Accurately predicting solubility curves via a thermodynamic cycle, machine learning, and solvent ensembles

Emad Al Ibrahim,[†] Nathan Morgan,[†] Simon Müller,[‡] Saikiran Motati,[†] and William H. Green[*,†]

[†]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, 02139, MA, USA.*

[‡]*Institute of Thermal Separation Processes (V8), Hamburg University of Technology, Hamburg, Germany.*

E-mail: whgreen@mit.edu

Abstract

Determining solubilities of organic molecules is critical in various fields such as pharmaceuticals, agrochemicals, and environmental science. Knowing how a solute will dissolve in different solvents and at different temperatures is essential for drug formulation, synthesis, purification, and crystallization. Hard-to-estimate solubility limits currently hinder the design of new processes, making innovation more expensive. We propose a fast and general method for predicting the solubilities of neutral organic molecules in a wide range of solvents and temperatures. Our method uses a thermodynamic fusion cycle to combine machine learning predictions of the activity coefficient, fusion enthalpy, and melting point temperature. This method was tested on a combined dataset with more than 100,000 experimental solubility values, showing better or

comparable performance to competing methods on many solubility benchmarks even at elevated temperatures. We also introduce reference ensembling to leverage all available experimental solubilities for a given solute in estimating its solubility in a different solvent. Reference ensembling is also shown to enhance the robustness of models trained directly on solubility data.

Introduction

The prediction of the solubility of organic molecules is a critical aspect of various fields such as pharmaceuticals,[1] agrochemicals,[2] environmental chemistry,[3] and materials science.[4] Accurate solubility prediction helps in understanding how a compound will behave in different solvents, which is essential for formulating drugs, gauging bioavailability, and developing new materials with desired properties. Solubility influences the efficacy and safety of drugs, since solubility strongly affects drug delivery, absorption, and excretion.[5] Additionally, solubility predictions aid in the design of chemical processes, where solubility determines the ease of purification and separation of compounds and predicts possible precipitation in flow processes.[6] Predicting solubility also supports environmental sciences by estimating the mobility and persistence of chemicals in natural environments.[3] Overall, screening based on reliable estimates of solubility can save time and resources by reducing the need for extensive experimentation, guiding the design and optimization of compounds with improved performance in their intended applications.

Theoretical calculations of solubility are often achieved using thermodynamic cycles.[7] Figure 1 shows two possible routes to go from a solid crystal to the solution phase. The first, called the sublimation cycle, includes the thermodynamics of transforming the solid to a gas through ($\Delta G_{sublimation}$) and then embedding the gaseous molecule into the solvent through ($\Delta G_{solvation}$). The second, called the fusion cycle,[8] follows the thermodynamics of melting the solid crystal (ΔG_{fusion}), and then mixing it into the solvent (ΔG_{mixing}). These thermodynamic cycles are often used to estimate solubility either directly[9] or via simplified relations like the General Solubility Equation.[10]

Solubility estimation includes numerous other techniques, with many applying semi-empirical/fragment-based equations (e.g. Hansen[11]/Hildebrand[12] solubility parameters, NRTL,[13] UNIQUAC,[14] and UNIFAC[15]), or molecular dynamics.[16,17] Quantum chemistry based equilibrium thermodynamics methods like COSMO-RS have also shown great success in the prediction of solvation properties that are related to solubility.[18] The COSMO-RS method

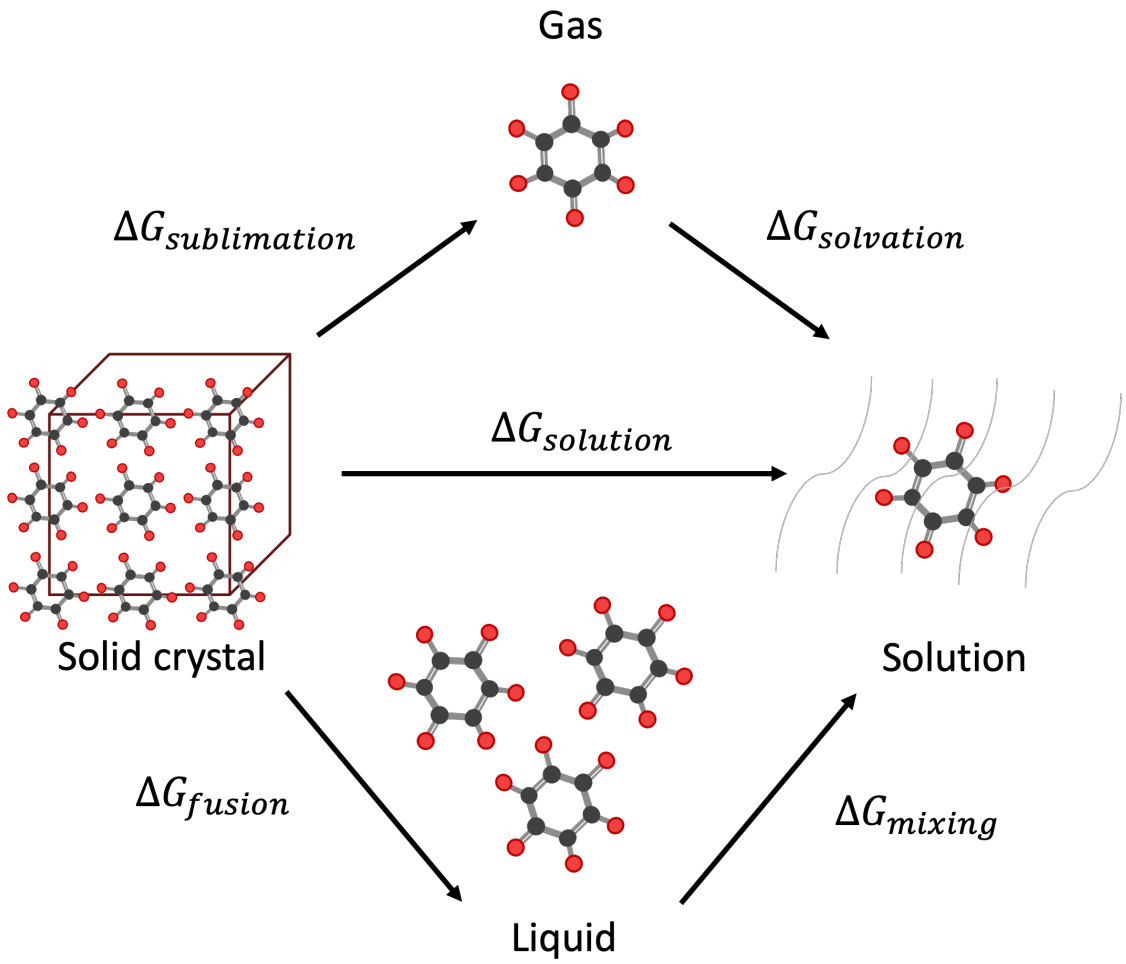


Figure 1: The thermodynamic fusion and sublimation cycles showing the change in free energy at constant temperature upon moving from a pure solid phase to a liquid or a gas to a solution at equilibrium (at a standard state of 1 mol/L).

is considered reliable in its calculations of activity coefficients and solvation energies.[19] However, published solubility predictions that rely on QSPR methods to estimate unknown fusion or sublimation energies are generally less accurate.[20] Moreover, COSMO-RS calculations require potentially expensive and inconvenient geometry optimizations. Another popular approach includes data-based methods like quantitative structure-property relationship (QSPR),[8,21,22] and machine learning techniques.[23–28] While data-based methods can provide good accuracy and flexibility, their extrapolative capabilities are limited by data availability/quality, with some of these methods also lacking thermochemical consistency. This is especially true in light of known solubility data quality problems such as measurement type, polymorphic effects, experimental limit of detection, and experimental variability.[26] These lead to significant aleatoric errors, making it hard to assess the accuracy of direct data-based models.[23]

Recent methods attempt to combine thermodynamic relations with machine learning models.[29,30] For example, the XGB-GSE method[30] relates estimations of the n-octanol-water partition coefficient (logP) from Crippen's method[31] and machine learning predictions of the melting point to aqueous solubility as shown in Equation 1:

$$\log_{10} S_{aq} = 0.5 - 0.01(MP - 25) - \log_{10} P \tag{1}$$

where S_{aq} is the aqueous solubility at 298K in mole/liter, and MP is the melting point in Celsius. This GSE is very popular because of its simplicity and ease of use.[10,30] However, it is limited to aqueous systems at 298K.

SolProp, a framework that uses several directed message-passing neural network models developed in 2022 by Vermeire and collaborators,[29] takes a different approach. SolProp predicts aqueous solubility at 298 K, which is then used as a reference value for other solvents. This is done by using machine learning predictions of solvation energies as shown in Equation 2:

$$\log_{10} S_{target} = \log_{10} S_{ref} - \frac{\Delta G_{solv}^{target} - \Delta G_{solv}^{ref}}{2.303RT} \tag{2}$$

where S_{target} is the solubility in [mol/L] in the target solvent, S_{ref} is the solubility in [mol/L] in the reference solvent (in this case water), ΔG_{solv} is the solvation energy, R is the universal gas constant, and T is the temperature. However, this is limited to predictions at 298K. To compute the temperature dependent solubility S_T, the temperature-dependent dissolution enthalpy $\Delta H_{diss,T}$ must be considered as shown in Equation 3.

$$\ln\left(\frac{S_T}{S_{298K}}\right) = \int_{298K}^{T} \frac{\Delta H_{diss,T}}{RT^2} dT \tag{3}$$

SolProp attempts to calculate the temperature dependence of solubility using two methods, one that accounts for temperature dependence of $\Delta H_{diss,T}$ through numerical integration (Equation 3), and a simpler approximation that neglects it by using ΔH_{solv} and $\Delta H_{sublimation}$ at 298K[29] (i.e. assumes $\Delta H_{diss,T} \approx \Delta H_{diss,298\ K}$).

Our method builds upon the progress of SolProp in using ML to relate sublimation properties to organic solubility and the XGB-GSE method in using ML to relate fusion properties to aqueous solubility. We propose a method to estimate solubility in both aqueous and non-aqueous solvents via a hypothetical supercooled liquid state for the solute. We start by compiling data and training machine learning models to predict the enthalpy of fusion ΔH_{fus}, melting point temperature T_{mp}, and activity coefficients γ_x^T as a function of the temperature and mole fraction. Then, we combine those predictions in Equation 4 (derivation shown in supporting information):

$$S_T \approx \rho_{solvent} x_{sat}^T \approx \frac{\rho_{solvent}}{\gamma_{sat}^T} \exp\left[\frac{\Delta H_{fus}}{R}\left(\frac{1}{T_{mp}} - \frac{1}{T}\right)\right] \tag{4}$$

Equation 4 will be used to iteratively calculate solid solubility based on predictions of ΔH_{fus}, T_{mp}, and γ_x^T coming from three individual ML models, and $\rho_{solvent}$ which is assumed to be known. Equation 4 conveniently models temperature dependence, which is a major

simplification compared to the workflows used in SolProp. If the solute is predicted to be a melt at the given temperature, the liquid solubility (i.e. the solubility of the liquid solute in the solvent-rich phase) will be calculated based on predicted activity coefficients.

The method is then validated on more than 100,000 experimental solubility labels and compared to competing methods (i.e. SolProp, XGB-GSE, and FastSolv) on both organic and aqueous solubility datasets. Reference ensembling is also used, which shows how the use of solubility data of the same solute in multiple reference solvents can enhance the prediction of the solute's solubility in the target solvent, showing the benefit of using multiple references over a single one. Reference ensembling is also used to map solubility predictions from reference to target solvents, enhancing the robustness of ML models trained directly on solubility data.

Methods

Datasets

This section describes the various datasets used for training (enthalpy of fusion, melting point, and activity coefficient), and testing (solubility).

Enthalpy of fusion data

Enthalpy of fusion is the amount of energy required to change one mole of a solid substance into a liquid at its melting point, at constant pressure. Data for fusion enthalpy were collected from the compilations of Acree and Chickos,[32] Yaws,[33] and the CRC.[34] Data points that were marked "unreliable" or that showed "possible dissociation" were removed, and duplicate measurements were averaged. The dataset includes 5,184 enthalpies of unique molecules reported in [kcal/mol], and is referred to as dHfus_DB. The data from Acree and Chickos[32] is approved for public release and shared in the digitally provided files.

Melting point temperature data

At the melting point temperature, the solid and liquid phases exist in equilibrium. The melting point data in this work is comprised of data from OCHEM,[35] PHYSPROP,[36] DrugBank,[37] Coley et al.,[38] and ChemInfo.[39] The combined dataset includes 273,404 melting points of unique molecules reported in [K], and is referred to as Tmp_DB, whereby duplicate measurements were averaged and only points with standard deviations less than 10 were kept. The compiled dataset, from sources that approved public release,[35,36,38] is shared in the digitally provided files.

Activity coefficient data

1. Data from theoretical calculations

COSMO-RS allows flexible calculation of activity coefficients across different conditions such as temperature, mole fraction, and chemical space. Many machine learning activity coefficient models in the literature are trained fully or partially on COSMO-RS data.[40–42] We use the published data from SolvGNN,[42] with a total of 560,000 labels where 80,000 solutes are sampled at (x=0.0,0.1,0.3,0.5,0.7,0.9,1.0), as an initial dataset. However, the dataset is limited to room temperature and is limited in chemical space (spanning random combinations of around 700 commonly used solvents). We extended the dataset by randomly sampling 1000 solutes, from the COSMObase 2023 database containing around 10,000 neutral compounds, for each solvent in their list of commonly used solvents. For each sampled solute-solvent pair, we also randomly sample a temperature in the [200-600 K] range from a uniform distribution and a mole fraction from an exponential distribution (to be more representative of realistic dilution conditions). To expand the diversity of the solutes, we also supplement the dataset with COSMO-RS calculations on ∼200 drug-like molecules listed in the supporting information. For each solvent from SolvGNN,[42] 50 drug-like solutes are sampled from the same temperature and mole fraction distributions to generate a total of 28300 additional data points. Distributions of the additional data sampled in this work are shown in Figures

S1 and S2 of the Supporting Information. The combined dataset of more than 1.1 million data points was used for pre-training and is referred to in this work as gamma_QM_DB.

2. Experimental activity coefficient data

The experimental dataset compiled in Wu et al.[43] is also used in this work and is referred to as gamma_exp_DB. Note that this dataset spans different temperatures but is all close to the infinite dilution limit (i.e. small mole fraction). To remedy this, the experimental data is augmented with data points at mole fraction of one where the activity coefficient equals one by definition (see the plot in Figure 2). An augmented data point of $\gamma(x = 1) = 1$ is added for each data point in gamma_exp_DB at the same solute, solvent, and temperature to create gamma_aug_DB. This is done to limit model drift in the fine-tuning process and ensure that the dependence on mole fraction is not forgotten. The details of the training datasets are summarized in Table 1.

Table 1: Summary of datasets used for training in this work. The columns f(T) and f(x) indicate if data is a function of temperature and mole fraction respectively.

Name	Description	Data entries	f(T)	f(x)	Source
dHfus_DB	Experimental data for enthalpy of fusion	5184 solutes	No	No	Compiled[32–34]
Tmp_DB	Experimental data for melting point temperatures	273404 solutes	No	No	Compiled[35–39]
gamma_QM_DB	Quantum chemical database (COSMO-RS) for activity coefficients	1162300 data 11572 solutes 685 solvents	Yes	Yes	COSMO-RS[18] & SolvGNN[42]
gamma_exp_DB	Experimental data for infinite dilution activity coefficients	21284 data 302 solutes 447 solvents	Yes	No	Wu et al.[43]
gamma_aug_DB	Augmented experimental data for infinite dilution activity coefficients	42568 data 302 solutes 447 solvents	Yes	Yes	Wu et al.[43] & augmented

Solubility data

To test our method, we use various experimental solubility datasets. The datasets we use include popular compilations of solubilities in various solvents and temperatures like BigSolDB,[44,45] CombiSolu-Exp,[29] CombiSolu-HighT-Exp,[29] and the handbook of solubility data for pharmaceuticals by Jouyban.[46] We also include datasets compiled in other machine learn-

ing studies like Bao et. al[28] and Boobier et. al.[27] Moreover, we use the Enabling Technologies Consortium (ETC) collaboration datasets of common solutes in a large variety of solvents around room temperature.[20] Due to the abundance of aqueous solubility data at 298K, we supplement our test datasets with popular aqueous solubility datasets, including drug like molecules (Drug_Aq),[47] solubility challenge data,[48] Ran & Yalkowsky data,[49] AqSolDB,[50] and a small dataset of PROTAC molecules.[51] Using both solubility and melting point data, we were able to find 1067 data points of liquid solubilties (i.e. $T_{mp} < T$) which are discussed separately in the results section. Note that solubility measurements can be intrinsic (maximum concentration of the uncharged compound) or apparent (relative population of dissolved microspecies at the buffer pH).[26] The Drug_Aq, solubility_challenge, and PROTAC datasets are specified as intrinsic, while the rest of the datasets are of unspecified types. Users should be cautious when using such data and keep in mind that most predictive models, including the one presented in this work, are designed for intrinsic solubility. However, when one can estimate the pK_a's, one can predict pH-dependent apparent solubility.[52]

We also provide a combination of all of these datasets, which we refer to as BiggerSolDB. Duplicates are first dropped, and then different solubility values are averaged (where only 208 data points had a standard deviation greater than one log unit). Any zwitterionic molecules are removed at this point. To our knowledge, this combined dataset is the largest public compilation of experimental solubility data with 118978 data points, with 10633 unique solutes and 243 unique solvents (100935 data points after dropping zwitterions). The dataset notably increases the number of available unique neutral solutes in organic solvents to 2200 from BigSolDB's 1160, which is important for approaches that train directly on solubility data.[23,27,28] A summary of the datasets used in this work is shown in Table 2. Data conversions are applied to standardize different units to base-10 logS in [mol/L]. To achieve this, temperature-dependent densities of solvents are used from the DIPPR database.[53] Note that temperature-dependent densities for 554 data points at specific temperatures were not found in the DIPPR database. For those solvents, the density at room temperature was used

if available and a warning label was given in the digitally provided files.

Models and training procedure

All three models in this work were trained using Chemprop v2,[54] a Python implementation of the directed message passing neural network (D-MPNN) architecture for molecular property prediction. The enthalpy of fusion and melting point models are both single-component models, while the activity coefficient model is a multi-component model with temperature and mole fraction supplied as extra data point descriptors.

To tune the models, a hyperparameter search of 30 steps was done on parameters that describe the structure of the model (i.e. depth, ffn_num_layers, dropout, message_hidden_dim, ffn_hidden_dim, max_lr, init_lr, final_lr, warmup_epochs, activation, aggregation, aggregation_norm, and batch_size). A random split of 80/10/10 train/validation/test was used, and training ran for 100 epochs. Due to the large size of the activity coefficient pre-training dataset (gamma_QM_DB), a randomly down-sampled version of 200,000 data points was used for hyperparameter tuning. After tuning, an ensemble of five models were trained for each model using the best configuration suggested by the hyperparameter search. Moreover, the feed forward network layers of the activity coefficient model were fine-tuned using gamma_aug_DB for an additional 20 epochs.

Solubility calculation

Solid solubility

The model predictions are then combined using Equation 4 to calculate solid solubility. Note, however, that the activity coefficient is a function of the mole fraction, while the mole fraction solubility is a function of the activity coefficient. In order to approximate the saturation condition, we start with an initial guess of zero (infinite dilution) and iteratively adjust the mole fraction until convergence (i.e. until the residual becomes sufficiently low or

Table 2: Summary of experimental solubility datasets used for testing in this work. The column f(T) indicates if data is a function of temperature. BiggerSolDB is made of the combination of all other datasets.

Name	Description	Data entries	f(T)
Handbook[46]	Pharmaceutical solubility data [logS]	5254 data 282 solutes 135 solvents	Yes
Bao et al.[28]	Solubility data for ML validation [logS]	3902 data 88 solutes 36 solvents	Yes
BigSolDB[44]	Compiled solubility data [logS]	54173 data 829 solutes 128 solvents	Yes
BigSolDBv2.0[45]	Compiled solubility data [logS] Version 2.0	103390 data 1448 solutes 191 solvents	Yes
CombiSolu-Exp[29]	Compiled solubility data - moderate T [logS]	4953 data 115 solutes 97 solvents	Yes
CombiSolu-Exp-highT[29]	Compiled solubility data - high T [logS]	1306 data 67 solutes 15 solvents	Yes
Boobier_non_Aq[27]	Solubility data for ML validation [logS]	1465 data 1131 solutes 3 solvents	Yes
ETC1[20]	Enabling Technologies Consortium data [logS]	346 data 10 solutes 44 solvents	Yes
ETC2[20]	Enabling Technologies Consortium data [logS]	340 data 15 solutes 42 solvents	No
Drug_Aq[47]	Aqueous drug solubility data [$\log S_{aq}^{298K}$]	72 solutes in water	No
Solubility_Challenge[48]	Aqueous "solubility challenge" data [$\log S_{aq}^{298K}$]	132 solutes in water	No
Ran & Yalkowsky[49]	Aqueous solubility for GSE validation [$\log S_{aq}^{298K}$]	148 solutes in water	No
Boobier_Aq[27]	Aqueous solubility for ML validation [$\log S_{aq}^{298K}$]	900 solutes in water	No
AqSolDB[50]	Compiled aqueous solubility data [$\log S_{aq}^{298K}$]	8613 solutes in water	No
PROTAC[51]	Aqueous "PROTAC" solubility data [$\log S_{aq}^{298K}$]	21 solutes in water	No
BiggerSolDB[20,27–29,44,46–51]	Combination of all neutrals w/o duplicates (no zwitterions)	100935 data 9785 solutes 231 solvents	Yes

the maximum number of iterations is reached). This process is shown visually in Figure 2.

Liquid solubility

If T_{mp} is predicted to be less than T, we solve for the solubility of the liquid solute in the solvent-rich phase by iteratively solving the following equation:

$$x_B^{\alpha}\gamma_B^{\alpha} = x_B^{\beta}\gamma_B^{\beta} \tag{5}$$

where x is the molar fraction, γ is the activity coefficient, B is the solute, α is the solvent-rich phase, and β is the solute-rich phase (the derivation of this equation is shown in the supporting information). We use an initial guess of $x_B^{\alpha} = 0$ and $x_B^{\beta} = 1$ to calculate x_B^{α}, and then update the prediction of γ_B^{α} to calculate x_B^{β}. Again, this is repeated until the residual becomes sufficiently low or the maximum number of iterations is reached.

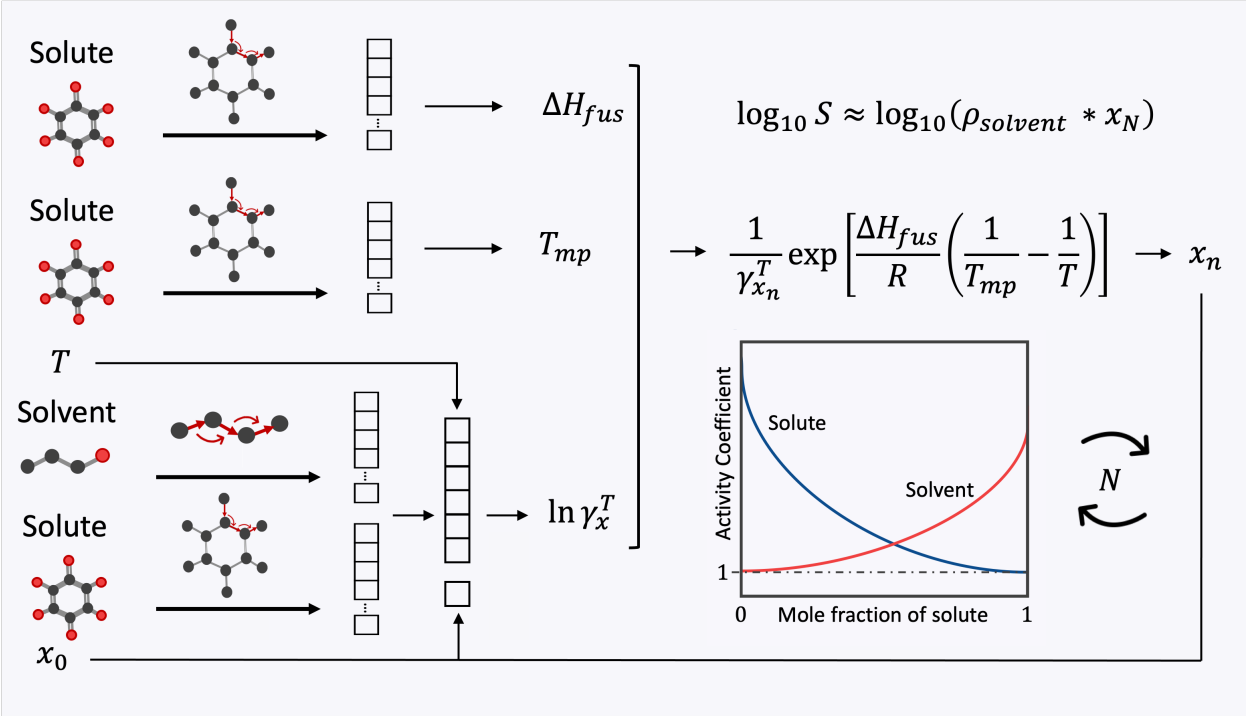
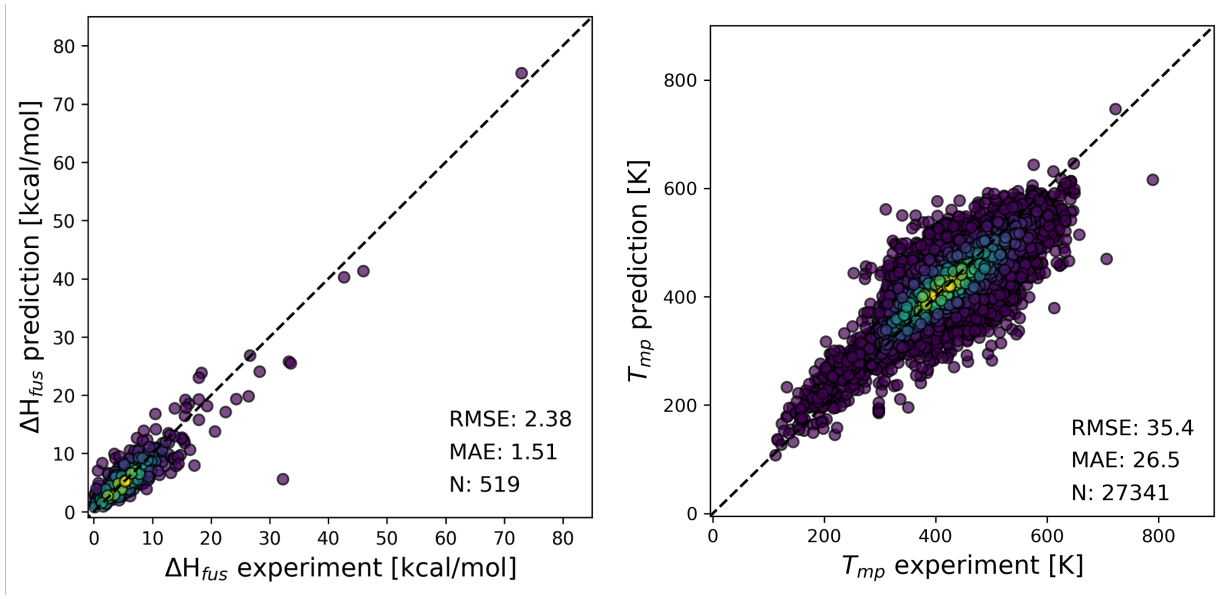


Figure 2: Visual depiction of the iterative process used for the solubility calculation. Infinite dilution is used as an initial guess of the mole fraction solubility (i.e. $x_0 = 0$). The calculated mole fraction x_n is then iteratively used as input to the activity coefficient model until converging to x_N after N iterations.

Results and discussion

Validation of ML model predictions

To validate the ML models, the enthalpy of fusion and melting temperature models were tested on a 10% randomly selected test set from dHfus_DB and Tmp_DB respectively. Figure 3 shows parity plots with mean absolute errors of 1.5 [kcal/mol] and 26.5 [K] for the enthalpy of fusion and melting point temperature respectively. Deviations between the model predictions and the experimental enthalpies of fusion and melting points are reasonable for most molecules given the expected experimental uncertainty, unaccounted polymorphic effects, and human errors in data recording[55] and are comparable to others models reported in literature.[38,55–59]

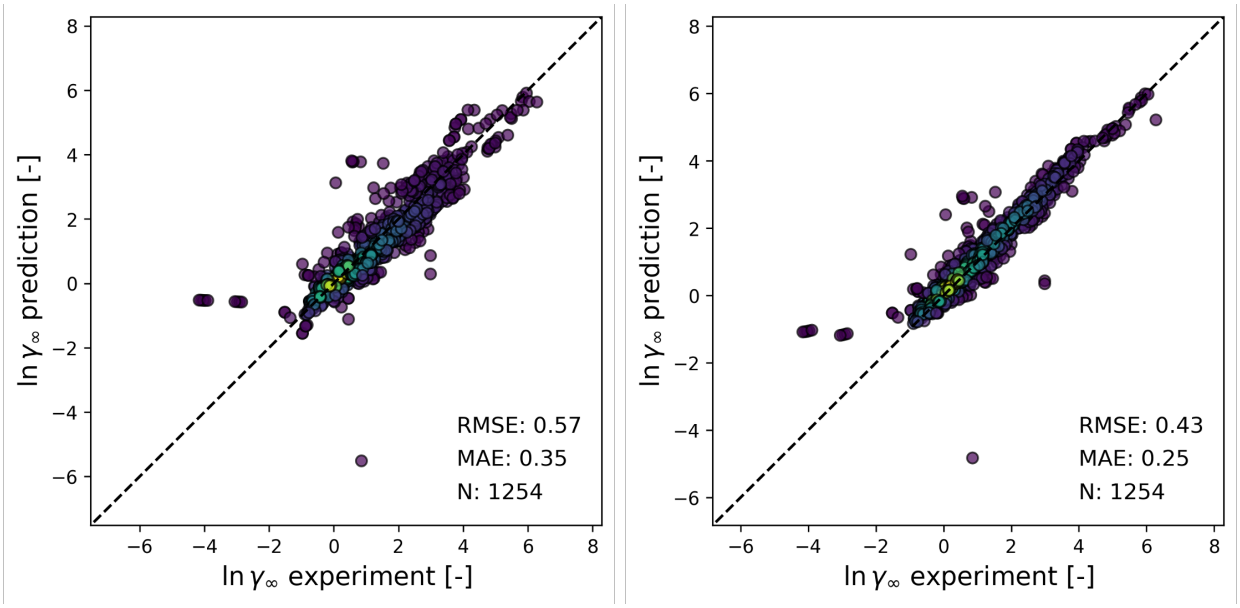


(a) (b)

Figure 3: Parity plot of the prediction on a 10% randomly selected test set of a) the enthalpy of fusion model, and b) the melting point model. Color shows density of the data points.

For the activity coefficient model, a 10% scaffold-split test set (Bemis-Murcko scaffolding on the solutes as implemented in Chemprop[54]) from gamma_exp_DB was used for evaluation. The test set was evaluated using both the pre-trained model that has only seen theoretical

COSMO-RS data from gamma_QM_DB, and the fine-tuned model that was trained on 80% and validated on 10% scaffold splits of the remaining experimental data in gamma_exp_DB. Figure 4 shows the importance of fine-tuning, where the mean absolute error drops from 0.35 to 0.25. This model is then retrained with a 90% train 10% validation splits on gamma_QM_DB for pre-training and using the complete gamma_exp_DB for fine-tuning.



(a) (b)

Figure 4: Parity plot of the prediction on a 10% scaffold-split test set of experimental activity coefficient data using a) pre-trained model, and b) fine-tuned model. Color shows density of the data points.

Testing of solubility predictions

This section assesses the validity of the solubility calculation process described in Figure 2. All calculations in this section ran for a total of ten iterations. A summary of the results are shown in Figure 5 where we compare the fusion cycle calculation to SolProp and the XGB-GSE methods. Note that SolProp was trained on aqueous solubility data at 298K, and so our new method is compared to the XGB-GSE method for the last six datasets instead. Parity plots for the 15 individual datasets is shown in Figure S3 of the supporting information. Our proposed solubility prediction method is competitive, with a performance that usually

matches or exceeds that of the other methods.

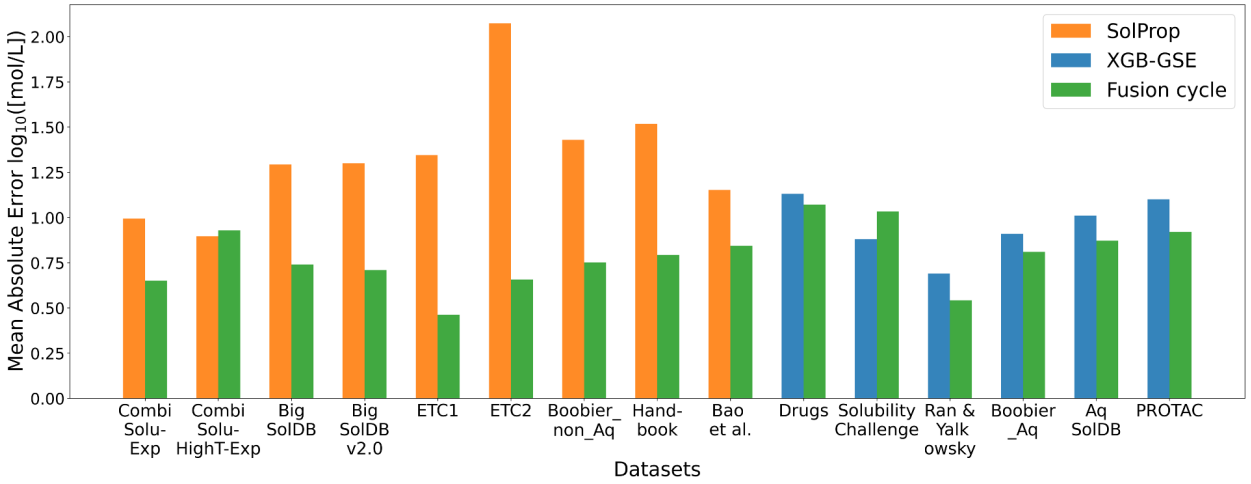


Figure 5: Benchmarking of solubility prediction against SolProp or XGB-GSE on the 15 datasets described in Table 2.

Although the performance of both the fusion cycle approach and SolProp worsens with increasing temperature, the fusion cycle seems to perform much better than SolProp at high temperatures (see model performance above at temperatures above 500K in Figure 6). This is possibly due to SolProp's temperature dependence relying on extra variables like the solvation and sublimation enthalpies.

While the iterative process is a more thermodynamically sound way to estimate saturation solubility, it significantly increases inference time (time increases by a factor of N iterations). The initial guess of infinite dilution is usually a good approximation for a fraction of the time required for the full calculation. Figure 7 compares the saturation solution (from ten iterations) and the infinite dilution solution. It shows that for the majority of data where the mole fraction solubility is less than 0.1, the iterative process does not enhance the solubility estimate. The iterative process seems to only matter at higher solubilities where we cannot assume infinite dilution. Note that for such high solubilities, techniques could be employed to force convergence when oscillating behavior is detected.[60] Such techniques were not implemented in this study for simplicity, but are expected to enhance predictions at high mole fractions. For our data, we observed oscillations of less than 0.1 logS unit so

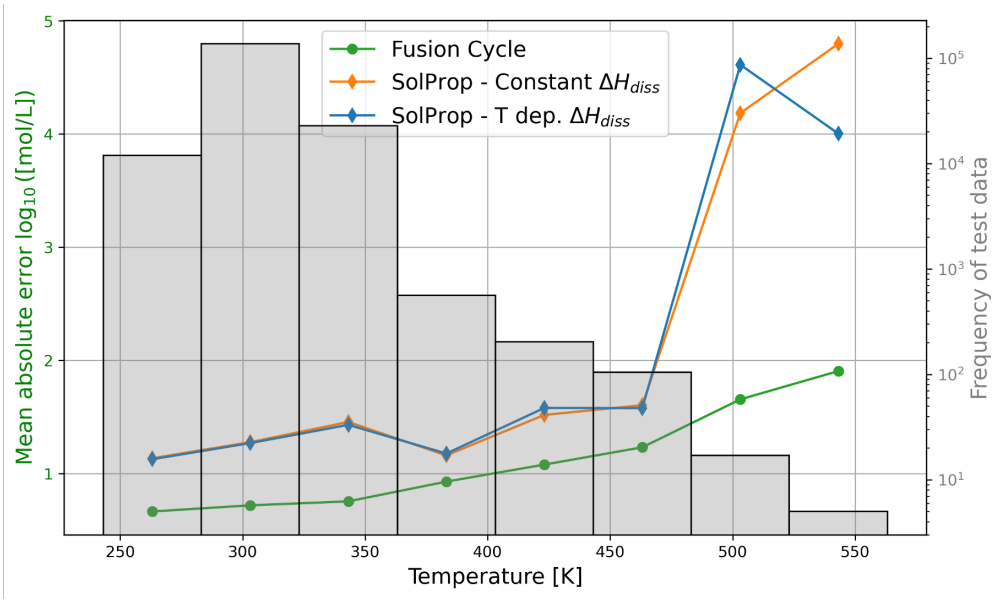


Figure 6: Dependence of mean absolute error in $\log_{10} S_T$ on temperature. Fusion cycle in circles is compared to the two SolProp methods with constant and temperature dependent dissociation enthalpies in diamonds. A histogram showing test data distribution is shown in gray.

this was not a concern.

The performance of the solubility calculation is also evaluated on BiggerSolDB (i.e. the combination of all 15 datasets). This gives a better sense of the fusion cycle's performance and generalizability as it is tested on the largest compiled dataset of more than 100,000 solubility labels. The parity plot in Figure 8.a shows that predictions are well-centered about the identity line. The histogram of the absolute error in Figure 8.b further supports this good agreement, with $\sim 76\%$ of the data predicted within ± 1 log unit and $\sim 94\%$ within ± 2 log units.

Liquid solubility

Note that the results in this section include some melts. As discussed in the methods section, liquid solubility is defined as the concentration of the solute in the solvent-rich phase. To demonstrate the applicability of our method to liquids, we present Figure 9.a, where only liquid solubilities (i.e. $T_{mp} < T$) are plotted against predicted solubility. The method works

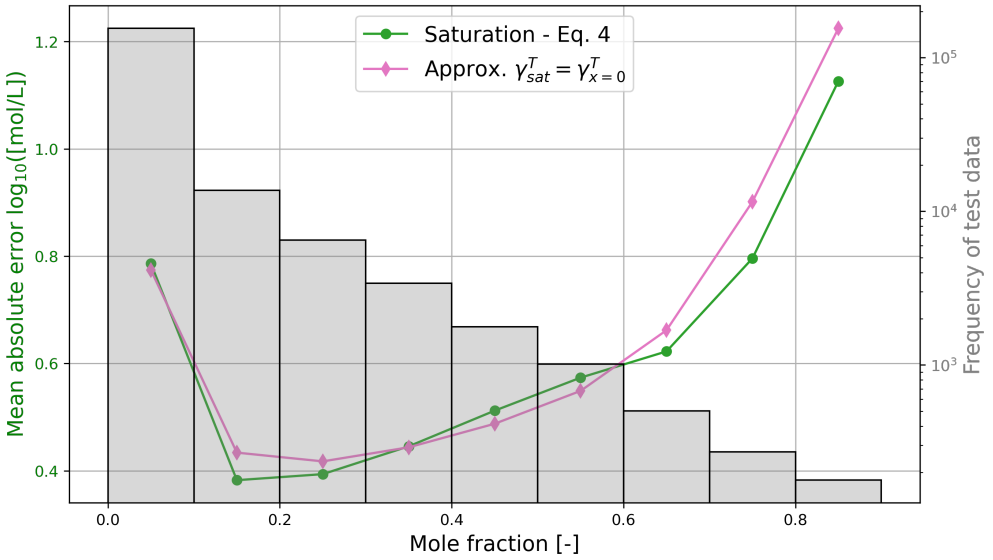
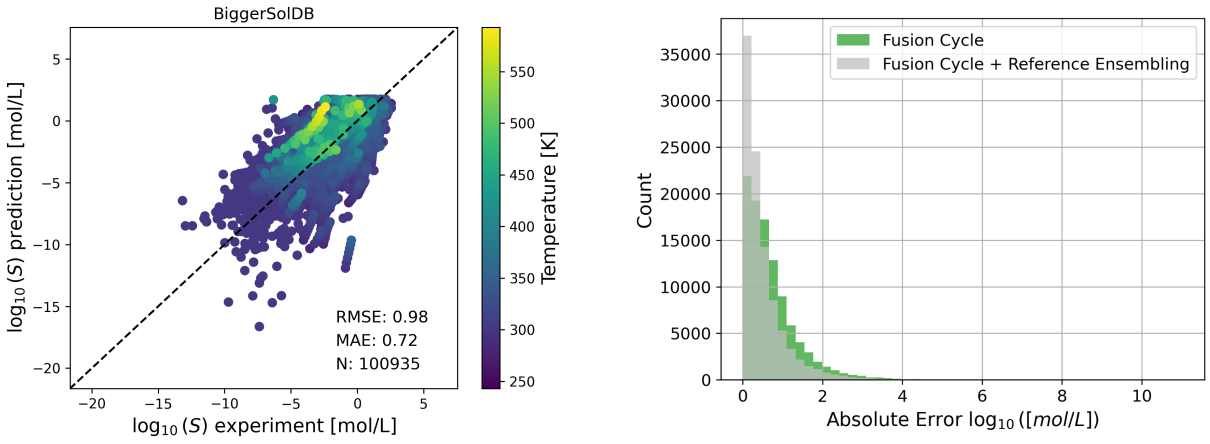


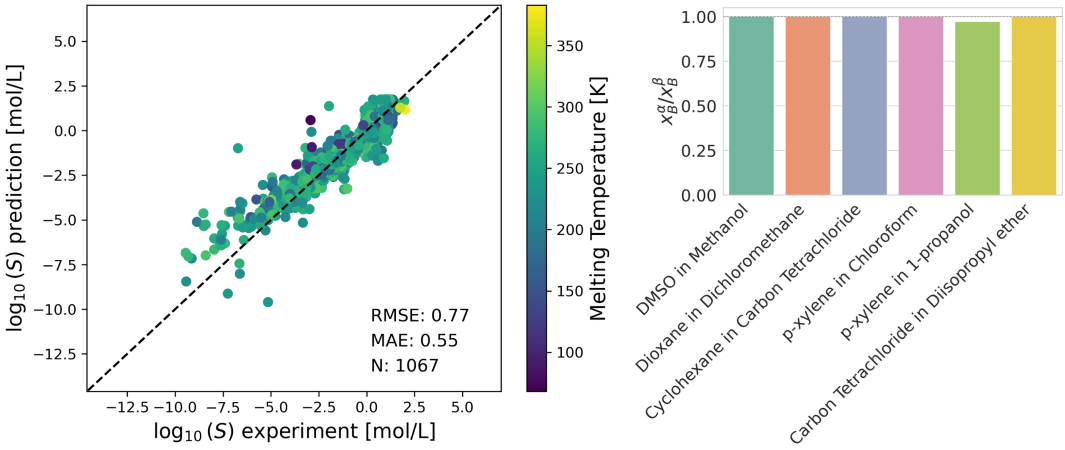
Figure 7: Dependence of the mean absolute error in $\log_{10} S_T$ on mole fraction. The saturation solution from ten iterations is compared to the approximation $\gamma^T_{sat} = \gamma^T_{x=0}$ (i.e. the initial guess). A histogram showing test data distribution is shown in gray.



(a) (b)

Figure 8: a) Parity plots of experimental vs. predicted solubility on BiggerSolDB. b) Histogram of the absolute error between experimental and calculated $\log_{10} S$ on BiggerSolDB. Using an ensemble of reference solvents significantly improves the predictions.

reasonably well for over 1000 liquids, with a mean absolute error of 0.55. The relatively better performance as compared to solids may be attributed to the use of equation 5, which does not depend on the predicted enthalpy of fusion and melting point temperature. Moreover, Figure 9.b shows the calculated x_B^α/x_B^β ratio for some miscible solvents. Miscible liquids mix together completely in all proportions and so only one liquid phase exists (i.e. $x_B^\alpha = x_B^\beta = x_B$ or $x_B^\alpha/x_B^\beta = 1$).



(a) (b)

Figure 9: a) Parity plots of experimental vs. predicted solubility in immiscible liquids (i.e. $T_{mp} < T$). b) Calculation of the x_B^α/x_B^β ratio for miscible solvents using equation 5 (x_B^α/x_B^β=1 if miscible).

Validation on logP data

The n-octanol-water partition coefficient is a partition coefficient for the two-phase system consisting of n-octanol and water. It serves as a measure of the relationship between lipophilicity (fat solubility) and hydrophilicity (water solubility) of a substance. The 'dry' logP at infinite dilution can be written in terms of solubilities as follow:

$$\log P = \log S_{n-octanol} - \log S_{aq} \tag{6}$$

where $\log S_{n-octanol}$ and $\log S_{aq}$ are solubilities at 298K in n-octanol and water respec-

tively. Figure 10 shows a comparison between logP calculated by Crippen's method[31] (an atom contribution method) and calculated via the Fusion Cycle on the corrected OPERA dataset with 12709 data points,[61] and the smaller SAMPL6[62]/SAMPL7[63] datasets. Although Crippen's method is fitted using logP data, the fusion cycle gives consistently better predictions without seeing any logP or logS data in its training (Figure 10). The parity plots of the predictions are shown in Figure S4 of the supporting information.

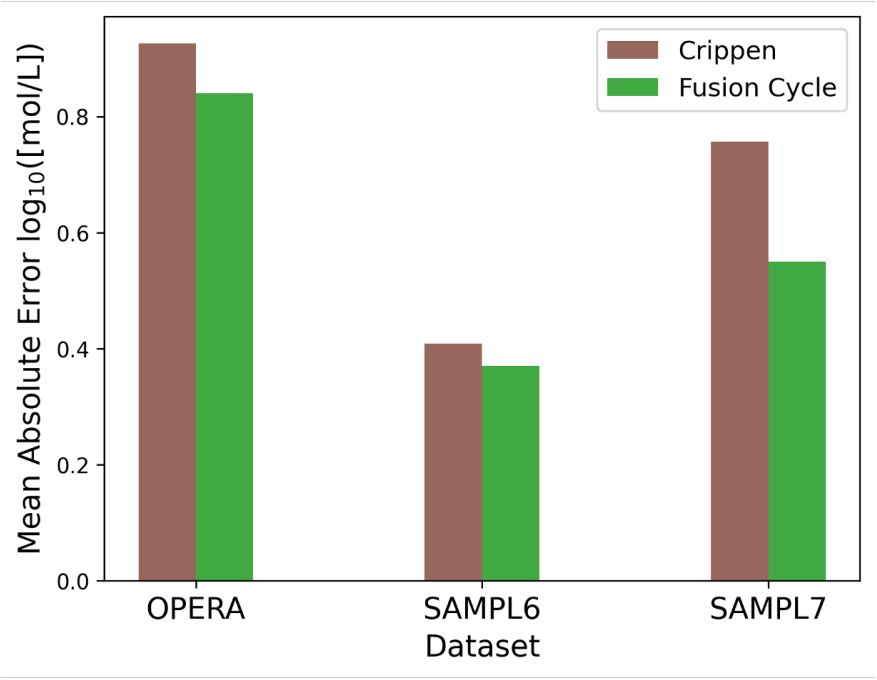


Figure 10: Bar chart comparing predictions using the Fusion Cycle presented here (Eq.6) with Crippen's method for the calculation of logP.

Reference ensembling

An interesting approach shown by Vermeire et al.[29] was the use of experimental solubility of a solute in a reference solvent to infer its solubility in a different target solvent. This circumvented the need to use predicted aqueous solubility as a reference, significantly reducing the associated error. To do this, SolProp used predictions of the solvation free energies at infinite dilution as follows:

$$\log_{10} S_{target}^{298K} = \log_{10} S_{ref}^{298K} - \frac{\Delta G_{solv,298K}^{target} - \Delta G_{solv,298K}^{ref}}{(2.303R)\,(298K)} \tag{7}$$

In this section, we use a similar approach but generalize it to N number of reference solvents and any temperature. First, we use the following relation between solvation energies and activity coefficients (derived in Leenhouts et al.[64] and valid for moderate temperatures and pressures):

$$\Delta G_{solv} = RT \ln \left[\gamma \frac{P_{sat}/RT}{\rho_{solvent}} \right] \tag{8}$$

One then can write the difference in solvation energies as the ratio in activity coefficients:

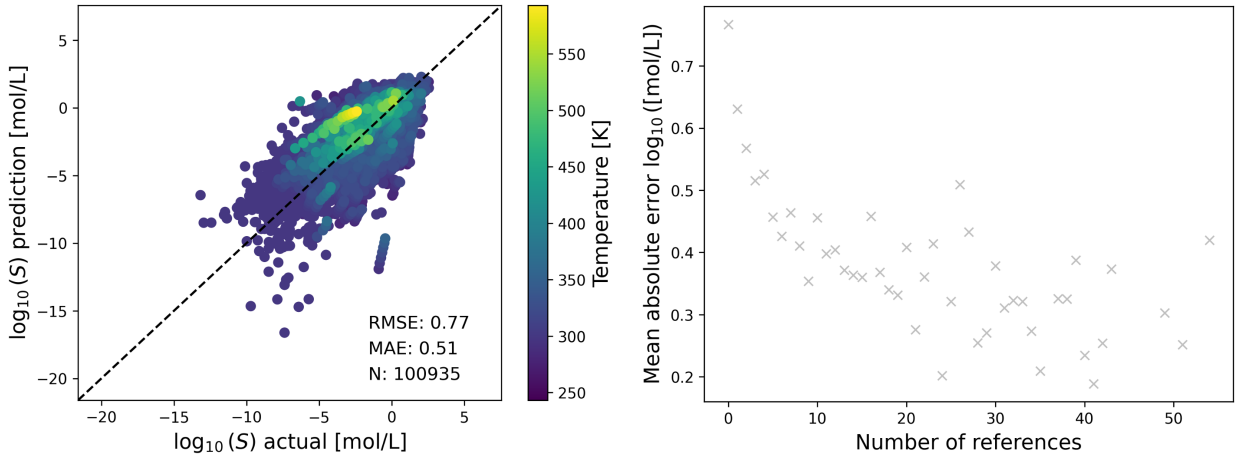
$$\Delta G_{solv}^{target} - \Delta G_{solv}^{ref} = RT \ln \left[\frac{\gamma_{target}\ \rho_{ref}}{\gamma_{ref}\ \rho_{target}} \right] \tag{9}$$

This can then be generalized to N number of references (similar to our previous approach for modeling acid dissociation constants[52]) as follows:

$$\log_{10} S = \frac{1}{N} \sum_i \left(\log_{10} S_{ref,i} - \log_{10} \left(\frac{\gamma_{target}\ \rho_{ref,i}}{\gamma_{ref,i}\ \rho_{target}} \right) \right) \tag{10}$$

While we are capable of estimating γ at the saturation limit, we choose to use the infinite dilution limit in this section since it is less computationally expensive and since most of our data is dilute. The results in Figure 11 show that as the number of references increases, the mean absolute error tends to decrease. This is expected since the error in the reference solubility and the predictions of γ_{ref} is being decreased by averaging N number of estimates. The final root mean squared errors of the fusion cycle and fusion cycle + reference ensembling are 0.98 and 0.77 respectively, which are within the commonly referred to 0.5-1.0 experimental limit of solubility measurements.[23,26] The histogram of the absolute error in Figure 8.b also compares the reference ensembling results to the fusion cycle estimate. The use of reference ensembling increases the predictions within ±1 log units to ∼ 87% and

predictions within ±2 log units to ∼ 97%.



(a) (b)

Figure 11: a) Parity plot of the results from reference ensembling on BiggerSolDB. b) The relation between number of references used and mean absolute error.

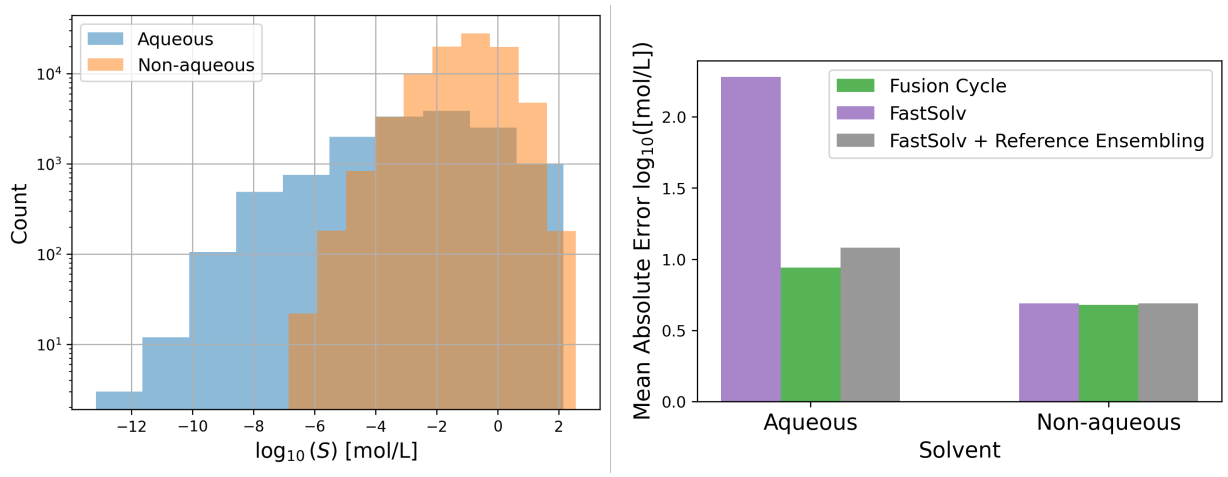
Comparison to FastSolv

This section compares FastSolv,[23] a recent neural network model trained on BigSolDB, to the Fusion Cycle methodology. FastSolv takes Mordred molecular descriptors[65] and temperature as input, and uses a Sobolev[66] loss which incorporates the target temperature derivatives in addition to the target values during training. This section is tested on BiggerSolDB_ext, a subset of BiggerSolDB that includes only solutes that do not appear in BigSolDB. Figure 12 shows that FastSolv matches the performance of the Fusion Cycle on non-aqueous solvents but does much worse on aqueous data.

The reason for this mismatch in performance is the difference in aqueous vs. non-aqueous data distributions, shown in Figure 12.a. Models trained directly on solubility data will always be limited by the training data noise and distribution. While methods relying on thermodynamic relations (e.g. Fusion Cycle, SolProp, and XGB-GSE) also have associated data biases, they seem to retain some physical intuition and perform better in out-of-distribution testing (e.g. generalizing to aqueous data which has a different distribution compared to

that of BigSolDB).

A possible point of synergy between methods that depend on thermodynamic cycles and those that train directly on solubility is the application of reference ensembling. A model that was trained directly on solubility will always perform better on some solvents compared to others due to data imbalance and varying uncertainties of solubility measurements in each solvent. One could use the model to predict solubility in a set of "good" solvents, and use those as reference values in equation 10. In our case, we use the 15 most represented solvents in BigSolDB (FastSolv's training data). Results in Figure 12.b show that this method can greatly enhance the robustness of a model for out-of-distribution solvents (water in this case). This of course comes at additional computational costs (scales with the number of reference solvents per solute) which can make inference time longer.



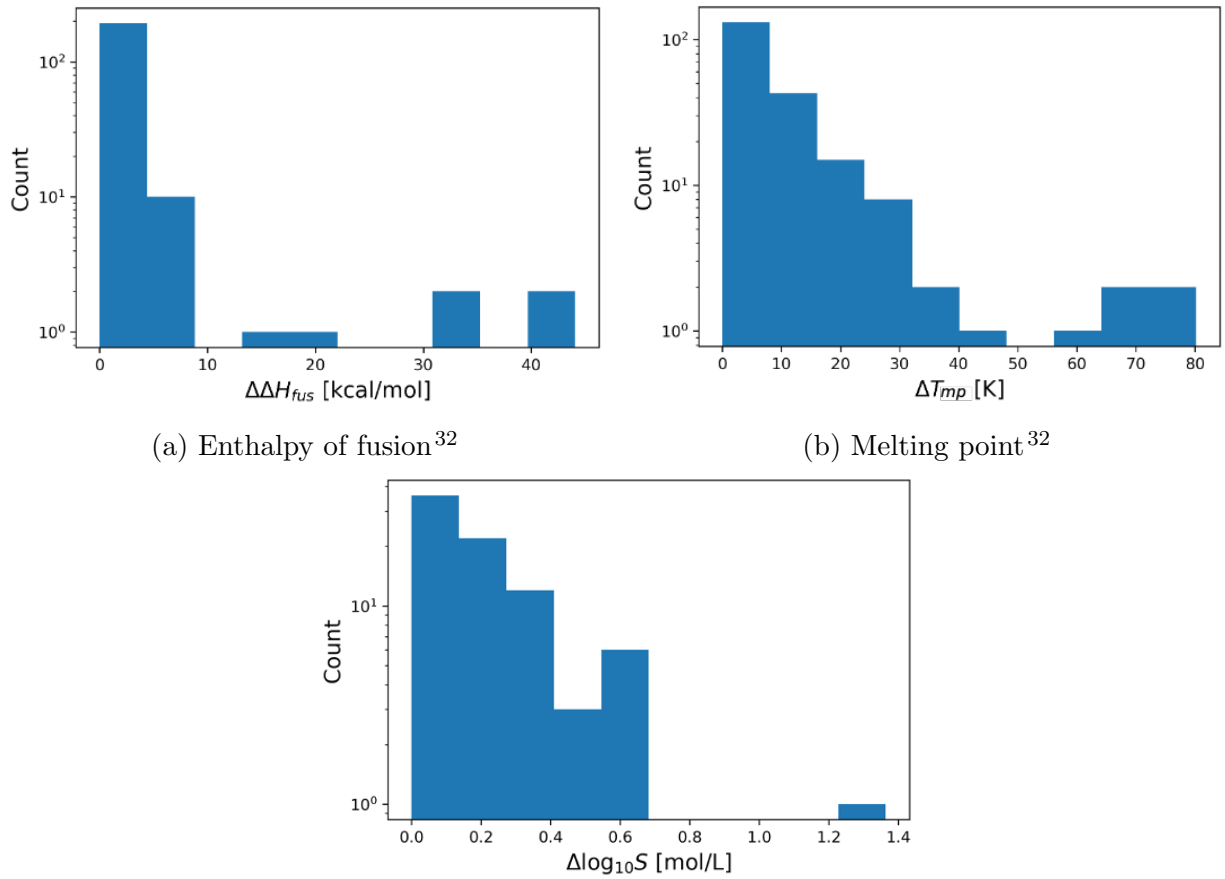
(a) (b)

Figure 12: a) Histogram showing the difference between aqueous and non-aqueous data distributions in BiggerSolDB. b) Mean absolute errors in predicted solubilities by the Fusion Cycle and FastSolv on aqueous and non-aqueous splits of BiggerSolDB_ext. Using reference solvents significantly improves FastSolv's ability to predict aqueous solubilities.

Limitations and future work

An obvious limitation both in this work and in most other methods is failing to account for polymorphic effects. A solid's crystal structure may vary in its packing to form amorphous

or crystalline structures. The more ordered a solid structure is, the higher the enthalpy of fusion; and the lower its solubility. Data on the effect of polymorphism on the properties used in this work are scarce. Polymorphs can have significantly different enthalpy of fusion, melting point temperatures, and solubility as seen in Figure 13. The models presented here are not capable of representing polymorphs, and most of our datasets do not include polymorph information. The use of polymorph-specific properties if known, with models that use that information is expected to give better results.[9,67]



(a) Enthalpy of fusion[32]

(b) Melting point[32]

(c) Solubility[68]

Figure 13: Distributions of molecules exhibiting multiple crystal polymorphs.

Another limitation is the method's inability to calculate solubilities in a mixture of solvents. This stems from the inability of the MPNN used to featurize a mixture of solvents. This can be remedied by using a molecular pooling function like MolPool.[64] This was recently

tested on solvation energies, and is expected to give similarly good performance if expanded to activity coefficients or solubilities. Finally, the methodology presented here is meant only for neutral molecules, and is not recommended for charged molecules, as discussed in the supporting information.

Conclusion

A method is proposed to predict solubilities of various solutes, solvents, and temperatures using a thermodynamic fusion cycle. Training datasets for enthalpy of fusion and melting point are compiled from many different sources and used to train ML models for the two properties. COSMO-RS calculations were generated to supplement existing public datasets of activity coefficients. A model to predict activity coefficients takes in the solute, solvent, temperature, and mole fraction, was then pre-trained on the COSMO-RS data and fine-tuned on augmented infinite dilution experimental values. Predictions from the three models were then used in as inputs to the thermodynamic equation to calculate the solubility of any neutral solute in a wide range of solvents over a wide temperature range. The method was tested on a total of 15 datasets, where it performed better than competing methods on 13 of them.

The combination of the 15 datasets is then used for further testing. The method maintains reasonable predictions at higher temperatures, something that competing methods sometimes struggle with. Reference ensembling is proposed as a method to leverage all available solubility data for a given solute in estimating its solubility in a unseen solvent. This enhances accuracy, dropping the mean absolute error, in $\log_{10}$(solubility), from 0.72 to 0.51 with a jump from 76 to 87% of data points being predicted within ± 1 log units. Furthermore, the method is demonstrated to work for the prediction of n-octanol-water partition coefficients without ever training directly on logS or logP data. Finally, reference ensembling is also used to enhance robustness and out-of-distribution generalization of methods directly

trained on solubility data.

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.

Acknowledgement

The authors acknowledge funding from the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS). E. A. acknowledges the support of the Ibn Rushd Post-doctoral Fellowship Program, administered by the King Abdullah University of Science and Technology (KAUST). Helpful conversations with Jackson Burns, Jonathan Zheng, Florence Vermiere, Roel Leenhouts, and Simona Buzzi are gratefully acknowledged.

Supporting Information Available

The supporting information includes derivations of key equations, details on data distributions, parity plots, and a discussion of error propagation.

Code implementation is available on github: `https://github.com/emadalibrahim/Fusion-Cycle`

Data files are available on Zenodo: `https://doi.org/10.5281/zenodo.16275416`

References

(1) Martínez, F.; Jouyban, A.; Acree Jr, W. E. Pharmaceuticals solubility is still nowadays widely studied everywhere. *Pharmaceutical Sciences* **2017**, *23*, 1.

(2) Khayet, M.; Fernández, V. Estimation of the solubility parameters of model plant surfaces and agrochemicals: a valuable tool for understanding plant surface interactions. *Theoretical biology and medical modelling* **2012**, *9*, 1–21.

(3) Letcher, T. *Thermodynamics, solubility and environmental issues*; Elsevier, 2007.

(4) Anand, S.; Wolverton, C.; Snyder, G. J. Thermodynamic guidelines for maximum solubility. *Chemistry of Materials* **2022**, *34*, 1638–1648.

(5) Faller, B.; Desrayaud, S.; Berghausen, J.; Laisney, M.; Dodd, S. How solubility influences bioavailability. *Solubility Pharm. Chem* **2020**, *113*, 113–132.

(6) Duong, D. T.; Walker, B.; Lin, J.; Kim, C.; Love, J.; Purushothaman, B.; Anthony, J. E.; Nguyen, T.-Q. Molecular solubility and hansen solubility parameters for the analysis of phase separation in bulk heterojunctions. *Journal of Polymer Science Part B: Polymer Physics* **2012**, *50*, 1405–1413.

(7) Fowles, D. J.; Connaughton, B. J.; Carter, J. W.; Mitchell, J. B. O.; Palmer, D. S. Physics-Based Solubility Prediction for Organic Molecules. *Chemical Reviews* **0**, *0*, null, PMID: 40728940.

(8) Abramov, Y. A. Major source of error in QSPR prediction of intrinsic thermodynamic solubility of drugs: solid vs nonsolid state contributions? *Molecular pharmaceutics* **2015**, *12*, 2126–2141.

(9) Mao, C.; Pinal, R.; Morris, K. R. A quantitative model to evaluate solubility relationship of polymorphs from their thermal properties. *Pharmaceutical research* **2005**, *22*, 1149–1157.

(10) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *Journal of pharmaceutical sciences* **2001**, *90*, 234–252.

(11) Hansen, C. M. *Hansen solubility parameters: a user's handbook*; CRC press, 2007.

(12) Hildebrand, J. H.; Scott, R. L. The solubility of nonelectrolytes. *The Journal of Physical Chemistry* **1950**,

(13) Renon, H.; Prausnitz, J. M. Local compositions in thermodynamic excess functions for liquid mixtures. *AIChE journal* **1968**, *14*, 135–144.

(14) Abrams, D. S.; Prausnitz, J. M. Statistical thermodynamics of liquid mixtures: a new expression for the excess Gibbs energy of partly or completely miscible systems. *AIChE journal* **1975**, *21*, 116–128.

(15) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal* **1975**, *21*, 1086–1099.

(16) Li, L.; Totton, T.; Frenkel, D. Computational methodology for solubility prediction: Application to the sparingly soluble solutes. *The Journal of chemical physics* **2017**, *146*.

(17) Boothroyd, S.; Anwar, J. Solubility prediction for a soluble organic molecule via chemical potentials from density of states. *The Journal of Chemical Physics* **2019**, *151*.

(18) Klamt, A.; Eckert, F.; Arlt, W. COSMO-RS: An Alternative to Simulation for Calculating Thermodynamic Properties of Liquid Mixtures. *Annual Review of Chemical and Biomolecular Engineering* **2010**, *1*, 101–122.

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

(20) Lovette, M. A.; Albrecht, J.; Ananthula, R. S.; Ricci, F.; Sangodkar, R.; Shah, M. S.; Tomasi, S. Evaluation of predictive solubility models in pharmaceutical process development an enabling technologies consortium collaboration. *Crystal Growth & Design* **2022**, *22*, 5239–5263.

(21) Yu, X.; Wang, X.; Wang, H.; Li, X.; Gao, J. Prediction of solubility parameters for polymers by a QSPR model. *QSAR & Combinatorial Science* **2006**, *25*, 156–161.

(22) Duchowicz, P. R.; Castro, E. A. QSPR studies on aqueous solubilities of drug-like compounds. *International journal of molecular sciences* **2009**, *10*, 2558–2577.

(23) Attia, L.; Burns, J. W.; Doyle, P. S.; Green, W. H. Organic Solubility Prediction at the Limit of Aleatoric Uncertainty. *In press, Nature Communications* **2025**,

(24) 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.

(25) Ramos, M. C.; White, A. D. Predicting small molecules solubility on endpoint devices using deep ensemble neural networks. *Digital Discovery* **2024**, *3*, 786–795.

(26) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Will we ever be able to accurately predict solubility? *Scientific Data* **2024**, *11*, 303.

(27) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature communications* **2020**, *11*, 5753.

(28) Bao, Z.; Tom, G.; Cheng, A.; Watchorn, J.; Aspuru-Guzik, A.; Allen, C. Towards the prediction of drug solubility in binary solvent mixtures at various temperatures using machine learning. *Journal of Cheminformatics* **2024**, *16*, 117.

(29) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. *Journal of the American Chemical Society* **2022**, *144*, 10785–10797.

(30) Zhu, X.; Polyakov, V. R.; Bajjuri, K.; Hu, H.; Maderna, A.; Tovee, C. A.; Ward, S. C. Building machine learning small molecule melting points and solubility models using CCDC melting points dataset. *Journal of Chemical Information and Modeling* **2023**, *63*, 2948–2959.

(31) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences* **1999**, *39*, 868–873.

(32) Acree, W.; Chickos, J. S. Phase transition enthalpy measurements of organic compounds. An update of sublimation, vaporization, and fusion enthalpies from 2016 to 2021. *Journal of Physical and Chemical Reference Data* **2022**, *51*.

(33) Yaws Carl, L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel, 2003.

(34) Lide, D. R. *CRC handbook of chemistry and physics*; CRC press, 2004; Vol. 85.

(35) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; others Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *Journal of computer-aided molecular design* **2011**, *25*, 533–554.

(36) Mansouri, K.; Grulke, C.; Richard, A.; Judson, R.; Williams, A. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. *SAR and QSAR in Environmental Research* **2016**, *27*, 911–937.

(37) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **2006**, *34*, D668–D672.

(38) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling* **2017**, *57*, 1757–1772.

(39) ChemInfo: Informationssystem Chemikalien des Bundes und der Länder. `https://recherche.chemikalieninfo.de/`, 2025; Accessed: 2025-07-20.

(40) Winter, B.; Winter, C.; Esper, T.; Schilling, J.; Bardow, A. SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients. arXiv (Chemical Physics), Sept 27. *arXiv preprint arXiv:2209.04135* **2022**,

(41) Rittig, J. G.; Felton, K. C.; Lapkin, A. A.; Mitsos, A. Gibbs–Duhem-informed neural networks for binary activity coefficient prediction. *Digital Discovery* **2023**, *2*, 1752–1767.

(42) Qin, S.; Jiang, S.; Li, J.; Balaprakash, P.; Van Lehn, R. C.; Zavala, V. M. Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium. *Digital Discovery* **2023**, *2*, 138–151.

(43) Wu, D.; Zhu, Z.; Zhang, J.; Wen, H.; Jin, S.; Shen, W. An Interpretable Solute–Solvent Interactive Attention Module Intensified Graph-Learning Architecture toward Enhancing the Prediction Accuracy of an Infinite Dilution Activity Coefficient. *Industrial & Engineering Chemistry Research* **2024**, *63*, 8741–8750.

(44) Krasnov, L.; Mikhaylov, S.; Fedorov, M.; Sosnin, S. BigSolDB: Solubility Dataset of Compounds in Organic Solvents and Water in a Wide Range of Temperatures. **2023**,

(45) Krasnov, L.; Malikov, D.; Kiseleva, M.; Tatarin, S.; Sosnin, S.; Bezzubov, S. BigSolDB 2.0, dataset of solubility values for organic compounds in different solvents at various temperatures. *Scientific Data* **2025**, *12*, 1236.

(46) Jouyban, A. *Handbook of solubility data for pharmaceuticals*; CRC press, 2009.

(47) Avdeef, A.; Kansy, M. Predicting solubility of newly-approved drugs (2016–2020) with a simple ABSOLV and GSE (Flexible-Acceptor) consensus model outperforming Random Forest regression. *Journal of Solution Chemistry* **2022**, *51*, 1020–1055.

(48) Llinas, A.; Oprisiu, I.; Avdeef, A. Findings of the second challenge to predict aqueous solubility. *Journal of chemical information and modeling* **2020**, *60*, 4791–4803.

(49) Ran, Y.; Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *Journal of chemical information and computer sciences* **2001**, *41*, 354–357.

(50) 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.

(51) García Jiménez, D.; Rossi Sebastiano, M.; Vallaro, M.; Mileo, V.; Pizzirani, D.; Moretti, E.; Ermondi, G.; Caron, G. Designing soluble PROTACs: strategies and preliminary guidelines. *Journal of Medicinal Chemistry* **2022**, *65*, 12639–12649.

(52) Zheng, J. W.; Al Ibrahim, E.; Kaljurand, I.; Leito, I.; Green, W. H. pKa prediction in non-aqueous solvents. *Journal of Computational Chemistry* **2025**, *46*, e27517.

(53) Thomson, G. H. The DIPPR databases. *International Journal of Thermophysics* **1996**, *17*, 223–232.

(54) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, 9–17, PMID: 38147829.

(55) Sivaraman, G.; Jackson, N. E.; Sanchez-Lengeling, B.; Vázquez-Mayagoitia, Á.; Aspuru-Guzik, A.; Vishwanath, V.; De Pablo, J. J. A machine learning workflow for molecular analysis: application to melting points. *Machine Learning: Science and Technology* **2020**, *1*, 025015.

(56) Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. How accurately can we predict the melting points of

drug-like compounds? *Journal of chemical information and modeling* **2014**, *54*, 3320–3329.

(57) Gharagheizi, F.; Salehi, G. R. Prediction of enthalpy of fusion of pure compounds using an artificial neural network-group contribution method. *Thermochimica acta* **2011**, *521*, 37–40.

(58) Gharagheizi, F.; Gohar, M. R. S.; Vayeghan, M. G. A quantitative structure–property relationship for determination of enthalpy of fusion of pure compounds. *Journal of thermal analysis and calorimetry* **2012**, *109*, 501–506.

(59) Leenhouts, R.; Jankelevitch, S.; Raike, R.; Müller, S.; Vermeire, F. Thermodynamics-informed Graph Neural Networks for Phase Transition Enthalpies. Proceedings of ESCAPE 35, Systems and Control Transactions, Vol.4. Ghent, Belgium, 2025; pp 1662–1669, LAPSE:2025.0419 (living archive), Creative CommonsCCBY-SA4.0.

(60) COSMOtherm Reference Manual. 2016; COSMOlogic GmbH & Co KG, Version C3.0 Release 17.01.

(61) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of cheminformatics* **2018**, *10*, 1–19.

(62) Loschen, C.; Reinisch, J.; Klamt, A. COSMO-RS based predictions for the SAMPL6 logP challenge. *Journal of computer-aided molecular design* **2020**, *34*, 385–392.

(63) Bergazin, T. D.; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L. Evaluation of log P, p K a, and log D predictions from the SAMPL7 blind challenge. *Journal of computer-aided molecular design* **2021**, *35*, 771–802.

(64) Leenhouts, R. J.; Morgan, N.; Al Ibrahim, E.; Green, W. H.; Vermeire, F. H. Pooling solvent mixtures for solvation free energy predictions. *Chemical Engineering Journal* **2025**, 162232.

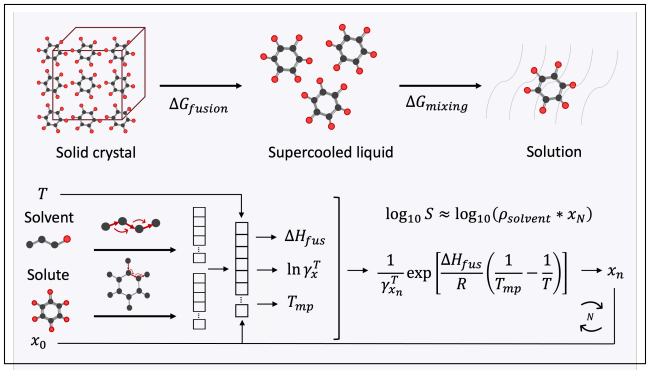
(65) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of cheminformatics* **2018**, *10*, 1–14.

(66) Czarnecki, W. M.; Osindero, S.; Jaderberg, M.; Swirszcz, G.; Pascanu, R. Sobolev training for neural networks. *Advances in neural information processing systems* **2017**, *30*.

(67) Gu, C.-H.; Grant, D. J. Estimating the relative stability of polymorphs and hydrates from heats of solution and solubility data. *Journal of pharmaceutical sciences* **2001**, *90*, 1277–1287.

(68) Pudipeddi, M.; Serajuddin, A. T. Trends in solubility of polymorphs. *Journal of pharmaceutical sciences* **2005**, *94*, 929–939.

TOC Graphic



ML models are used to iteratively calculate solubility through the fusion cycle.