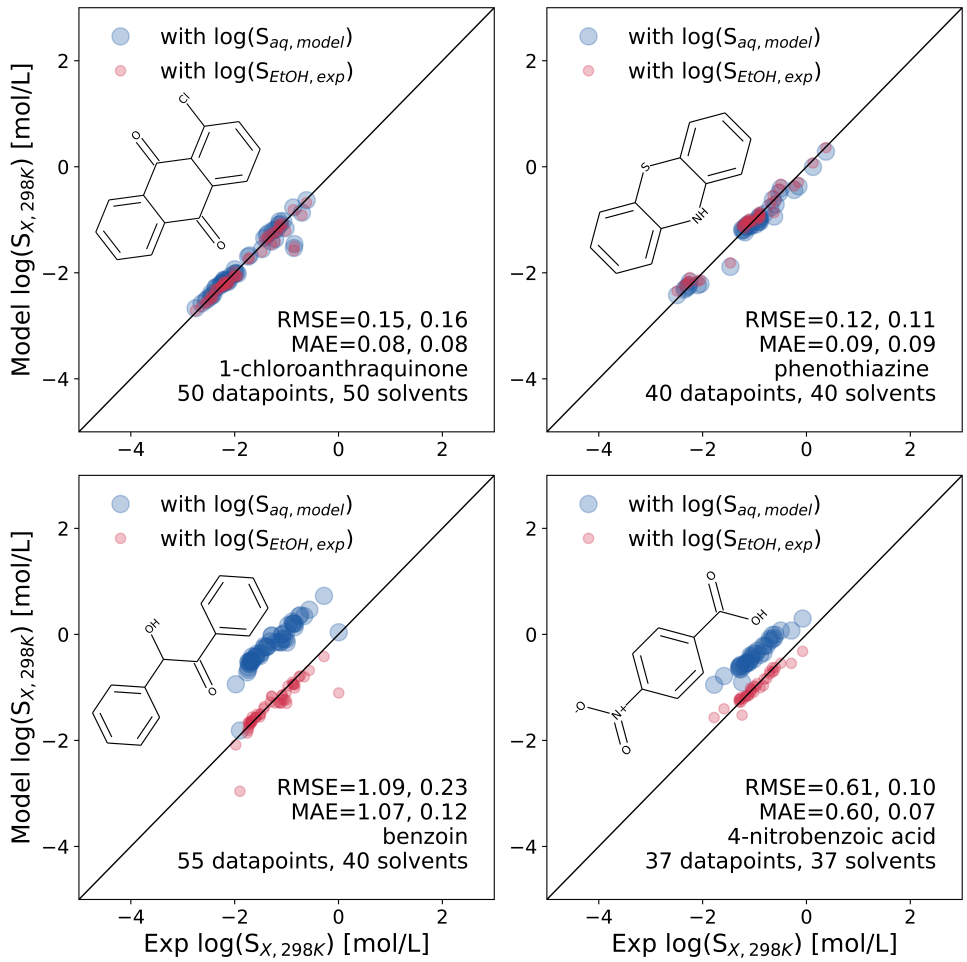

Figure 4: Parity plot for the performance of the solid solubility model at 298K for four individual solutes in a variety of organic solvents. The model uses model predictions for the aqueous solid solubility $\log(S_{aq,model})$ (blue), or experimental data for the solid solubility in ethanol $\log(S_{EtOH,exp})$ at 298K (red).

A comparison between the model predictions and experimental measurements is given in Figure 3 (right, red). The overall predictions for $\log(S_{X,298\mathrm{K}})$ improve to a RMSE/MAE of 0.29/0.16 log(mol/L). Using experimental data in other common solvents (methanol, acetonitrile, ethyl acetate, and toluene) as a reference gives similar results that are presented in the supporting information (Figure S7). In Figure 4, values for $\log(S_{X,298\mathrm{K}})$ calculated using $\log(S_{EtOH,exp})$ are indicated in red. The calculated solid solubility shifts towards the parity line compared to the calculations that use the model predictions for $\log(S_{aq,model})$. In general, using experimental data for the solid solubility in one solvent as a reference in Eq.

3 significantly improves the model predictions on the solid solubility in different solvents for the same solute. However, one should note that the experimental data used as a reference often originate from the same source as the other experimental data that are used to evaluate the model performance. As a result, some consistent errors specific to the experimental unit or procedure are compensated for and the model performance might be overestimated. This is demonstrated in the supporting information (Figure S8) by using experimental aqueous solid solubility data from both the CombiSolu-Exp and the AqueousSolu-Exp database as a reference. With aqueous solid solubility experimental data from the CombiSolu-Exp database, similar errors are obtained as compared to using experimental data measured in ethanol (RMSE/MAE equal to 0.27/0.06). When experimental aqueous solid solubility from a different source (AqueousSolu-Exp) is used, the model error increases significantly (RMSE/MAE of 0.56/0.35). The difference between both is related to the experimental uncertainty for aqueous solid solubility data. When comparing the overlapping experimental aqueous solid solubility data between the CombiSolu-Exp and the AqueousSolu-Exp databases, a RMSE/MAE of 0.27/0.17 is obtained (Figure S9).

Predicting Temperature-Dependent Solid Solubility

First method: neglecting the temperature dependence of $\Delta H_{diss}(T)$

The temperatures in the CombiSolu-Exp database range from 243 to 364K. Since these temperatures are close to room temperature (298K), it is assumed reasonable to neglect the temperature dependence of the dissolution enthalpy. To calculate $\log(S_{X,T})$ at temperature T with Eq. 9, the solid solubility at 298K from section is used together with the dissolution enthalpy $\Delta H_{diss,298\mathrm{K}}$. The dissolution enthalpy at 298K is calculated from (i) the new transfer learning model predictions for $\Delta H_{solv,298\mathrm{K}}$, (ii) the machine model predictions for solute parameters,[31] and (iii) the correlation for $\Delta H_{sub,298\mathrm{K}}$ published by Abraham and Acree[30] (Eq. 10).

The calculation of $\log(S_{X,T})$ is done for the complete CombiSolu-Exp database and the

results are given in Figure 5. First, $\log(S_{aq,model,298\mathrm{K}})$ is used to calculate $\log(S_{X,298\mathrm{K}})$ unless the standard deviation of the ensemble of model predictions for $\log(S_{X,298\mathrm{K}})$ is too large. For 4922 datapoints with 115 unique solutes and 95 unique solvents, the results are given in Figure 5 (left) and the solid solubility is predicted with a RMSE/MAE of 1.49/0.99 log(mol/L). Similar to the calculation of $\log(S_{X,298\mathrm{K}})$ (see Figure 3), the results significantly improve if experimental data for the solid solubility at 298K is used. This is demonstrated using the experimental solid solubility in ethanol at 298K if available in the CombiSolu-Exp database for that solute. The calculated values of $\log(S_{X,T})$ using $\log(S_{EtOH,exp,298\mathrm{K}})$ are compared to experimental data in Figure 5 (right) for 3071 datapoints with 44 unique solutes and 83 unique solvents. The model performance improves to a RMSE/MAE of 0.44/0.29 log(mol/L).

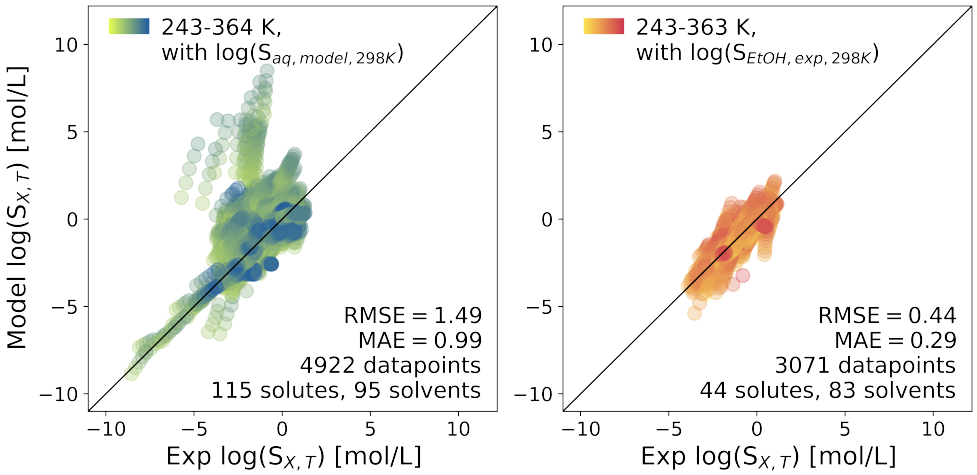

Figure 5: Parity plot for the performance of the solid solubility model at different temperatures (243-333K) using a model prediction for the aqueous solid solubility at 298K $\log(S_{aq,model,298\mathrm{K}})$ (left, blue-yellow gradient), and using experimental data for the solid solubility in ethanol at 298K $\log(S_{EtOH,exp,298\mathrm{K}})$ when available (right, red-yellow gradient).

Similar to $\log(S_{X,298\mathrm{K}})$, trends are observed when model calculations are compared to the experimental data for individual solutes. This is demonstrated in Figure 6 for n-acetylglycine, benzoin, bezafibrate, and chlorpropamide. The experimental data for these solvents originate from the publications by Dai et al.,[32] Guo et al.,[42] Guo et al.,[43] Liu et al.,[44] Liu et al.,[45] Stephens et al.,[38] Strickland et al.,[39] Yang et al.,[46] Yang et al.,[40] and Zhu et al.[41] For the

40 most common solutes in the CombiSolu-Exp database, the same analysis is added to the supporting information (Figure S10). Replacing $\log(S_{aq,model,298\mathrm{K}})$ with experimental data $\log(S_{EtOH,exp,298\mathrm{K}})$ for the calculation of $\log(S_{X,298\mathrm{K}})$ shifts the prediction results to the parity line, which is in line with the results for $\log(S_{X,298\mathrm{K}})$. The solid solubility trends with respect to temperature are well captured for most solutes in Figure 6. In case of bezafibrate, a slope that deviates from the parity line is observed when the model calculations are compared to the experimental data as a function of temperature. Because the slope is very similar for the solid solubility of bezafibrate in 16 solvents, we infer that the deviating slope is caused by a misprediction of the sublimation enthalpy of bezafibrate.

Second method: using the temperature dependent $\Delta H_{diss,T}$

The temperature dependence of ΔH_{diss} can be neglected when the temperature does not deviate too much from room temperature. In Figures 5 and 6 it is demonstrated that this method works for a temperature between 243 and 364K. However, to calculate the solid solubility at higher temperatures, the temperature dependence of the dissolution enthalpy has to be considered through numerical integration, Eq. 11. For this method, we use (i) the machine learning model predictions for the solute parameters[31] (ii) the correlations published by Abraham and Acree[29,30] for $\Delta H_{sub,298\mathrm{K}}$, $C_{p,g}$, and $C_{p,s}$, (iii) the transfer learning model predictions for $\Delta G_{solv,298\mathrm{K}}$[22] and $\Delta H_{solv,298\mathrm{K}}$, and (iv) the method published by Chung et al.[47] to calculate the temperature dependent solvation enthalpy using critical properties of the solvent for calculation of the saturation density. Since the critical properties of the solvent are only collected for a limited set of solvents considered in this work, this method is currently restricted to ∼100 solvents with tabulated critical properties.

The calculated $\Delta H_{diss}(T)$ is used to predict the experimental solid solubility in the CombiSolu-Exp database. A RMSE/MAE of 1.49/0.98 log(mol/L) is achieved in case $\log(S_{aq,model,298\mathrm{K}})$ is used to determine $\log(S_{X,298\mathrm{K}})$, and 0.44/0.29 log(mol/L) in case $\log(S_{EtOH,exp,298\mathrm{K}})$ is used. No significant performance improvement is observed compared to Figure 5 where $\Delta H_{diss,T} \approx \Delta H_{diss,298\mathrm{K}}$, which confirms that the temperature dependence of the dissolution

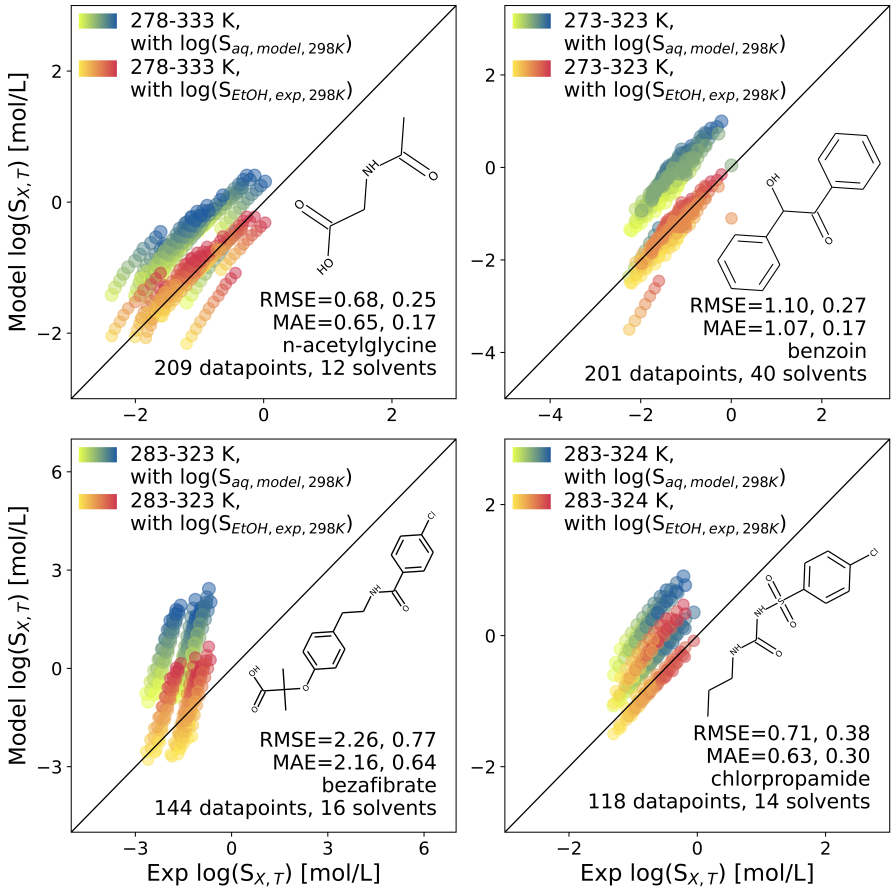

Figure 6: Parity plot for the performance of the solid solubility model at different temperatures (273-364K) for four individual solutes in a variety of organic solvents. The model uses model predictions for the aqueous solid solubility at 298K $\log(S_{aq,model,298\mathrm{K}})$ (blue-yellow gradient), or experimental data for the solid solubility in ethanol at 298K $\log(S_{EtOH,exp,298\mathrm{K}})$ (red-yellow gradient).

enthalpy can be safely neglected between 243 and 364K.

To compare the performance of both methods at higher temperatures, model predictions are compared to the experimental data in the CombiSolu-HighT-Exp database. Compared to the CombiSolu-Exp database, many of the solutes in this database are not drug-like components but *e.g.* polyaromatic hydrocarbons, and the majority of the data are measured in water rather than other organic solvents. A comparison of the different methods is given in Figure 7 for selected solvent/solute pairs (isophathalic acid in water, 4-formylbenzoic acid in acetic acid, phenothiazine in water, and 2-methylanthracene in water). The experimental

data for those solvent/solute pairs originate from Sheehan et al.,[48] Long et al.,[49] Han et al.,[50] Cheng et al.,[51] Lyu et al.,[52] Li et al.,[53] Sun et al.,[54] Ahmadian et al.,[55] Karasek et al.,[56,57] and Dohanyosova et al.[58]

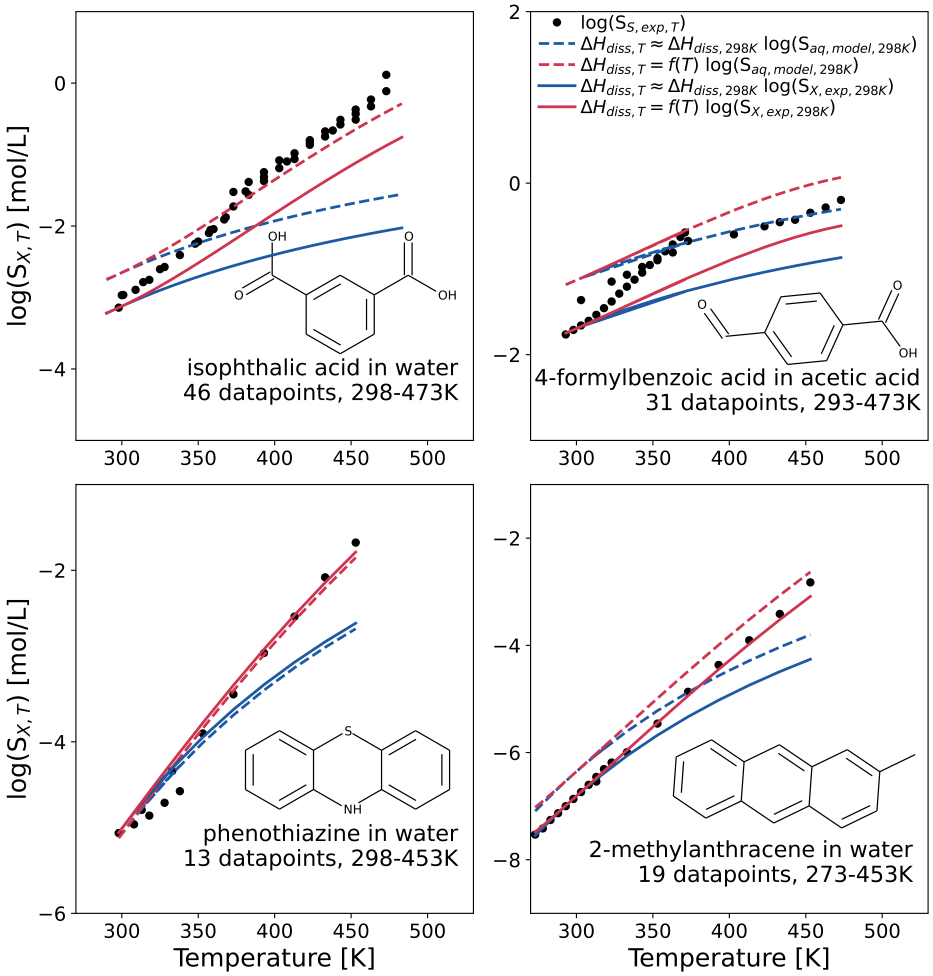

Figure 7: Solid solubility as a function of temperature for four solvent/solute pairs with four different methods. Model predictions are done with the temperature dependent dissolution enthalpy ($\Delta H_{diss} = f(T)$, red) or the dissolution enthalpy at 298K ($\Delta H_{diss,T} \approx \Delta H_{diss,298\mathrm{K}}$, blue). The model uses experimental data for the solid solubility at 298K ($\log(S_{X,exp,298\mathrm{K}})$, full lines) or the model predicted solid solubility at 298K based on the aqueous solid solubility and solvation energy ($\log(S_{aq,model,298\mathrm{K}})$, dashed lines)

In Figure 7, four different methods for the prediction of $\log(S_{X,T})$ are compared to experimental data as a function of temperature. If the temperature dependence of the dissolution enthalpy is accounted for ($\Delta H_{diss} = f(T)$) and the experimental solid solubility at 298K for that solvent/solute pair ($\log(S_{X,exp,298\mathrm{K}})$) is used for Eq. 11, the best results are obtained

(Figure 7, full red line). In case the experimental solid solubility at 298K is replaced by the model solid solubility $\log(S_{X,model,298\mathrm{K}})$ calculated based on the aqueous solid solubility $\log(S_{aq,model,298\mathrm{K}})$ and $\Delta G_{solv,298\mathrm{K}}$ predictions, the trends of $\log(S_{X,model,T})$ as a function of temperature are well captured, but the absolute values are off in some cases (Figure 7, dashed red line). The blue lines in Figure 7 do not account for the temperature dependence of the dissolution enthalpy ($\Delta H_{diss,T} \approx \Delta H_{diss,298\mathrm{K}}$). For most solvent/solute pairs, this approximation causes significant deviations in $\log(S_{X,T})$ predictions above 350K.

The different methods are compared for 60 additional solvent/solute pairs in the supporting information (Figure S11). While the temperature dependence of the solid solubility is predicted well for most pairs, there is one pair, namely 5-fluorouracil in water, whose predicted temperature gradient shows an opposite trend ($\Delta H_{diss,T} < 0$) with the experimental data. It is found that the sublimation enthalpy of 5-fluorouracil is substantially underpredicted by our model as 20.8 kcal/mol compared to the reported experimental value of 31.8 kcal/mol.[59] This experimental $\Delta H_{sub,298\mathrm{K}}$ value is used to recalculate $\log(S_{X,T})$ according to Eq. 11, which significantly improved the model performance (see supporting information Figure S12).

Alternatively, it is possible to derive a more reliable value for $\Delta H_{sub,298\mathrm{K}}$ from the measured solid solubilities across a range of temperatures. The dissolution enthalpy at 298K ($\Delta H_{diss,298\mathrm{K}}$) can be estimated from the temperature gradient of the experimental solid solubility data near 298K using Eq. 8 and $\Delta H_{sub,298\mathrm{K}}$ can be subsequently estimated using Eq. 9 based on the estimated $\Delta H_{diss,298\mathrm{K}}$ and the predicted $\Delta H_{solv,298\mathrm{K}}$. New predictions are made based on this approach for the solute/solvent pairs whose temperature gradient at 298K could be approximated from the experimental data. The resulting plots can be found in the supporting information (Figure S13). As can be seen, substantial improvements are made using this approach for several pairs such as budesonide in water, 4-formylbenzoic acid in 1-methyl-2-pyrrolidinone, paracetamol in water, and propazine in water. However, solid solubility data sometimes have high experimental uncertainties and accurate measurements

are needed to correctly estimate the temperature gradient. Moreover, if the main source of error is the $\Delta H_{solv,T}$ term, this approach will not be able to improve the predictions at high temperatures.

To conclude this discussion, the contribution of each term of the dissolution enthalpy in Eq. 7 is further analyzed. The dissolution enthalpy is calculated using the four thermodynamic processes involving $C_{p,s}$, $\Delta H_{sub,298\mathrm{K}}$, $C_{p,g}$, and $\Delta H_{solv,T}$, respectively. The individual contribution of these terms to the predicted solid solubility is plotted as a function of temperature for four solute/solvent pairs and presented in the supporting information (Figure S14). From the the plot, it can be seen that the contribution from the $C_{p,s}$ and $C_{p,g}$ terms (heating of the solid- and gas-phase solutes) is much smaller than that of the $\Delta H_{sub,298\mathrm{K}}$ and $\Delta H_{solv,T}$ terms. Subsequently, the solid solubility prediction is made using the $\Delta H_{diss,T}$ that includes only $\Delta H_{sub,298\mathrm{K}}$ and $\Delta H_{solv,T}$ and excludes the heat capacity terms and compared with the prediction made using the full $\Delta H_{diss,T}$ expression. The resulting plot can be found in the supporting information (Figure S15). The two predictions using the partial and full $\Delta H_{diss,T}$ expressions are nearly identical, confirming that the heat capacity terms, and especially the temperature dependency of the heat capacity, can be safely neglected when computing $\Delta H_{diss,T}$.

Conclusions

Thermodynamics, machine learning, and known correlations are combined to build a predictive tool for the solid solubility of neutral solutes in many organic solvents over a wide temperature range. This is the first modeling tool that can predict the solid solubility for a broad range of solvents and temperatures without the necessity for empirical parameters. Data scarcity for the solid solubility in different solvents and at different temperatures has limited the use of machine learning for the direct prediction of this property. A software package and an user-friendly web interface are designed and publicly available for the prediction

of solubility of gaseous and solid solutes.

Three new datasets are compiled and provided as part of this work. The CombiSolu-Exp database contains 4953 solid solubility datapoints for 115 unique solutes and 97 unique solvents. The CombiSolu-HighT-Exp database contains 1306 high temperature solid solubility datapoints with 67 unique solutes and 15 unique solvents. The AqueousSolu-Exp database provides data specifically for aqueous solid solubility at 298K, and contains 11804 unique solutes. All datasets used for validation and construction of the models in this work are provided as part of the SolProp data collection.

Based on only molecular identifiers for a solute and a solvent, our new machine learning models predict $\log(S_{aq,298\mathrm{K}})$ and $\Delta H_{solv,298\mathrm{K}}$. Those new models are combined with existing models for $\Delta G_{solv,298\mathrm{K}}$, $\Delta H_{sub,298\mathrm{K}}$, $C_{p,g}$, $C_{p,s}$, and the temperature dependence of ΔG_{solv} to construct a model for the solid solubility of any neutral organic solute in many organic solvents X over a range of temperatures T; $\log(S_{X,T})$. One of the strengths of this method is that intermediate outputs are also provided to the user. Several of those are interesting properties by themselves. They are further used for the calculation of $\log(S_{X,T})$, but if experimental data are available for any of those properties the solid solubility output can be corrected to gain more accurate predictions. The main downside of this method is that it relies on the predictions of several submodels to calculate $\log(S_{X,T})$. If one of those submodels fails, also the prediction of $\log(S_{X,T})$ will be off unless experimental input can compensate.

The new model predicts the solid solubility at 298K $\log(S_{X,298\mathrm{K}})$ with a RMSE/MAE of 0.89/0.62. Even in cases where the prediction of the absolute solid solubility is poor, the model captures the relative solubility between different solvents accurately. If experimental data are available for the solid solubility of the solute in at least one different solvent at 298K, the RMSE/MAE improve to 0.29/0.16. The solid solubility at temperatures below approximately 350K can be calculated for the same broad range of solutes and solvents based on only the molecular identifiers. Below 350K, the temperature dependence of the

dissolution enthalpy can be neglected. In this case, machine learning models are used for the prediction of $\Delta H_{solv,298\mathrm{K}}$ and solute parameters E, S, A, B. Together with $\log(S_{X,298\mathrm{K}})$ and the published correlation for $\Delta H_{sub,298\mathrm{K}}$, the temperature dependent solid solubility can be calculated. In case the model for $S_{aq,298\mathrm{K}}$ is used for the calculation, $\log(S_{X,T})$ can be predicted with a RMSE/MAE of 1.49/0.99. If experimental data for the solid solubility of the solute in at least one solvent are available, the RMSE/MAE of $\log(S_{X,T})$ improve to 0.44/0.29. At temperatures higher than 350K, the temperature dependence of the dissolution enthalpy has to be accounted for. This requires the calculation of the saturated solvent density and information on the critical properties of the solvent. If those are available, our method can calculate the solid solubility of solvent/solute pairs at elevated temperatures even approaching the critical point of the solvent.

The developed model has an excellent performance in predicting solubility trends in different solvents at 298K and as a function of temperature. Even though the absolute solid solubility prediction can be off in some cases, the ability to predict solubility trends has tremendous applications, including but not limited to pharmaceutical production and solvent selection, the design of redox flow batteries, and organophotocatalysis.

Computational Methods

First method: neglecting the temperature dependence of $\Delta H_{diss,T}$

For the first method, we neglect the temperature variation of the dissolution enthalpy ($\Delta H_{diss,T} \approx \Delta H_{diss,298\mathrm{K}}$). This assumption results in Eq. 8 and is valid as long as the temperature T does not deviate too much from 298K. For the calculation of $\Delta H_{diss,298\mathrm{K}}$, we need the sublimation enthalpy at 298K ($\Delta H_{sub,298\mathrm{K}}$) and the solvation enthalpy at 298K ($\Delta H_{solv,298\mathrm{K}}$). Abraham and Acree[30] developed a correlation to calculate $\Delta H_{sub,298\mathrm{K}}$ based on solute parameters as shown in Eq. 10. The first four solute parameters (E, S, A, B) can be predicted with the machine learning models developed during our earlier work,[31] and

the solute parameter V can be calculated using the McGowan method.[60] The additional parameters introduced by Abraham and Acree,[30] *i.e.* I(OH,adj), I(OH,non), and I(NH) in Eq. 10, are indicator variables for the presence of aliphatic diols with adjacent OH groups, aliphatic diols with non-adjacent OH groups, and aliphatic amine groups, respectively. In our earlier work,[31] we also published a model for the prediction of $\Delta H_{solv,298\mathrm{K}}$. Even though a good model performance was achieved, to obtain an improved performance for solutes with a higher molar mass, a new model is built for $\Delta H_{solv,298\mathrm{K}}$ using a transfer learning methodology similar to the one we used for the prediction of $\Delta G_{solv,298\mathrm{K}}$.[22] In the supporting information (section S.3) more information can be found on this transfer learning model architecture and the quantum chemical and experimental databases employed.

$$\ln(\frac{S_T}{S_{298\mathrm{K}}}) = \frac{-\Delta H_{diss,298\mathrm{K}}}{R}\left(\frac{1}{T} - \frac{1}{298\mathrm{K}}\right) \tag{8}$$

$$\ln(\frac{S_T}{S_{298\mathrm{K}}}) = \frac{-\left(\Delta H_{sub,298\mathrm{K}} + \Delta H_{solv,298\mathrm{K}}\right)}{R}\left(\frac{1}{T} - \frac{1}{298\mathrm{K}}\right) \tag{9}$$

$$\begin{aligned}\Delta H_{sub,298\mathrm{K}}[\mathrm{kJ/mol}] = 9.96 - 2.10\ E + 24.10\ S + 13.70\ A + 0.79\ B \\ + 38.71\ V - 1.36\ S \cdot S + 36.90\ A \cdot B \\ + 1.86\ V \cdot V - 10.89\ I(\mathrm{OH,adj}) \\ + 14.74\ I(\mathrm{OH,non}) + 9.69\ I(\mathrm{NH})\end{aligned} \tag{10}$$

Neglecting the temperature dependence of the dissolution enthalpy allows us to use the machine learning models developed in earlier work for a broad range of organic solvents; however, this method is expected to yield a high error if the temperature deviates significantly from 298K.

Second method: using the temperature dependent $\Delta H_{diss}(T)$

With the second method, the temperature dependence of $\Delta H_{diss,T}$ is accounted for through numerical integration and the solid solubility can be calculated for a broader temperature range, but a limited amount of solvents. To calculate the solid solubility at temperature T with the second method, the expression for $\Delta H_{diss,T}$ given in Eq. 7 is plugged into the integral in Eq. 5. After integration, this yields Eq. 11. $\Delta H_{sub,298\mathrm{K}}$ is defined as the enthalpy of sublimation at 298K and is independent of temperature.

$$\ln(\frac{S_T}{S_{298\mathrm{K}}}) = \frac{-C_{p,s} \cdot 298\mathrm{K} - \Delta H_{sub,298\mathrm{K}} + C_{p,g} \cdot 298\mathrm{K}}{R} \left(\frac{1}{T} - \frac{1}{298\mathrm{K}}\right) + \frac{-C_{p,s} + C_{p,g}}{R} \cdot \ln\left(\frac{T}{298\mathrm{K}}\right) + \int_{298\mathrm{K}}^{T} \frac{\Delta H_{solv,T}}{RT^2}\, dT \quad (11)$$

Similar as with the previous method, $\Delta H_{sub,298\mathrm{K}}$ is estimated from the correlation published by Abraham and Acree[30] (Eq. 10). Also $C_{p,g}$ and $C_{p,s}$ at 298K are estimated using a correlation recently published by Abraham and Acree[29] (Eqs. 12 and 13) using the same solute parameters as in Eq. 10. Other methods such as the group additivity method in RMG[61] are available for estimating heat capacities of gaseous compounds across temperatures. However, as the temperature dependence of heat capacities is neglected, $C_{p,g}$ at 298K from Eq. 12 is used instead in this work.

$$\begin{aligned} C_{p,g}[\mathrm{J/K/mol}] = {} & -8.62 - 24.33\ E - 15.83\ S + 12.35\ A + 13.27\ B \\ & + 160.00\ V + 10.66\ S \cdot S - 2.11\ A \cdot B + 0.41\ V \cdot V \end{aligned} \quad (12)$$

$$\begin{aligned} C_{p,s}[\mathrm{J/K/mol}] = {} & 11.63 - 34.18\ E - 1.20\ S - 1.09\ A + 12.28\ B \\ & + 181.69\ V + 2.32\ S \cdot S + 4.24\ A \cdot B - 1.85\ V \cdot V \\ & - 28.50\ I(\mathrm{OH,non}) \end{aligned} \quad (13)$$

$\Delta H_{solv,T}$ is predicted using our previous work on the temperature dependence of solvation free energies.[47] This method can predict temperature-dependent solvation free energies $\Delta G_{solv,T}$ up to the critical temperature of a solvent along the saturation curve. This can be done for any solvent-solute pair if $\Delta G_{solv,298\mathrm{K}}$, $\Delta H_{solv,298\mathrm{K}}$, and the temperature-dependent saturation density of the solvent in the gas and liquid phase are known. Both $\Delta G_{solv,298\mathrm{K}}$ and $\Delta H_{solv,298\mathrm{K}}$ can be predicted by the transfer learning models from our previous[22] and current work. The solvent saturation density is estimated using the open-source thermophysical property software CoolProp[62] and the critical temperature and density of the solvent. Details on the estimation method for the saturation density of the solvent can be found in the supporting information (section S.4). Based on this method, $\Delta G_{solv,T}$ is predicted across temperatures, and $\Delta H_{solv,T}$ is computed using the temperature gradient of $\Delta G_{solv,T}$ and the relationship $\Delta H = \Delta G - T\frac{\mathrm{d}\Delta G}{\mathrm{d}T}$. Since $\Delta H_{solv,T}$ has a complex temperature dependence, the last term of Eq. 11 is obtained from numerical integration. The downside of this second method is that the critical temperature and critical density of the solvent are required to calculate $\Delta H_{solv,T}$. We have collected the critical properties of approximately 100 organic solvents from various literature sources for this work.

Although the method by Chung et al.[47] evaluates the temperature-dependent solvation free energy at a solvent's saturation pressure, they showed that the method could also provide a reasonable prediction for a range of pressures as long as the temperature is not too close to the critical temperature of the solvent (*i.e.* $T < 0.8T_\mathrm{c}$). They indicated that the pressure effect becomes more pronounced at elevated temperatures and that the method should be used for pressures close to the saturation pressure in the high temperature region (*i.e.* $P < 1.3P_\mathrm{sat}$ for $0.8T_\mathrm{c} < T < T_\mathrm{c}$). Hence our method using Eq. 11 is expected to be valid for similar conditions.

Acknowledgement

The authors acknowledge the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS), and the DARPA Accelerated Molecular Discovery (AMD) program (DARPA HR00111920025) for funding. To perform quantum calculations, we used HPC resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. The MIT SuperCloud Lincoln Laboratory Supercomputing Center is acknowledged for providing HPC resources to train the deep neural networks that have contributed to the research results reported within this paper. We thank Charles McGill for suggestions that improved this manuscript.

Supporting Information Available

The software used to construct the models for $\Delta G_{solv,298\mathrm{K}}$, $\Delta H_{solv,298\mathrm{K}}$, and $\log(S_{aq,298\mathrm{K}})$, and to calculate $\log(S_{X,T})$ is open-source available on github (`https://github.com/fhvermei/SolProp_ML`). The software together with the trained models are compiled in a conda package that can be used for predictions of $\log(S_{X,T})$ (`https://anaconda.org/fhvermei/solprop_ml`). A notebook with examples for the calculation of $\log(S_{X,T})$ is provided on github. This notebook also provides examples of how experimental data can be used to improve model performance. The SolProp data collection (see Table 1) is available as part of the supporting information and on Zenodo (`https://zenodo.org/record/5970538`), also the machine learning models and solid solubility predictions for the CombiSolu-Exp and CombiSolu-HighT-Exp are made available. The databases are open access and distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license (`https://creativecommons.org/licenses/by/4.0/`). Citations should refer directly to this manuscript. A user-friendly interface for the calculation of $\log(S_{X,T})$ and related properties using the different methodologies described in this work are also available on our

website (`https://rmg.mit.edu/database/solvation/searchSolubility/`).

The supporting information contains a comprehensive discussion on the curation of datasets and on the training of the machine learning submodels for the prediction of aqueous solid solubility and solvation enthalpy. Additional validation of the models for the calculations of solid solubility is presented for more solvent-solute pairs at various temperatures.

References

(1) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data* **2019**, *6*, 143.

(2) Kuentz, M.; Bergström, C. A. S. Synergistic Computational Modeling Approaches as Team Players in the Game of Solubility Predictions. *Journal of Pharmaceutical Sciences* **2021**, *110*, 22–34.

(3) Renon, H. N R T L: An empirical equation or an inspiring model for fluids mixtures properties? *Fluid Phase Equilibria* **1985**, *24*, 87–114.

(4) Abraham, M. H.; Smith, R. E.; Luchtefeld, R.; Boorem, A. J.; Luo, R.; Acree, W. E. Prediction of Solubility of Drugs and Other Compounds in Organic Solvents. *J. Pharm. Sci.* **2010**, *99*, 1500–1515.

(5) Lusci, A.; Pollastri, G.; Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling* **2013**, *53*, 1563–1575.

(6) Cui, Q.; Lu, S.; Ni, B.; Zeng, X.; Tan, Y.; Chen, Y. D.; Zhao, H. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Frontiers in Oncology* **2020**, *10*, 121.

(7) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, *11*, 1–10.

(8) Francoeur, P. G.; Koes, D. R. SolTranNet–A Machine Learning Tool for Fast Aqueous Solubility Prediction. *Journal of Chemical Information and Modeling* **2021**, *61*, 2530–2536.

(9) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(10) Lim, H.; Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chemical Science* **2019**, *10*, 8306–8315.

(11) Lim, H.; Jung, Y. MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning. *Journal of Cheminformatics* **2021**, *13*, 56.

(12) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *Journal of Chemical Information and Modeling* **2019**, *59*, 1338–1346.

(13) Pathak, Y.; Mehta, S.; Priyakumar, U. D. Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks. *Journal of Chemical Information and Modeling* **2021**, *61*, 689–698.

(14) Hille, C.; Ringe, S.; Deimel, M.; Kunkel, C.; Acree, W. E.; Reuter, K.; Oberhofer, H. Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme. *The Journal of Chemical Physics* **2018**, *150*, 041710.

(15) Sunsandee, N.; Hronec, M.; Štolcová, M.; Leepipatpiboon, N.; Pancharoen, U. Thermodynamics of the solubility of 4-acetylbenzoic acid in different solvents from 303.15 to 473.15 K. *Journal of Molecular Liquids* **2013**, *180*, 252–259.

(16) Teoh, W. H.; Vieira De Melo, S. A.; Mammucari, R.; Foster, N. R. Solubility and solubility modeling of polycyclic aromatic hydrocarbons in subcritical ethanol and water mixtures. *Industrial and Engineering Chemistry Research* **2014**, *53*, 10238–10248.

(17) Fowles, D. J.; Palmer, D. S.; Guo, R.; Price, S. L.; Mitchell, J. B. O. Toward Physics-Based Solubility Computation for Pharmaceuticals to Rival Informatics. *Journal of Chemical Theory and Computation* **2021**, *17*, 3700–3709.

(18) Robinson, S. G.; Yan, Y.; Hendriks, K. H.; Sanford, M. S.; Sigman, M. S. Developing a Predictive Solubility Model for Monomeric and Oligomeric Cyclopropenium-Based Flow Battery Catholytes. *Journal of the American Chemical Society* **2019**, *141*, 10171–10176.

(19) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Accounts of Chemical Research* **2008**, *41*, 760–768.

(20) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *The Journal of Physical Chemistry A* **1998**, *102*, 5074–5085.

(21) Moine, E.; Privat, R.; Sirjean, B.; Jaubert, J.-N. Estimation of Solvation Quantities from Experimental Thermodynamic Data: Development of the Comprehensive CompSol Databank for Pure and Mixed Solutes. *Journal of Physical and Chemical Reference Data* **2017**, *46*, 33102.

(22) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021**, *418*, 129307.

(23) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual Computational Chemistry Laboratory – Design and Description. *Journal of Computer-Aided Molecular Design* **2005**, *19*, 453–463.

(24) Wishart, D. S. et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(25) Mitchell, J.; McDonagh, J.; Boobier, S. DLS-100 Solubility Dataset. *Figshare* **2017**, DOI: 10.6084/m9.figshare.5545639.v1.

(26) Mansouri, K.; Grulke, C.; Judson, R.; Williams, A. Free online access to experimental and predicted properties through the EPA's CompTox Chemistry Dashboard. *Figshare* **2018**, DOI: 10.23645/epacomptox.5179045.

(27) Sushko, I. et al. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533–554.

(28) Grant, D. J.; Mehdizadeh, M.; Chow, A. H.; Fairbrother, J. E. Non-linear van't Hoff solubility-temperature plots and their pharmaceutical interpretation. *International Journal of Pharmaceutics* **1984**, *18*, 25–38.

(29) Abraham, M. H.; Acree, W. E. Estimation of heat capacities of gases, liquids and solids, and heat capacities of vaporization and of sublimation of organic chemicals at 298.15 K. *Journal of Molecular Liquids* **2020**, *317*, 113969.

(30) Abraham, M. H.; Acree, W. E. Estimation of enthalpies of sublimation of organic, organometallic and inorganic compounds. *Fluid Phase Equilibria* **2020**, *515*, 112575.

(31) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *Journal of Chemical Information and Modeling* **2022**, *62*, 433–446.

(32) Dai, J.; Eddula, S.; Jiang, C.; Zhang, A.; Liu, K.; Zhu, S.; Wang, S.; Gupta, A.; Churchill, B.; Garcia, E.; Acree, W. E.; Abraham, M. H. Abraham model correlations

for describing solute transfer processes into diethyl carbonate. *Physics and Chemistry of Liquids* **2021**, *59*, 26–39.

(33) Flanagan, K. B.; Hoover, K. R.; Garza, O.; Hizon, A.; Soto, T.; Villegas, N.; Acree, W. E.; Abraham, M. H. Mathematical correlation of 1-chloroanthraquinone solubilities in organic solvents with the Abraham solvation parameter model. *Physics and Chemistry of Liquids* **2006**, *44*, 377–386.

(34) Grubbs, L. M.; Saifullah, M.; De La Rosa, N. E.; Ye, S.; Achi, S. S.; Acree, W. E.; Abraham, M. H. Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents. *Fluid Phase Equilibria* **2010**, *298*, 48–53.

(35) Hoover, K. R.; Coaxum, R.; Pustejovsky, E.; Stovall, D. M.; Acree, W. E.; Abraham, M. H. Thermochemical behavior of dissolved carboxylic acid solutes: part 4 – mathematical correlation of 4-nitrobenzoic acid solubilities with the abraham solvation parameter model. *Physics and Chemistry of Liquids* **2004**, *42*, 339–347.

(36) Hoover, K. R.; Acree JR, W. E.; Abraham, M. H. Mathematical correlation of phenothiazine solubilities in organic solvents with the Abraham solvation parameter model. *Physics and Chemistry of Liquids* **2006**, *44*, 367–376.

(37) Saifullah, M.; Ye, S.; Grubbs, L. M.; De La Rosa, N. E.; Acree, W. E.; Abraham, M. H. Abraham Model Correlations for Transfer of Neutral Molecules to Tetrahydrofuran and to 1,4-Dioxane, and for Transfer of Ions to Tetrahydrofuran. *J. Solution Chem.* **2011**, *40*, 2082–2094.

(38) Stephens, T. W.; Loera, M.; Calderas, M.; Diaz, R.; Montney, N.; Acree, W. E.; Abraham, M. H. Determination of Abraham model solute descriptors for benzoin based on measured solubility ratios. *Physics and Chemistry of Liquids* **2012**, *50*, 254–265.

(39) Strickland, S.; Ocon, L.; Zhang, A.; Wang, S.; Eddula, S.; Liu, G.; Tirumala, P.; Huang, J.; Dai, J.; Jiang, C.; Acree, W. E.; Abraham, M. H. Abraham model correlations for describing dissolution of organic solutes and inorganic gases in dimethyl carbonate. *Physics and Chemistry of Liquids* **2021**, *59*, 181–195.

(40) Yang, Y.; Tang, W.; Li, X.; Han, D.; Liu, Y.; Du, S.; Zhang, T.; Liu, S.; Gong, J. Solubility of Benzoin in Six Monosolvents and in Some Binary Solvent Mixtures at Various Temperatures. *Journal of Chemical & Engineering Data* **2017**, *62*, 3071–3083.

(41) Zhu, Y.; Cheng, C.; Chen, G.; Zhao, H. Solubility Modeling and Mixing Properties for Benzoin in Different Monosolvents and Solvent Mixtures at the Temperature Range from 273.15 to 313.15 K. *Journal of Chemical & Engineering Data* **2018**, *63*, 341–351.

(42) Guo, X.; Cheng, Y. W.; Wang, L. J.; Li, X. Solubility of terephthalic acid in aqueous N-methyl pyrrolidone and N,N-dimethyl acetamide solvents at (303.2 to 363.2) K. *Journal of Chemical and Engineering Data* **2008**, *53*, 1421–1423.

(43) Guo, Y.; He, H.; Huang, H.; Qiu, J.; Han, J.; Hu, S.; Liu, H.; Zhao, Y.; Wang, P. Solubility Determination and Thermodynamic Modeling of N-Acetylglycine in Different Solvent Systems. *Journal of Chemical & Engineering Data* **2021**, *66*, 1344–1355.

(44) Liu, H.; Wang, S.; Qu, C.; Li, M.; Qu, Y. Solid–Liquid Equilibrium of Chlorpropamide in 14 Pure Solvents at Temperature of 283.15 to 323.15 K. *Journal of Chemical & Engineering Data* **2020**, *65*, 2859–2871.

(45) Liu, M.; Wang, S.; Qu, C.; Zhang, Z.; Qu, Y. Solubility Determination and Thermodynamic Properties of Bezafibrate in Pure and Binary Mixed Solvents. *Journal of Chemical & Engineering Data* **2020**, *65*, 2156–2169.

(46) Yang, W. Thermodynamic Models for Determination of the Solubility of N-Acetylglycine in (Methanol+Acetonitrile) Binary Solvent Mixtures. *Journal of Thermodynamics and Catalysis* **2016**, *7*, 1–6.

(47) Chung, Y.; Gillis, R. J.; Green, W. H. Temperature-dependent vapor–liquid equilibria and solvation free energy estimation from minimal data. *AIChE J.* **2020**, *66*, e16976.

(48) J. Sheehan, R. *Ullmann's Encyclopedia of Industrial Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2011; pp 17–28.

(49) Long, B. W.; Wang, L. S.; Wu, J. S. Solubilities of 1,3-benzenedicarboxylic acid in water + acetic acid solutions. *Journal of Chemical and Engineering Data* **2005**, *50*, 136–137.

(50) Han, N.; Zhu, L.; Wang, L.; Fu, R. Aqueous solubility of m-phthalic acid, o-phthalic acid and p-phthalic acid from 298 to 483 K. *Separation and Purification Technology* **1999**, *16*, 175–180.

(51) Cheng, Y.; Huo, L.; Li, X. Solubilities of isophthalic acid in acetic acid + water solvent mixtures. *Chinese Journal of Chemical Engineering* **2013**, *21*, 754–758.

(52) Lyu, Q.; Zhang, W.; Sun, W.; Xu, Z.; Zhao, L. Determination and correlation of solubility and dissolution thermodynamics of m-toluic acid and isophthalic acid in binary (water plus ethanoic acid) solvent mixtures. *Canadian Journal of Chemical Engineering* **2018**, *96*, 1814–1819.

(53) Li, D. Q.; Evans, D. G.; Duan, X. Solid-liquid equilibria of terephthalaldehydic acid in different solvents. *Journal of Chemical and Engineering Data* **2002**, *47*, 1220–1221.

(54) Sun, W.; Qu, W.; Zhao, L. Solubilities of 4-formylbenzoic acid in ethanoic acid, water, and ethanoic acid/water mixtures with different compositions from (303.2 to 473.2) K. *Journal of Chemical and Engineering Data* **2010**, *55*, 4476–4478.

(55) Ahmadian, S.; Panahi-Azar, V.; Fakhree, M. A.; Acree, W. E.; Jouyban, A. Solubility of phenothiazine in water, ethanol, and propylene glycol at (298.2 to 338.2) K and their

binary and ternary mixtures at 298.2 K. *Journal of Chemical and Engineering Data* **2011**, *56*, 4352–4355.

(56) Karásek, P.; Planeta, J.; Roth, M. Aqueous solubility data for pressurized hot water extraction for solid heterocyclic analogs of anthracene, phenanthrene and fluorene. *Journal of Chromatography A* **2007**, *1140*, 195–204.

(57) Karásek, P.; Planeta, J.; Roth, M. Solubilities of triptycene, 9-phenylanthracene, 9,10-dimethylanthracene, and 2-methylanthracene in pressurized hot water at temperatures from 313 K to the melting point. *Journal of Chemical and Engineering Data* **2008**, *53*, 160–164.

(58) Dohányosová, P.; Dohnal, V.; Fenclová, D. Temperature dependence of aqueous solubility of anthracenes: Accurate determination by a new generator column apparatus. *Fluid Phase Equilibria* **2003**, *214*, 151–167.

(59) Szterner, P.; Kaminski, M.; Zielenkiewicz, A. Vapour pressures, molar enthalpies of sublimation and molar enthalpies of solution in water of five halogenated derivatives of uracil. *Journal of Chemical Thermodynamics* **2002**, *34*, 1005–1012.

(60) Abraham, M. H.; McGowan, J. C. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia* **1987**, *23*, 243–246.

(61) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J.; Blondal, K.; West, R. H.; Goldsmith, C. F.; Green, W. H. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(62) Bell, I. H.; Wronski, J.; Quoilin, S.; Lemort, V. Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library coolprop. *Ind. Eng. Chem. Res.* **2014**, *53*, 2498–2508.

Graphical TOC Entry