Designing solvent systems in chemical processes using self-evolving solubility databases and graph neural networks

Yeonjoon Kim, Hojin Jung, Sabari Kumar, Alex Claiborne, Robert S. Paton, Seonah Kim*

Department of Chemistry, Colorado State University, Fort Collins, CO 80523, United States
*Corresponding author. seonah.kim@colostate.edu

Abstract

Designing solvent systems is the key to achieving the facile synthesis and separation of desired products from chemical processes. In this regard, many machine-learning models have been developed to predict the solubilities of given solute-solvent pairs. However, breakthroughs in developing predictive models for solubility are needed, which can be accomplished through a remarkable expansion and integration of experimental and computational solubility databases. To maximize predictive accuracy, these two databases should not be separately trained when developing ML models. In addition, they should not be simply combined without reconciling the discrepancies between different magnitudes of errors and uncertainties. Here, we introduce self-evolving solubility databases and graph neural networks developed through semi-supervised self-training approaches. Solubilities from quantum-mechanical calculations are referred to during semi-supervised learning, but they are not directly added to the database. Such methodologies enable the augmentation of databases while correcting the discrepancy between experiments and computation and improving the predictive accuracy against experimental solubilities. The resulting model was successfully applied to two practical examples relevant to solvent selection in organic reactions and separation processes: (i) linear relationship between reaction rates and solvation free energy for three organic reactions, and (ii) partition coefficients for lignin-derived monomers and drug-like molecules.

Introduction

Solubility has been touted as the key molecular property to consider in designing various chemical reactions and processes. It provides the control of reactivity, catalytic activity, separation ability, and other molecular properties. In chemical synthesis, solvent selection controls the solubilities of chemical species involved in reactions and determines their catalytic activity and product selectivity. It is one of the crucial factors in designing homogeneous catalytic reactions pertinent to pharmaceutical synthesis in the solution phase, such as the functionalization of organic molecules through C-H activation.[1-6] In this regard, linear relationships have been elucidated between solvent properties (permittivity, polarity, etc.) and stability of reactants/products, and thus reaction rates for various organic reactions in different solvents.[7-9] Such linear solvation energy relationships (LSERs) inform the solvent selection, leading to the maximal yield of target products.

In the pharmaceutical industry, solubilities in water and organic solvents are essential properties to consider during the entire process development, including screening and synthesis of drug candidates.[10, 11] The candidates having sufficient water solubility should be identified to achieve high bioavailability in oral administration.[12] Water solubility is also relevant to the toxic effects of drugs and pesticides on human health and the environment.[13-15] Solubilities in organic solvents have to be measured as well as water solubilities, especially for assessing *in vivo* efficacy and safety of intravenous drugs dissolved in non-toxic organic solvents.[11, 16, 17] Specifically, solubilities of drug-like molecules in chloroform and diethyl ether have been investigated for the simplified modeling of the polar environment around proteins, and membranes.[18, 19] In addition, solubility plays a critical role in emerging research areas to confront the challenges of climate change, such as sustainable chemistry and renewable energy. For instance, the solvent selection is conducted in biomass upgrading to biofuels and renewable polymers to maximize catalytic activity.[20-22] The optimal water-organic solvent systems enhance not only the conversion to target products but also their extraction from separation processes.[20, 21] Meanwhile, developing organic redox flow batteries is another promising research area for renewable energy storage, and it is important to design electrolytes highly soluble in water or organic solvents for high charge densities.[23-25]

To date, the solubilities of various solutes in water and organic solvents have been measured experimentally, and databases of experimental solubilities have been released. The available databases include AqSolDB,[26] Open Notebook Scientific Challenge,[27] Minnesota Solvation Database,[28-30] FreeSolv,[31] CompSol,[32] and solubility challenge database.[12, 33, 34] Many computational methods have also been developed, enabling *in silico* screening of solvents and solutes through solubility prediction before experiments. Such methods include quantum mechanics (QM) or density functional theory (DFT) with implicit solvation models (e.g., Solvation Model based on Density - SMD),[35] molecular dynamics (MD) simulations, or QM-based thermodynamic equilibrium methods, e.g., the Conductor-like Screening Model (COSMO).[36-38] For more rapid and accurate solubility predictions, various predictive models have been actively developed by analyzing quantitative structure-property relationship (QSPR)[34, 39-44] or adopting machine learning (ML) techniques.[34, 42, 45-55] Particularly, current advanced ML models used graph neural networks (GNNs) combined with interaction layers[47, 53] recurrent neural networks with attention layers,[45] and natural language processing-based transformers.[54] These models achieved accuracies close to experimental uncertainties. Furthermore, the development of ML models has been expanded to the prediction of solubility limits at different temperatures,[52] solvation enthalpy, LSER, and solute parameters,[51] and generative models for designing molecules having optimal aqueous solubility.[55]

Despite the dramatic advancement discussed above, further improvement is needed to accomplish accurate solubility predictions for the broader chemical space of solvents and solutes. There are around 10,000 data points of Gibbs solvation free energies (ΔG_{solv}) in the current largest experimental database, but more data points (around >100,000) would be desirable for training reliable GNNs.[53, 56] In this respect, there have been attempts for pre-training against computational databases followed by transferring the trained model and re-training against the experimental data.[53, 57] Employing such transfer learning approaches is advantageous in utilizing the extensive computational database and refining the model by correcting the discrepancies between theory and experiment. However, transfer learning can diminish the prediction accuracy of the extensive pre-trained computational database after the model is re-trained against the small experimental database. A comprehensive and theory-experiment integrated database would provide another opportunity to accomplish balanced accuracy simultaneously for the chemical space covered by both experiments and computations.

To build an integrated database, discrepancies between theoretical and experimental solubilities should be rectified. In other words, computational solubilities should have a fidelity as high as experimental ones. Accuracies of computational methods depend on the molecule size, constituent elements, functional groups, etc. Therefore, it is not feasible to merely combine experimental and computational databases and train the model. Each database has a different source and magnitude of errors and uncertainties,[58-61] which would deteriorate the accuracy of predictive models. For reliable integration of databases from different sources, state-of-the-art techniques for data augmentation and self-training have been developed, such as noisy student self-distillation and semi-supervised distillation (SSD). The overall procedure of these approaches is as follows; first, the 'Teacher' model is trained against the small but reliable database. Second, predictions are carried out for larger data, creating a new database. Third, the 'Student' model is trained using the database combining the initial database and that from the prediction of the 'Teacher', with or without introducing noise to the model. This procedure is iterated for the gradual addition of reliable data points to the integrated database. These methods have been successfully applied to various ML predictive models for image classification,[62, 63] natural language processing,[64] and protein structures.[65]

In this contribution, SSD was introduced to GNN predictions of solubilities, leading to an augmented database and accurate predictive model encompassing broader chemical space than that covered by experimental measurements. The solute-solvent pairs for the data augmentation were obtained from the **CombiSolv-QM** database, the largest existing database of ΔG_{solv} calculated using COSMO-RS.[53] For reliable data integration and distillation, we referred to the solubilities calculated using COSMO-RS and M06-2X with SMD implicit solvation model, but these values were not included in the database. Instead, ΔG_{solv} values refined through SSD were considered in model development to correct the discrepancies between the experiment and theory. It was found that the databases augmented from SSD enhance the accuracy for predicting experimental solubilities, manifesting the effectiveness of our approach.

Moreover, we successfully applied our model to two practical examples related to solvent system design in reaction kinetics and separation. First, the linear relationship was elucidated between ΔG_{solv} of reactants/products and reaction rates for five chemical reactions. Second, 370 water-organic partition coefficients were predicted for 30 lignin-derived monomers and 17 drug-like molecules and compared with experimental values. These examples demonstrate the potential of our ML approaches in enabling the chemistry-informed design of solvent systems.

Results and Discussion

Graph neural networks and quantum-mechanical methods for model development.

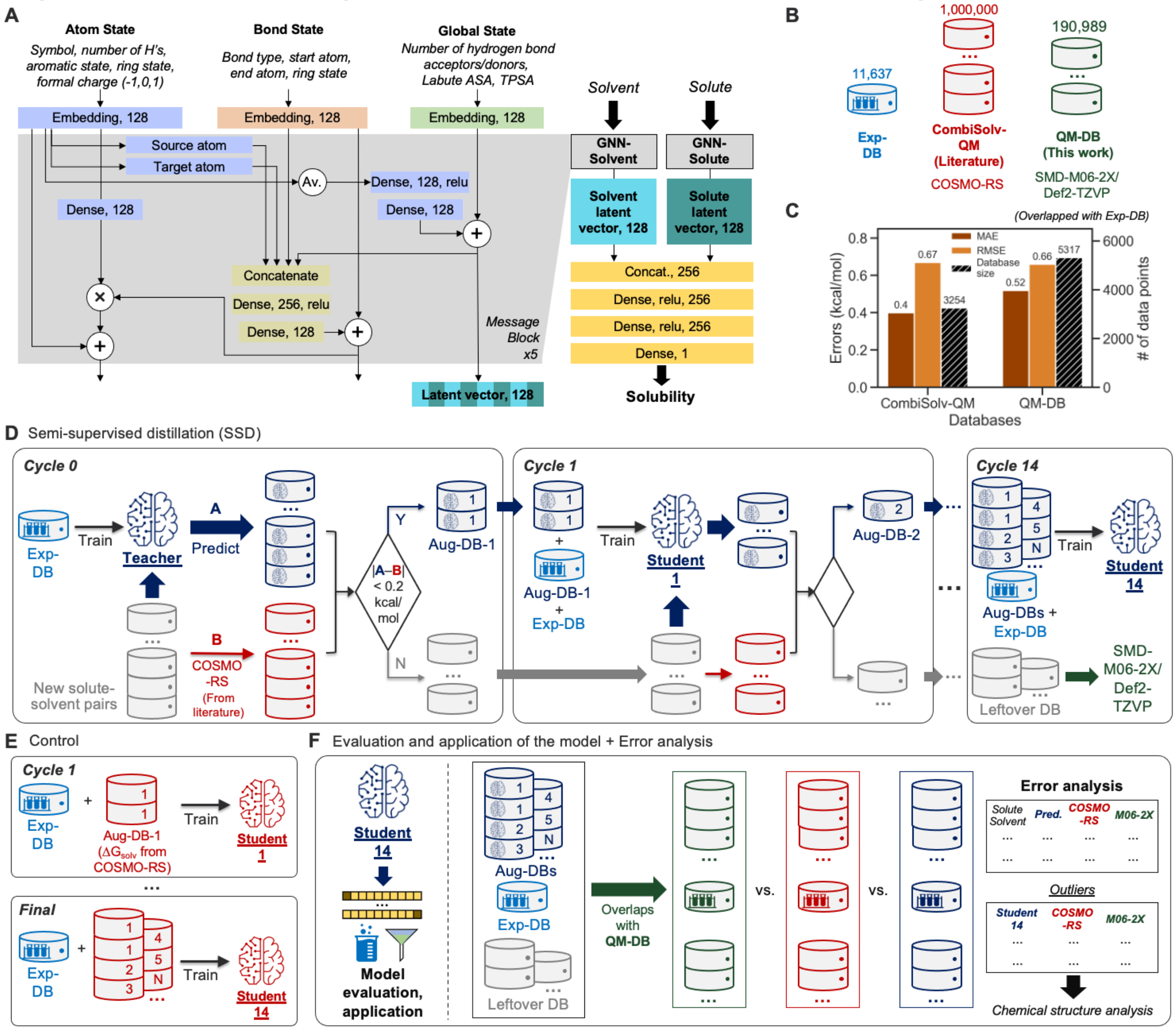


Figure 1. (A) Architecture of the graph neural network for solubility. (B) Description of three databases used to evaluate theoretical methods against experimental solubilities. (C) Comparison of accuracies of **CombiSolv-QM** and **QM-DB** for the data points overlapping with Exp-DB. (D) Semi-supervised distillation (SSD) for self-evolving solubility databases and graph neural networks. (E) Control for comparing the accuracies of models with and without SSD. (F) A schematic description of evaluation, application, and error analysis of the model obtained from SSD.

To execute data augmentation and self-training, first, a GNN was constructed, as shown in Fig. 1A. The model takes 2D molecular structures (SMILES strings) of solvent and solute as inputs, and each of them undergoes a message passing GNN. The overall architecture of two GNNs (GNN-Solvent and GNN-Solute) is similar to our previous GNNs for predicting bond dissociation enthalpy and cetane number.[56, 66] It consists of three blocks representing the atom, bond, and global state of a molecule. Initial atom, bond, global features are embedded as 128-dimensional vectors and pass through five message-passing layers. In each layer, mathematical operations among feature vectors lead to their mutual updates so that the model captures implications regarding the influence of local atom/bond environments and global molecular structures on solubility. Each GNN then outputs a 128-dimensional latent vector for solvent and solute, respectively. These two vectors are concatenated and undergo additional dense layers to take solute-solvent interactions into account, and finally, ΔG_{solv} is predicted. Other operations as well as concatenation have also been reported in previous studies to consider molecular interactions, such as global convolution among molecules and graph-of-graphs neural networks.[67, 68] However, the concatenation of latent vectors was sufficient to achieve accuracy close to experimental uncertainty: mean absolute error (MAE) of ΔG_{solv} around 0.2 kcal/mol (*vide infra*).

The GNN shown in Fig. 1A was inspired by the recent state-of-the-art GNN model for ΔG_{solv} developed by Vermeire *et al.*,[53] but it has differences as follows. First, we have attempted to minimize the number of atom and bond features, leading to a fewer number of atom and bond features than their model. Second, the dimensions of hidden layers were also minimized while maintaining accuracy. Our GNN has hidden layers with 128 and 256 nodes before and after concatenation, respectively, whereas they used 200 and 500-dimensional hidden layers. Third, a separate global state block was built in our GNN, and it participated in feature updates while they concatenated global features after undergoing the message-passing layers. We selected four global features after testing various molecular descriptors; two surface area descriptors were utilized in Vermeire et al.,[53] and two hydrogen bond descriptors were adopted in our predictive model for cetane number.[66] Of note, accuracies comparable to Vermeire et al. were still achieved (Details in the next section) after the hyperparameter tuning, truncation, and modification of the model explained above.

Next, we evaluated QM methods that will provide reference solubility values during the data augmentation using SSD by comparing experimental and calculated ΔG_{solv}. Experimental ΔG_{solv} values were collected from various data sources, and they were curated, resulting in **Exp-DB** consisting of 11,637 data points.(Fig. 1B) Most data points in **Exp-DB** overlap with those in **CombiSolv-Exp**,[53] but it has additional 1,419 data points accounting for 'self-solvation' where the solvent and solute are identical. COSMO-RS and SMD-M06-2X/Def2-TZVP were then benchmarked against **Exp-DB**. To assess COSMO-RS, we adopted **CombiSolv-QM**, the most extensive ΔG_{solv} database consisting of one million data points obtained from COSMO-RS calculations.[53] SMD-M06-2X/Def2-TZVP was elected among plenty of theoretical methods since it provided reliable results from calculations of molecular properties pertinent to solvation. For example, it showed the best accuracy in evaluating the redox potentials of 174 organic molecules in water and acetonitrile among 33 different combinations of density functionals, basis sets, and solvation models.[25] In this work, a new database (**QM-DB**) was built by calculating ΔG_{solv} for 190,989 solute-solvent pairs in Exp-DB and **CombiSolv-QM**. Not all pairs were calculated due to the limited availability of SMD solvent parameters (dielectric constant, refractive index, surface tension, Abraham hydrogen bond acidity, and basicity).

Fig. 1C compares the number of solute-solvent pairs in **CombiSolv-QM** and **QM-DB** that are overlapped with **Exp-DB** and their MAEs and root-mean-square errors (RMSEs) against **Exp-DB**. There are 3,254 common solute-solvent pairs in **CombiSolv-QM** which show an MAE and RMSE of 0.4 and 0.67 kcal/mol with respect to **Exp-DB**. Meanwhile, an RMSE comparable to **CombiSolv-QM** (0.66 kcal/mol) was observed from **QM-DB** with more overlapped data points (5,317). These results manifest the reliability of M06-2X in providing further explanation regarding the errors of QM methods and ML models after the model development (Details in the next section).

Self-training graph neural networks based on semi-supervised distillation and data augmentation.

Building the GNN model and databases was followed by training the model based on SSD (Fig. 1D). The SSD is initiated by training the 'teacher' model using **Exp-DB** (Cycle 0). The trained model is then used for augmenting the database; new solute-solvent pairs are gathered from **CombiSolv-QM**, and their ΔG_{solv} is predicted using the 'teacher' model. The predicted values are compared with COSMO-RS solubilities stored in **CombiSolv-QM**. If the absolute difference between these two is below 0.2 kcal/mol, the corresponding data points are stored in the augmented database (**Aug-DB-1**) with teacher-predicted solubility values. It should be emphasized that the values from ML prediction are saved instead of those from COSMO-RS. This is for refining data points based on the solubility trends learned from **Exp-DB** while maintaining reliability by referring to QM solubility values. The threshold value was set to 0.2 because the uncertainty of experimental measurements of ΔG_{solv} is typically up to 0.2 kcal/mol.[58-61] If the deviation between ML and QM is below 0.2, it can be assumed that the difference is mainly from experimental uncertainty, and the prediction from 'teacher' is credible.

Next, the 'Student 1' model is trained using the database combining **Aug-DB-1** and **Exp-DB** (Cycle 1), and the same procedure is carried out for the solute-solvent pairs that remained after extracting **Aug-DB-1**. 'Student 1' performs ΔG_{solv} prediction for the remaining ones, and the predicted values are subject to the 0.2 kcal/mol cutoff, resulting in **Aug-DB-2**. These cycles were repeated 14 times, and such iterations enabled the self-training of ML models. The database is grown gradually, and subsequent student models learn larger databases that contain ΔG_{solv} values refined based on the guidance from previous student models and COSMO-RS solubilities. We found that such gradual integration shows better accuracy than using the Teacher-predicted values for the whole **CombiSolv-QM** and re-training at once. This is because the model should be slowly trained so that it can steadily transmit the trend it learned from **Exp-DB** while minimizing the discrepancy between experiments and theory. It should be noted that no trained weights of the GNN model were transferred from the previous cycle when training the Student model in the current cycle. Only **Aug-DB**s and **Exp-DB** are transferred, and each student is trained from scratch. This is to verify that the new **Aug-DB** is integrated well with the databases cumulated from previous cycles, and it shows no significant discrepancies and anomalies during the training.

Ultimately, the 14th cycle yields the 'Student 14' model and the integrated database containing **Exp-DB** and 14 **Aug-DB**s. The cycle was terminated at the 14th cycle because the MAE for the test set of **Exp-DB** significantly increased (Detailed results in Fig. 2A, *vide infra*). This stopping criterion was applied since the leftover data points in **CombiSolv-QM** no longer synchronized well with the large **Aug-DB**s cumulated during previous cycles. The solute-solvent pairs not included in **Aug-DB**s were stored in **Leftover DB**. Accuracies of the Student models from SSD were compared with those from the control models trained by the database simply combining ΔG_{solv} values from experiments and COSMO-RS (Fig. 1E).

The resulting Student 14 model was then subject to subsequent evaluation, error analysis, and applications (Fig. 1F). To evaluate the model's accuracy, mean absolute errors (MAEs) and distributions of errors were investigated. For additional error analysis, we obtained the solute-solvent pairs in **QM-DB** that overlap with those in other databases (**Aug-DB**s, **Exp-DB**, **Leftover DB**). Next, we compared their ΔG_{solv} values acquired from four different sources: Experiments (if available), predictions from Student 14, SMD-M06-2X/Def2-TZVP, and COSMO-RS calculations. Outliers were identified from this comparison, and their chemical structures were analyzed to assess the strengths and weaknesses of each QM method or ML model. Also, the model was applied to two practical examples of solvent selections in chemistry: (i) Elucidation of the relationship between reaction rate and ΔG_{solv}, (ii) partition coefficients of lignin-derived monomers and drug-like molecules. Detailed results are discussed in the following sections.

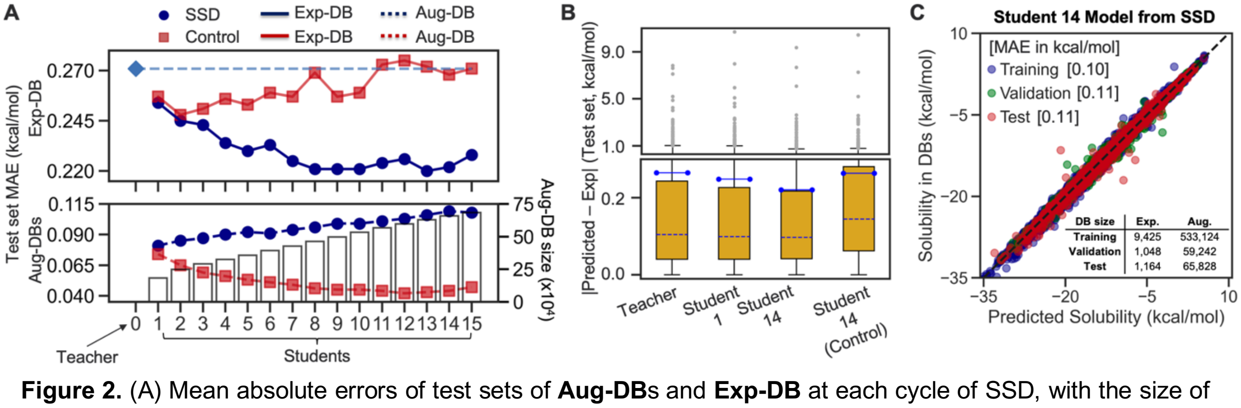


Figure 2. (A) Mean absolute errors of test sets of **Aug-DB**s and **Exp-DB** at each cycle of SSD, with the size of cumulated **Aug-DB**s. (B) Box plots of absolute error distributions for the test set of **Exp-DB**, for four representative models from SSD. (Yellow box: interquartile range, blue line: mean, blue dotted line: median, lower/upper bound of the error bar: 5th/95th percentile, gray dots: outliers beyond the 95th percentile.) (C) Parity plot of solubility values from the prediction and database for the 14th student model.

Model performance.

Fig. 2A illustrates the results from the SSD training (Fig. 1D) of the GNN shown in Fig. 1A. The initial training to obtain the Teacher model showed the MAE of 0.27 kcal/mol for the test set of **Exp-DB**. As the SSD cycles proceeded, the sizes of **Aug-DB**s gradually increased. Interestingly, the MAE of the **Exp-DB** test set decreased as **Aug-DB**s grew, until Student 14, even though no data points in **Aug-DB**s are from experiments. The MAEs increased from Student 13 to Student 14, but the increment is only 0.002 kcal/mol. The Student 14 model achieved an MAE of 0.222 kcal/mol for the **Exp-DB** test set. This indicates that the SSD scheme works properly in data augmentation while it still captures experimental solubility trends. Meanwhile, the test set MAEs of **Aug-DB**s remain relatively constant, which is another indication of the feasibility of SSD. However, a higher MAE was shown in Student 15 (0.229 kcal/mol) than in Student 14 (0.222 kcal/mol) for the test set of **Exp-DB** (Fig. 2A), which needs further analysis using other QM methods besides COSMO-RS (**Outlier analysis**, *vide infra*).

On the contrary, the Control models showed a gradual increase in MAEs of the **Exp-DB** test set, demonstrating that simply merging solubilities from experiments and COSMO-RS is not advantageous for maintaining the accuracy of the ground-truth **Exp-DB**. In addition, Control shows overfitting to **Aug-DB**s, as test set MAEs are decreasing for **Aug-DB**s, whereas those for **Exp-DB** are increasing. These MAEs diverge rather than approaching the irreducible experimental uncertainty of 0.2 kcal/mol. It is arguable that there is a difference of only around 0.05 kcal/mol between test set MAEs from SSD (0.22 kcal/mol) and Control (0.27 kcal/mol). However, the Control model shows a discrepancy of 0.23 kcal/mol between test set MAEs of **Exp-DB** and **Aug-DB**s (0.27 vs. 0.04), whereas that from SSD is around 0.1 kcal/mol (0.22 vs. 0.12).

Moreover, the box plot in Fig. 2B demonstrates that the SSD approach is promising. For the test set of **Exp-DB**, Student 1 shows more significant outliers (gray dots) with higher errors than the Teacher. This outlying behavior is remedied in Student 14, with a lower MAE (blue line) and a more narrow interquartile range (yellow box) than Teacher. The higher accuracy of Student 14 than Teacher shows the effectiveness of SSD. In contrast, Student 14 from Control does not show significant accuracy improvement compared to Teacher, and outliers also show higher errors. Student 14 from SSD showed high and balanced accuracies for the training, validation, and test sets of the integrated database, with overall MAEs of 0.10, 0.11, and 0.11 kcal/mol, respectively (Fig. 2C).

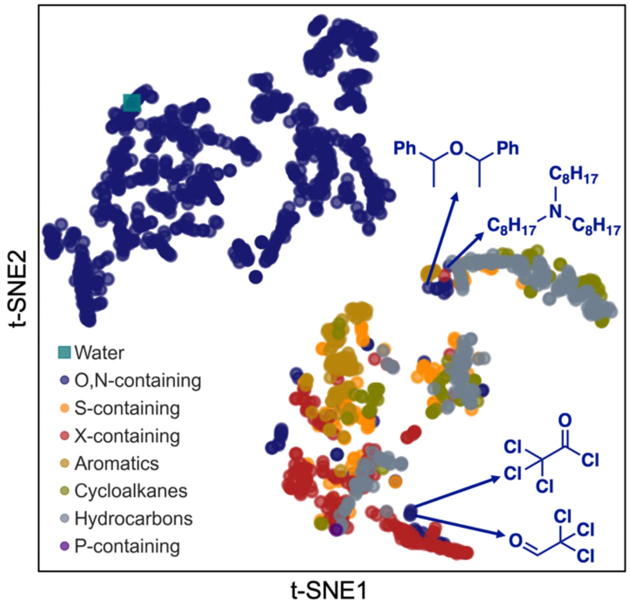
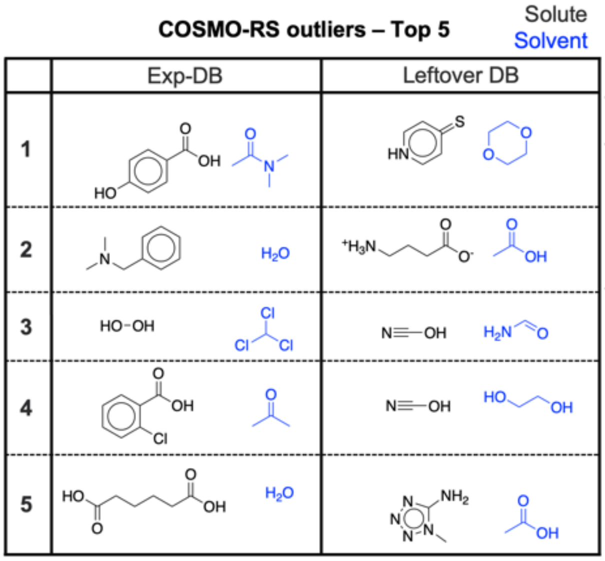


Figure 3. 2D plot of t-distributed stochastic neighbor embeddings (t-SNEs) for the latent vectors of 1,447 solvents obtained from Student 14 model.

Of note, we tested other variants of semi-supervised learning methods, such as noisy student self-distillation (NSSD). During the training using NSSD, noises are introduced to the model by applying dropout and stochastic depth methods to the hidden layers of the model. NSSD was effective in ML models for image classification because partially dropping the information from hidden layers would be helpful for handling the variance among different images with the same label.[62, 63] In this regard, we also tested multiple NSSD models in solubility predictions with different dropout rates and survival probabilities of stochastic depth. However, in all cases, NSSD showed higher prediction errors than SSD (i.e., no noise introduced to the model). That is because dropout and stochastic depth can presumably cause errors in recognizing a molecule. The model can miss the information about key structural features related to solubility due to introducing the noise to the model. In contrast, for images, if some part is lost, the model can still recognize and classify them. As a result, the SSD method was chosen throughout this study instead of NSSD for the development of self-evolving solubility databases and GNNs.

We also carried out the clustering analysis of t-distributed stochastic neighbor embeddings (t-SNEs) of latent vectors for 1,447 solvents included in all the databases shown in Fig. 1B. This analysis is to further verify the chemical feasibility of the Student 14 model (Fig. 3). 2D t-SNE coordinates were obtained for these solvents, and each solvent was categorized according to the priority of categories listed in the legend of Fig. 3. For example, if a solvent contains both O and S, it is classified as ‘O,N-containing’ because O has higher priority than S. We found clear clustering patterns among several categories: O,N-containing (upper left), halogen (X)-containing (lower center), and hydrocarbon solvents (the rest of them). O,N-containing solvents exclusively occupy their region, possibly because they are solvents that can participate in hydrogen bonds and show characteristic solubility trends.

However, some O,N-containing solvents are located close to other molecular groups, such as aromatics, hydrocarbons, and X-containing ones. We found that such solvents contain oxygen or nitrogen with the atoms corresponding to the molecular groups they are close to. For instance, Bis(alpha-phenylethyl) ether shown has oxygen with aromatic carbons, so it is placed around the Aromatics cluster, and alkyl groups exist in trioctylamine, leading to its position around Hydrocarbons. Two solvents having a carbonyl group and three chlorine atoms can also be found near the X-containing cluster (Fig. 3).



ΔG_{solv} (kcal/mol)

Exp-DB	Exp.	M06-2X	COSMO-RS
1	-11.51	-11.65	-15.98
2	-5.05	-5.04	-2.63
3	-4.70	-4.84	-2.80
4	-10.37	-10.21	-12.18
5	-12.86	-13.04	-14.33

Leftover DB	Student 14 Predicted	M06-2X	COSMO-RS
1	-8.69	-8.74	-17.70
2	-10.91	-10.99	-17.65
3	-5.94	-5.84	-12.46
4	-6.02	-6.18	-12.04
5	-10.84	-10.85	-16.84

Figure 4. Top 5 outliers of COSMO-RS when comparing ΔG_{solv} with (1) **Exp-DB** and (2) the data points that were not included in **Aug-DB**s after training until Student 14 (**Leftover DB**).

Outlier analysis.

As introduced in Fig. 1F, the error and outlier analysis were carried out by using the **QM-DB** solubilities calculated in the SMD-M06-2X/Def2-TZVP level of theory. Here, we focused on analyzing ΔG_{solv} calculated using COSMO-RS since these values were referred to when we applied the SSD approach (Fig. 1D). The top five outliers of COSMO-RS against **Exp-DB** were found from the comparison of ΔG_{solv} in **CombiSolv-QM**[53] (Fig. 1B) with those in **Exp-DB**. Their absolute errors range from 1.5-4.5 kcal/mol, whereas ΔG$_{solv}$ for these five solute-solvent pairs in **QM-DB** showed an error range of only 0.01-0.2 kcal/mol. The same analysis was performed for **Leftover DB** which contains the data points that were not included in **Aug-DB**s after 14 cycles of SSD. In this case, QM solubility values were compared with the predicted values from Student 14 model since no experimental ΔG$_{solv}$ is available for these outliers. The ΔG$_{solv}$ values from the M06-2X level were much closer to Student-predicted values compared to those from COSMO-RS. It should be noted that only COSMO-RS solubility values were used during SSD, whereas M06-2X values had not been seen.

In addition, common structural features can be found in these outliers. All solute and solvent molecules contain hydrogen bond donors and acceptors. Some molecules also contain halogens. The higher accuracy of M06-2X with SMD for those molecules may be in part attributed to the halogenicity, hydrogen bond acidity, and basicity parameters used by SMD. Further analysis was performed for all 190,989 data points in **QM-DB**; among them, 117,605 were already merged into **Aug-DB**s. The remaining 73,384 are in Leftover DB, and the SMD-M06-2X/Def2-TZVP method showed better accuracy than COSMO-RS for 17,819 of them. It should be emphasized that the outlier analysis discussed above does not indicate that COSMO-RS is inappropriate for the model training. The SSD guided by COSMO-RS solubilities led to the Student 14 model with the self-evolving database consisting of 658,194 data points in total (Fig. 2C). However, the outlier analysis implies that employing multiple QM methods would further improve the ML model and augmented database obtained from SSD since each method shows higher accuracy than others for a certain group of molecules.

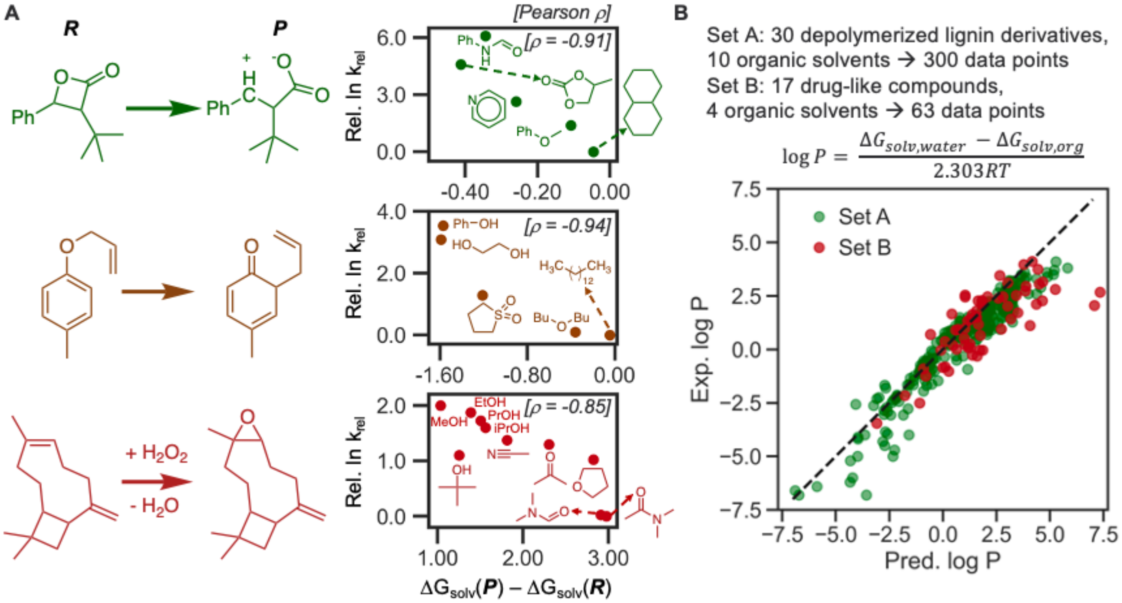


Figure 5. (A) A linear relationship between ΔG$_{solv}$ of reactants and products vs. reaction rates for three organic reactions. (B) A parity plot showing the correlation between experimental and predicted log P values for 363 data points.

Application 1 – Linear relationships between solvation free energy and reaction rates of organic reactions.

It is crucial to find a linear solvation free energy relationship (LSER) between the property relevant to solvents and reaction rates of organic reactions since it informs solvent selections in chemical process design. Previous studies have elucidated the LSER between reaction rates and experimentally measured solvent properties such as dielectric constant and polarity.[7-9] Here, we demonstrate new directions to discover the LSER pertinent to the kinetics of organic reactions through ML. For three organic reactions, the linear relationship was found (Fig. 5A) between reaction rates,[69, 70] and ΔG$_{solv}$ differences between product and reactant. The high negative Pearson correlation coefficients were shown for all three cases (ρ = -0.94 ~ -0.85). These negative correlations indicate that the solvation stabilization of the product leads to a higher reaction rate, while the reactant should be less stabilized to undergo reactions.

The first reaction is a ring opening to decarboxylate the reactant and form an alkene whose reaction rates were measured in five solvents. A nonpolar solvent, decalin, shows the lowest reaction rate, whereas the fastest reaction was observed in a polar N-phenylforamide solvent. This is consistent with the fact that the zwitterionic product (***P***) is more polar than the reactant (***R***), so a polar solvent would be favorable to stabilize the product more than the reactant. The second reaction, Cope rearrangement, in five different solvents was investigated. Two solvents with hydroxyl groups (ethylene glycol and phenol) showed higher reaction rates than other solvents. This is because the ketone group in the product can form hydrogen bonds with alcoholic solvents, leading to product stabilization and faster reactions. Our ML model also showed reliable and chemically explainable results in the complex third reaction example, the epoxidation of β-Caryophyllene investigated in 10 different solvents.

Notably, the above results manifest that the ΔG_{solv} difference of only around 1 kcal/mol can lead to a large difference in reactivity predictions, demanding a fast and accurate ML model. Such ML-driven design of solvent systems is promising because it can save time taken in expensive QM calculations while being accurate. The linear relationship can be extrapolated to the new solvents for which experiments were not performed yet, leading to the design of solvent systems toward a higher reaction rate. Using the ML-predicted quantities would facilitate solvent selections in designing chemical reactions. However, the above three reactions were not performed at room temperature, whereas the ML model gives the solubilities at room temperature. Considering the temperature dependence of solubility would be one of the ways to further improve ML models, although the results in Fig. 5A already show decent correlations.

Application 2 – Prediction of partition coefficients for lignin-derived monomers and drug-like molecules.

Table 1. Comparison of prediction accuracies of 363 partition coefficients for COSMO-RS and ML model.

		Kendall tau rank coefficient	**RMSE**
Set A	COSMO-RS	0.77	0.50
	ML	0.82	0.77
Set B	COSMO-RS	0.77	1.00
	ML	0.58	1.41

As the second application example, we examined our GNN model by calculating the 363 water-organic partition coefficients (log P) of which experimental values are available from the literature.[71] The dataset of log P values is divided into two sets (Set A and Set B). Set A consists of log P measured for 30 depolymerized lignin derivatives dissolved in 10 organic solvents and water. There are log P values for 17 drug-like compounds dissolved in four organic solvents and water, making up 63 data points. Fig. 5B depicts the parity plot of ML-predicted log P vs. experimental ones. Overall, the model shows predictions close to experimental ones while overestimating log P in some cases. We compared the accuracy of our model with log P calculated using COSMO-RS. Table 1 summarizes Kendall tau rank coefficients and root-mean-square errors (RMSEs) for COSMO-RS and our ML model. The GNN showed rank coefficients of 0.82 and 0.58 for Set A and Set B, respectively, whereas those for COSMO-RS are 0.77 for both sets.[71] Our GNN showed a better correlation for Set A than COSMO-RS, while COSMO-RS performed better in Set B. In terms of RMSE, COSMO-RS showed better accuracy in both sets.

These results indicate that the ML model reliably captures the rank of solubilities in different organic solvents, although relatively less accuracy was shown in predicting the solubility value itself. However, it should be emphasized that calculating log P using ML takes less than one second and yields an accuracy comparable to QM methods, whereas QM and COSMO-RS calculations of log P are computationally demanding. Some acidic/basic solutes can be ionized into cations/anions in the solution. In addition, an organic solvent can dissolve water and vice versa. A detailed consideration of these effects would further improve the accuracy. Rapid and reliable log P predictions using ML would lead to the computational design of solvent systems for separation processes in organic, pharmaceutical synthesis, and renewable energy industries.

Conclusions

Solubility is a critical molecular property to consider when designing chemical processes such as synthesis and separation in organic, pharmaceutical, and sustainable chemistry. Many ML models have been developed, but one should have a reliable integration of experimental and computational solubility databases to maximize the database size and, thus prediction accuracy. To reduce the discrepancies among different data sources, here, semi-supervised self-training methodologies were adopted in solubility predictions, leading to self-evolving solubility databases and GNN predictive models. The resulting model showed reliable accuracy. It was also applied to practical examples of solvent selection in chemical reactions and separation processes. All these results demonstrate the practical applicability of the developed model to the design of solvent systems in chemical processes. Such approaches can be further improved by employing multiple QM methods during the data augmentation process. Considering temperature effects on solubility in ML models should also be pursued to achieve the application of the model to a broader scope of chemistry. Predicting solubilities in multicomponent solvents is another challenge in the expansion of ML models, which would lead to the realistic modeling of mixtures utilized in various chemical reactions and separation processes.

References

1. Dalton, T.; Faber, T.; Glorius, F., C–H Activation: Toward Sustainability and Applications. *ACS Cent. Sci.* **2021,** *7*, 245-261.
2. Dyson, P. J.; Jessop, P. G., Solvent effects in catalysis: rational improvements of catalysts via manipulation of solvent interactions. *Catal. Sci. Technol.* **2016,** *6*, 3302-3316.
3. Huxoll, F.; Jameel, F.; Bianga, J.; Seidensticker, T.; Stein, M.; Sadowski, G.; Vogt, D., Solvent Selection in Homogeneous Catalysis—Optimization of Kinetics and Reaction Performance. *ACS Catal.* **2021,** *11*, 590-594.
4. Hailes, H. C., Reaction Solvent Selection: The Potential of Water as a Solvent for Organic Transformations. *Org. Process Res. Dev.* **2007,** *11*, 114-120.
5. Varghese, J. J.; Mushrif, S. H., Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *React. Chem. Eng.* **2019,** *4*, 165-206.
6. Moseley, J. D.; Murray, P. M., Ligand and solvent selection in challenging catalytic reactions. *J. Chem. Tech. Biotech.* **2014,** *89*, 623-632.
7. Slakman, B. L.; West, R. H., Kinetic solvent effects in organic reactions. *J. Phys. Org. Chem.* **2019,** *32* (3), e3904.
8. Sherwood, J.; Parker, H. L.; Moonen, K.; Farmer, T. J.; Hunt, A. J., N-Butylpyrrolidinone as a dipolar aprotic solvent for organic synthesis. *Green Chem.* **2016,** *18* (14), 3990-3996.
9. Dyson, P. J.; Jessop, P. G., Solvent effects in catalysis: rational improvements of catalysts via manipulation of solvent interactions. *Catalysis Science & Technology* **2016,** *6* (10), 3302-3316.
10. Pinho, S. P.; Macedo, E. A., Chapter 20 Solubility in Food, Pharmaceutical, and Cosmetic Industries. In *Developments and Applications in Solubility*, The Royal Society of Chemistry: 2007; pp 305-322.
11. Jouyban, A., Review of the cosolvency models for predicting solubility of drugs in water-cosolvent mixtures. *J. Pharm. Pharm. Sci.* **2008,** *11*, 32-58.
12. Llinàs, A.; Glen, R. C.; Goodman, J. M., Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008,** *48*, 1289-1303.
13. Bergström, C. A. S.; Charman, W. N.; Porter, C. J. H., Computational prediction of formulation strategies for beyond-rule-of-5 compounds. *Adv. Drug Deliv. Rev.* **2016,** *101*, 6-21.

14. Bergström, C. A. S.; Larsson, P., Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int. J. Pharm.* **2018,** *540*, 185-193.

15. Fioressi, S. E.; Bacelo, D. E.; Rojas, C.; Aranda, J. F.; Duchowicz, P. R., Conformation-independent quantitative structure-property relationships study on water solubility of pesticides. *Ecotoxicol. Environ. Saf.* **2019,** *171*, 47-53.

16. Nayak, A. K.; Panigrahi, P. P., Solubility Enhancement of Etoricoxib by Cosolvency Approach. *ISRN Phys. Chem.* **2012,** *2012*, 820653.

17. Seedher, N.; Kanojia, M., Co-solvent solubilization of some poorly-soluble antidiabetic drugs. *Pharm. Dev. Technol.* **2009,** *14*, 185-192.

18. Newmister, S. A.; Li, S.; Garcia-Borràs, M.; Sanders, J. N.; Yang, S.; Lowell, A. N.; Yu, F.; Smith, J. L.; Williams, R. M.; Houk, K. N.; Sherman, D. H., Structural basis of the Cope rearrangement and cyclization in hapalindole biogenesis. *Nat. Chem. Biol.* **2018,** *14* (4), 345-351.

19. Kraml, J.; Hofer, F.; Kamenik, A. S.; Waibl, F.; Kahler, U.; Schauperl, M.; Liedl, K. R., Solvation Thermodynamics in Different Solvents: Water–Chloroform Partition Coefficients from Grid Inhomogeneous Solvation Theory. *J. Chem. Inf. Model.* **2020,** *60* (8), 3843-3853.

20. Esteban, J.; Vorholt, A. J.; Leitner, W., An overview of the biphasic dehydration of sugars to 5-hydroxymethylfurfural and furfural: a rational selection of solvents using COSMO-RS and selection guides. *Green Chem.* **2020,** *22* (7), 2097-2128.

21. Huber, G. W.; Chheda, J. N.; Barrett, C. J.; Dumesic, J. A., Production of liquid alkanes by aqueous-phase processing of biomass-derived carbohydrates. *Science* **2005,** *308* (5727), 1446-1450.

22. Shen, Z.; Van Lehn, R. C., Solvent Selection for the Separation of Lignin-Derived Monomers Using the Conductor-like Screening Model for Real Solvents. *Ind. Eng. Chem. Res.* **2020,** *59* (16), 7755-7764.

23. Hollas, A.; Wei, X.; Murugesan, V.; Nie, Z.; Li, B.; Reed, D.; Liu, J.; Sprenkle, V.; Wang, W., A biomimetic high-capacity phenazine-based anolyte for aqueous organic redox flow batteries. *Nat. Energy* **2018,** *3* (6), 508-514.

24. Kucharyson, J. F.; Cheng, L.; Tung, S. O.; Curtiss, L. A.; Thompson, L. T., Predicting the potentials, solubilities and stabilities of metal-acetylacetonates for non-aqueous redox flow batteries using density functional theory calculations. *J. Mat. Chem. A* **2017,** *5* (26), 13700-13709.

25. S. V, S. S.; Law, J. N.; Tripp, C. E.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; St. John, P. C., Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat. Mach. Intell.* **2022,** *4* (8), 720-730.

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

27. Bradley, J.-C.; Neylon, C.; Guha, R.; Williams, A.; Hooker, B.; Lang, A.; Friesen, B.; Bohinski, T.; Bulger, D.; Federici, M.; Hale, J.; Mancinelli, J.; Mirza, K.; Moritz, M.; Rein, D.; Tchakounte, C.; Truong, H., Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents. *Nat. Preced.* **2010**.

28. Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., Minnesota Solvation Database (MNSOL) version 2012. Retrieved from the Data Repository for the University of Minnesota, https://doi.org/10.13020/3eks-j059. **2020**.

29. Kelly, C. P.; Cramer, C. J.; Truhlar, D. G., SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute−Water Clusters. *J. Chem. Theory Comput.* **2005,** *1*, 1133-1152.

30. Thompson, J. D.; Cramer, C. J.; Truhlar, D. G., New Universal Solvation Model and Comparison of the Accuracy of the SM5.42R, SM5.43R, C-PCM, D-PCM, and IEF-PCM Continuum Solvation Models for Aqueous and Organic Solvation Free Energies and for Vapor Pressures. *J. Phys. Chem. A* **2004,** *108*, 6532-6542.

31. Mobley, D. L.; Guthrie, J. P., FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **2014,** *28* (7), 711-720.

32. 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. *J. Phys. Chem. Ref. Data* **2017,** *46* (3), 033102.

33. Llinas, A.; Avdeef, A., Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets. *J. Chem. Inf. Model.* **2019,** *59*, 3036-3040.
34. Llinas, A.; Oprisiu, I.; Avdeef, A., Findings of the Second Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2020,** *60*, 4791-4803.
35. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009,** *113* (18), 6378-6396.
36. Boothroyd, S.; Kerridge, A.; Broo, A.; Buttar, D.; Anwar, J., Solubility prediction from first principles: a density of states approach. *Phys. Chem. Chem. Phys.* **2018,** *20*, 20981-20987.
37. Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V., First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory Comput.* **2012,** *8*, 3322-3337.
38. Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O., A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* **2015,** *17*, 6174-6191.
39. Ran, Y.; He, Y.; Yang, G.; Johnson, J. L. H.; Yalkowsky, S. H., Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* **2002,** *48*, 487-509.
40. Palmer, D. S.; Mitchell, J. B. O., Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Mol. Pharm.* **2014,** *11*, 2962-2972.
41. Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N., Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020,** *11*, 5753.
42. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R., Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019,** *59*, 3370-3388.
43. Qiu, J.; Albrecht, J.; Janey, J., Solubility Behaviors and Correlations of Common Organic Solvents. *Org. Process Res. Dev.* **2020,** *24*, 2702-2708.
44. Lovrić, M.; Pavlović, K.; Žuvela, P.; Spataru, A.; Lučić, B.; Kern, R.; Wong, M. W., Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *J. Chemom.* **2021,** *35*, e3349.
45. Lim, H.; Jung, Y., Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* **2019,** *10*, 8306-8315.
46. 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. *Front. Oncol.* **2020,** *10*.
47. Pathak, Y.; Laghuvarapu, S.; Mehta, S.; Priyakumar, U. D., Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-Like Molecules. *Proc. AAAI Conf. AI* **2020,** *34*, 873-880.
48. Sorkun, M. C.; Koelman, J. M. V. A.; Er, S., Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **2020,** *24*, 101961-101961.
49. Francoeur, P. G.; Koes, D. R., SolTranNet–A Machine Learning Tool for Fast Aqueous Solubility Prediction. *J. Chem. Inf. Model.* **2021,** *61*, 2530-2536.
50. Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D., A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J. Cheminform.* **2020,** *12*, 15.
51. 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. *J. Chem. Inf. Model.* **2022,** *62* (3), 433-446.
52. Vermeire, F. H.; Chung, Y.; Green, W. H., Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022,** *144* (24), 10785-10797.
53. Vermeire, F. H.; Green, W. H., Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021,** *418*, 129307.
54. Yu, J.; Zhang, C.; Cheng, Y.; Yang, Y.-F.; She, Y.-B.; Liu, F.; Su, W.; Su, A., SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. **2022.**

55. Bilodeau, C.; Jin, W.; Xu, H.; Emerson, J. A.; Mukhopadhyay, S.; Kalantar, T. H.; Jaakkola, T.; Barzilay, R.; Jensen, K. F., Generating molecules with optimized aqueous solubility using iterative graph translation. *React. Chem. Eng.* **2022,** *7* (2), 297-309.

56. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020,** *11*, 2328.

57. Panapitiya, G.; Girard, M.; Hollas, A.; Murugesan, V.; Wang, W.; Saldanha, E., Predicting Aqueous Solubility of Organic Molecules Using Deep Learning Models with Varied Molecular Representations. *arXiv preprint arXiv:2105.12638* **2021**.

58. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009,** *113*, 6378-6396.

59. Kelly, C. P.; Cramer, C. J.; Truhlar, D. G., SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute−Water Clusters. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1133-1152.

60. Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S., Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **2008,** *51* (4), 769-779.

61. Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J., The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput. Aided Mol. Des.* **2010,** *24* (4), 259-279.

62. Xie, Q.; Luong, M.-T.; Hovy, E.; Le, Q. V. In *Self-training with noisy student improves imagenet classification*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp 10687-10698.

63. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; Li, C.-L., Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process Syst.* **2020,** *33*, 596-608.

64. He, J.; Gu, J.; Shen, J.; Ranzato, M. A., Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788* **2019**.

65. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.

66. Kim, Y.; Cho, J.; Naser, N.; Kumar, S.; Jeong, K.; McCormick, R. L.; St. John, P.; Kim, S., Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proc. Combust. Inst.* **2022**, Accepted.

67. Qin, S.; Jiang, S.; Li, J.; Balaprakash, P.; Van Lehn, R.; Zavala, V., Capturing Molecular Interactions in Graph Neural Networks: A Case Study in Multi-Component Phase Equilibrium. **2022**.

68. Wang, H.; Lian, D.; Zhang, Y.; Qin, L.; Lin, X., Gognn: Graph of graphs neural network for predicting structured entity interactions. *arXiv preprint arXiv:2005.05537* **2020**.

69. Welton, T.; Reichardt, C., *Solvents and solvent effects in organic chemistry*. John Wiley & Sons: 2011.

70. Steenackers, B.; Neirinckx, A.; De Cooman, L.; Hermans, I.; De Vos, D., The Strained Sesquiterpene β-Caryophyllene as a Probe for the Solvent-Assisted Epoxidation Mechanism. *ChemPhysChem* **2014,** *15* (5), 966-973.

71. Tshepelevitsh, S.; Hernits, K.; Leito, I., Prediction of partition and distribution coefficients in various solvent pairs with COSMO-RS. *J. Comput. Aided Mol. Des.* **2018,** *32* (6), 711-722.