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, 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, 45-56] Particularly, current advanced ML models used graph neural networks (GNNs) combined with interaction layers[47, 53, 57] recurrent neural networks with attention layers,[45] and natural language processing-based transformers.[54, 58] 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, 59] 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, 60] 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. In addition, QM solubilities systematically deviate from experimental ones. 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,[61-64] 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,[65, 66] natural language processing,[67] reaction classification,[68] and protein structures.[69]

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 11 chemical reactions. Second, 363 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.

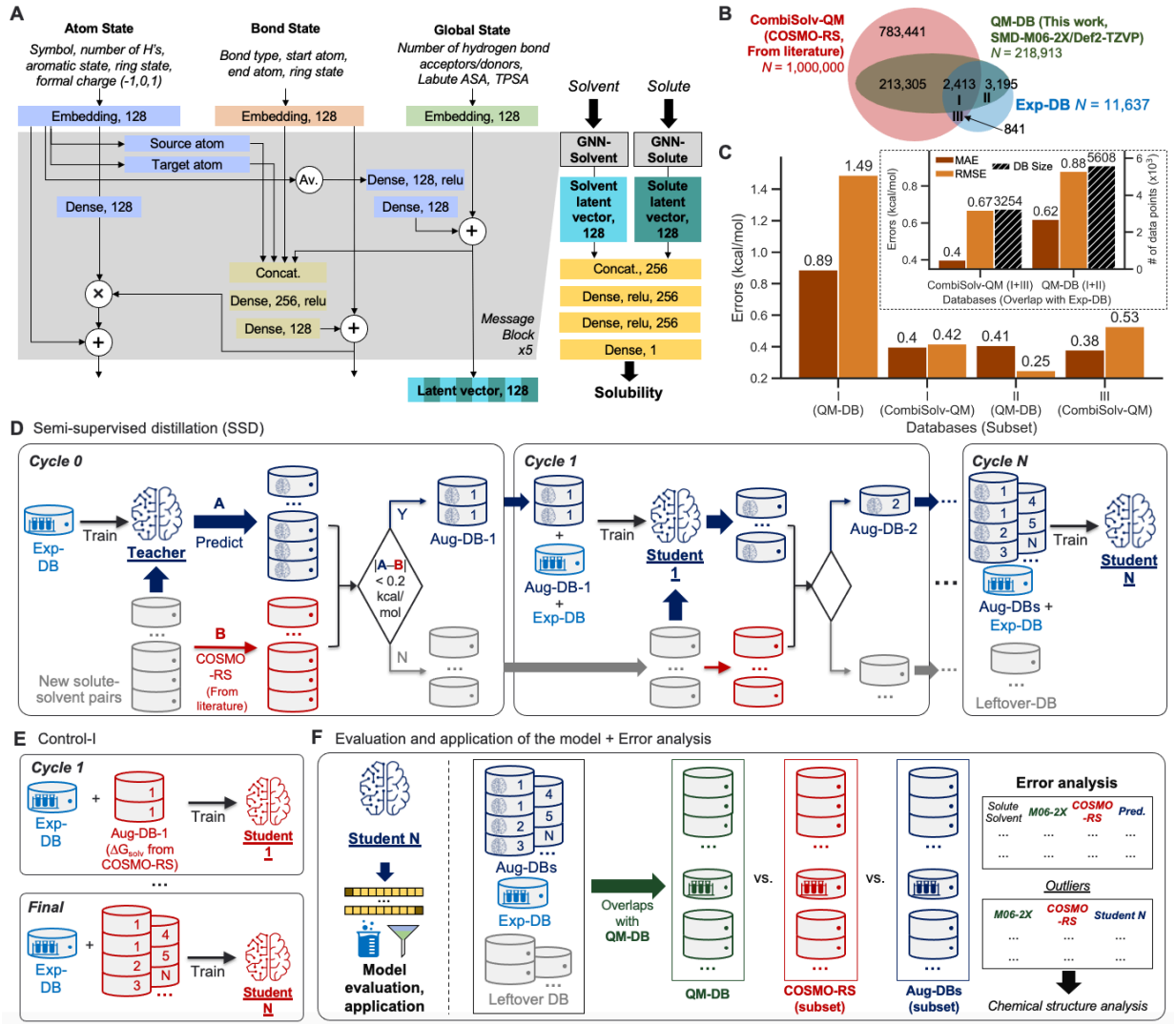


Figure 1. (A) Architecture of the graph neural network for solubility. (B) Description of three databases used to evaluate theoretical methods against experimental solubilities. (C) Comparison of accuracies of **CombiSolv-QM** and **QM-DB** for the data points overlapping with **Exp-DB**. (D) Semi-supervised distillation (SSD) for self-evolving solubility databases and graph neural networks. (E) Control-I 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.[59, 70] It consists of three blocks representing the atom, bond, and global state of a molecule. Initial atom, bond, and 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.[71, 72] 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 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.[70] 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. More details about the hyperparameter tuning procedure are available in the Methods section.

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 calculating molecular properties pertinent to solvation, e.g., the redox potentials of 174 organic molecules in water and acetonitrile.[25] In this work, a new database (**QM-DB**) was built by calculating ΔG_{solv} for 218,913 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). These three databases have 2,413 overlapped data points (Region **I**), whereas we calculated 3,195 more ΔG_{solv} values using the SMD-M06-2X/Def2-TZVP method, compared to **CombiSolv-QM** (Region **II**). Calculated solubilities of 841 solute-solvent pairs in **Exp-DB** are available only in **CombiSolv-QM** (Region **III**) due to the unavailability of some solvents in SMD calculations

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,** showing an MAE and RMSE of 0.4 and 0.67 kcal/mol with respect to **Exp-DB**. Meanwhile, the MAE and RMSE higher than **CombiSolv-QM** (0.62 and 0.88 kcal/mol, respectively) were observed from **QM-DB**. However, the error values were not significantly increased for the 5,351 data points, which are about 1.7 times higher than the 3,254 overlapped data points in **CombiSolv-QM**. Accuracies of the two theoretical methods were further analyzed in terms of different overlapping regions (Regions **I**-**III**) shown in Fig. 1B. For Region **I**, the COSMO-RS-calculated values in **CombiSolv-QM** are more accurate than the DFT-calculated ones in **QM-DB**. Nonetheless, the M06-2X DFT shows notably high accuracy for the 3,195 data points in Region **II**, with an MAE and RMSE of 0.41 and 0.25 kcal/mol, respectively. Meanwhile, comparable accuracy was achieved with the COSMO-RS method in Region **III**.

These results manifest that each theoretical method has strengths and weaknesses in terms of computational costs, the scope of molecules available for calculations, and accuracies for different functional groups of solutes and solvents, etc. Detailed analysis of the functional groups is discussed in the **Error analysis** Section (*vide infra*). It is noteworthy that the results in Fig. 1C do not necessarily indicate the superiority of one method in evaluating ΔG_{solv} compared to the other. Although **QM-DB** is less extensive than **CombiSolv-QM**, the M06-2X functional can be used as a complementary method of COSMO-RS for explaining the errors of QM methods and ML models after the model development. COSMO-RS is typically a more cost-efficient option for high-throughput calculations than DFT with implicit solvation models because it needs DFT calculations of a charge density(σ) profile only once per one solute/solvent. Then, in principle, COSMO-RS can readily calculate ΔG_{solv} for any combinations of solute-solvent pairs whose σ profiles are available.

In contrast, SMD-DFT methods such as SMD-M06-2X/Def2-TZVP need multiple geometry optimization and thermochemistry calculations for the same solute when a solvent is changed, which is computationally demanding. SMD parameters have been tabulated for only 179 solvents, limiting the molecular scope for estimating ΔG_{solv}. However, SMD-M06-2X/Def2-TZVP can show higher accuracies than COSMO-RS for certain functional groups. The 'committee' of multiple theoretical methods would lead to more reliable development and evaluation of databases and predictive models than utilizing only one method. More details are discussed in the following sections.

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 the reliability gained 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.[61-64] 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 the 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 remain 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 multiple times, enabling 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. Such gradual integration leads to better accuracy than combining the whole **CombiSolv-QM** with **Exp-DB** 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 are transferred from the previous cycle when training the Student model in the current cycle. Only the databases (**Aug-DB**s and **Exp-DB**) are transferred, and each Student is trained from scratch at each cycle. In other words, the current Student is totally blind to the training results of previous Students. Therefore, at each cycle, the model learns new relationships between chemical structures and solubility that are not biased by previous cycles but are comprehensively applicable to all molecules from the previous and current cycles. This SSD scheme is to ensure that the new **Aug-DB-*i*** at the *i*-th cycle is integrated well with the databases cumulated from previous cycles, and it shows no significant discrepancies and anomalies during the training.

Ultimately, the 35th cycle yields the 'Student 35' model and the integrated database containing **Exp-DB** and 35 **Aug-DB**s. The cycle was terminated at the 35th cycle because the RMSE for the test set of **Exp-DB** does not show any more significant improvement (Detailed results in Fig. 2, *vide infra*). This

stopping criterion was applied since the leftover data points in **CombiSolv-QM** (so-called **Leftover-DB**) 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 models trained by the databases simply combining ΔG_{solv} values from experiments and COSMO-RS (Control-I, Fig. 1E). The analysis on Control-I was performed at every SSD cycle to compare the increasing/decreasing trends of MAEs and RMSEs when the models are trained without/with SSD. All these Control models are examined to demonstrate that the SSD approach in Fig. 1D is optimal for maximizing the database size while minimizing the discrepancy between experimental and computational ΔG_{solv} and achieving the best accuracy.

The resulting Student 35 model was then subject to subsequent evaluation, error analysis, and applications (Fig. 1F). To evaluate the model's accuracy, mean absolute errors (MAEs), root-mean-square errors (RMSEs), 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 35, SMD-M06-2X/Def2-TZVP, and COSMO-RS calculations. Outliers were identified from this comparison, and their chemical structures were analyzed to assess the

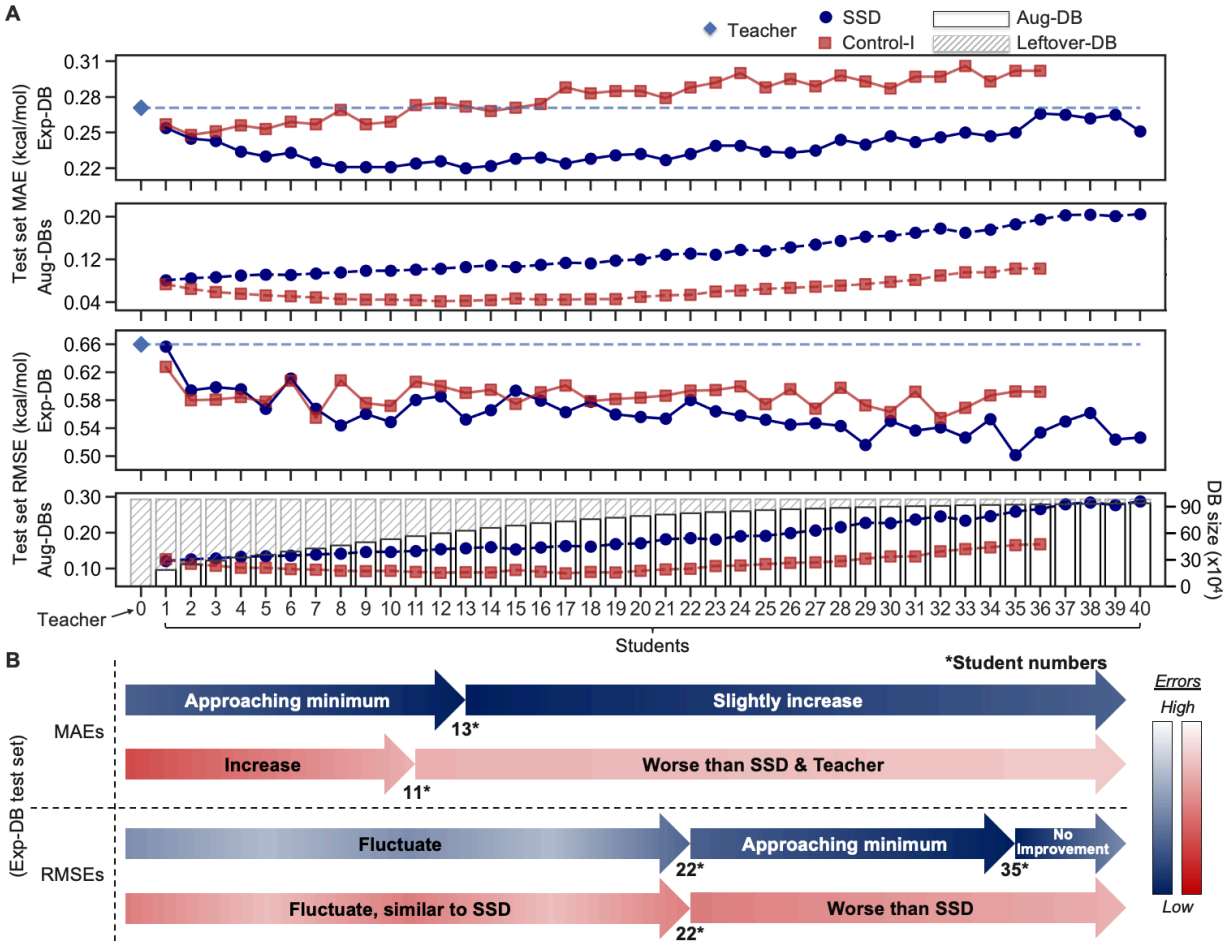


Figure 2. (A) Mean absolute errors (MAEs) and root-mean-square errors (RMSEs) of test sets of **Aug-DB**s and **Exp-DB** at each cycle of SSD, with the size of cumulated **Aug-DB**s. (B) A schematic summary of the increasing and decreasing trends of MAEs and RMSEs for SSD and Control-I models at each cycle.

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.

Model performance.

This section explains how the optimal number of SSD cycles was determined as 35. 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 resulted in 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 for the **Exp-DB** test set reached a minimum when **Aug-DB**s was expanded until Student 13, even though no data points in **Aug-DB**s are from experiments. The database was grown from 11,637 (Teacher) to 639,925 (Student 13) data points. The Student 13 model achieved an MAE of 0.22 kcal/mol for the **Exp-DB** test set. This indicates that the SSD scheme works properly in the data augmentation while the model still captures experimental solubility trends. On the contrary, the Control-I models show a rise 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**. The **Exp-DB** test set MAE was not improved in the Control-I Student 13 model (0.272 kcal/mol) compared to that for Teacher (0.271 kcal/mol).

It is arguable that there is only a small difference of around 0.05 kcal/mol between **Exp-DB** test set MAEs from SSD (0.22 kcal/mol) and Control-I (0.27 kcal/mol) in Student 13. However, we also need to consider the gap between the test set prediction error of **Exp-DB** and that of **Aug-DB**s, because one of the key goals of SSD is the reconciliation between experimental and computational data. At the 13th SSD cycle, Control-I 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 only 0.11 kcal/mol (0.22 vs. 0.11). When proceeding from Teacher to Student 13, more severe overfitting to **Aug-DB**s occurred for Control-I than SSD. In other words, the test set MAEs decreased for **Aug-DB**s, while those for **Exp-DB** increased. These MAEs diverged rather than approaching the irreducible experimental uncertainty of 0.2 kcal/mol. The SSD models show the opposite trend; the test set MAEs for **Exp-DB** decreased, whereas those for **Aug-DB**s increased during the 13 SSD cycles. Although The test set MAE slightly rose from 0.08 kcal/mol (Student 1) to 0.11 kcal/mol (Student 13) for **Aug-DB**s, it is still within the experimental uncertainty range (0.2 kcal/mol) while alleviating the overfitting to **Aug-DB**s.

After Student 13, the MAEs of the **Exp-DB** test set gradually increased; however, it is hard to guarantee that 13 SSD cycles are sufficient to obtain the best model. More Student models should be analyzed because all SSD models until Student 40 still show lower MAEs than the Teacher model, and the MAE is not the only metric for evaluating the accuracy. In addition, more SSD cycles extend **Aug-DB**s, which is advantageous in terms of acquiring the larger integrated database. In this regard, we analyzed RMSEs of Student models that show more irregular trends than MAEs. The fluctuating RMSEs until Student 13 necessitate further SSD cycles to investigate whether adding more data points leads to the improvement and convergence in RMSEs. This oscillation of RMSEs persists till the 22nd Student

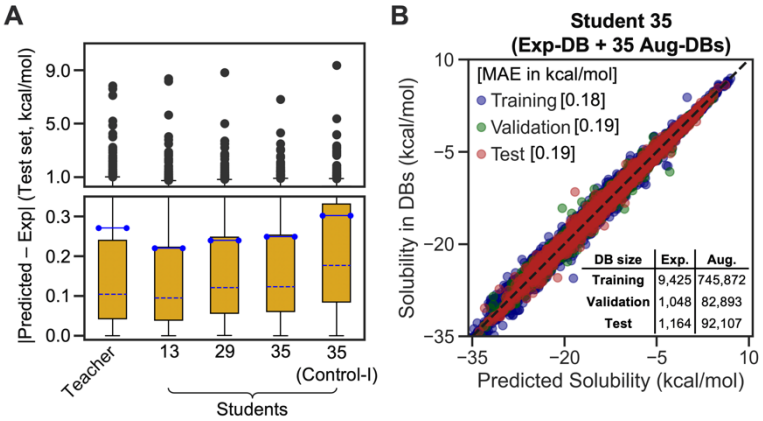


Figure 3. (A) Box plots of absolute error distributions for the test set of **Exp-DB**, for the four representative models from SSD and one Control-I model. (Yellow box: interquartile range, blue line: mean, blue dotted line: median, lower/upper bound of the error bar: 5th/95th percentile, gray dots: outliers beyond the 95th percentile.) (B) Parity plot of solubility values in the databases vs. those from the predictions of the best-case Student 35 model.

model, and RMSEs of SSD models are not significantly lower than those of Control-I models. The accuracy of SSD models begins to surpass Control-I models after Student 22.

The RMSEs decrease with less fluctuation, and the best accuracy was achieved in Student 35 with an RMSE of 0.50 kcal/mol, whereas the RMSE of the 35th Control-I model is 0.59 kcal/mol. Although MAE slightly increased from Student 13 to Student 35 (0.22→0.25 kcal/mol), RMSE reaches a minimum with the larger database compared to Student 13 (639,925→932,509 data points). The SSD cycles after Student 35 did not effectively improve the accuracy. The RMSE was minimized at Student 35 at the expense of slightly increasing MAEs. RMSE is a good metric for penalizing large errors of outliers, indicating that Student 35 effectively alleviates prediction errors of **Exp-DB** outliers while maintaining reliable accuracy for other data points. It should be emphasized that MAE was used for the loss function (Details in the Methods section), but RMSE was also minimized during the later stages of SSD. This result implies the importance of including a large amount of data to minimize high prediction errors of outliers by iterating the SSD loop multiple times. Moreover, the best accuracy was obtained in Student 35 when the prediction accuracy of the models was assessed against the 'external data set' of 371 experimental partition coefficients (*vide infra* for details).

Meanwhile, the test set RMSEs of **Aug-DB**s increased (0.12-0.29 kcal/mol) from Teacher to Student 40. However, they are still lower than the lowest test set RMSEs of **Exp-DB** (0.50 kcal/mol), and the RMSE difference between **Exp-DB** and **Aug-DB**s decreased, indicating the mitigation of overfitting. For example, the RMSE difference in Students 1, 13, and 35 is 0.54, 0.40, and 0.24 kcal/mol, respectively. This result is another manifestation of the feasibility of SSD and is analogous to what we obtained from MAEs. Fig. 2B summarizes the results from analyzing the trends of MAEs and RMSEs of **Exp-DB**. SSD showed decreasing and increasing MAEs, until and after Student 13, respectively. The Control-I models' MAEs continued to increase, and their accuracies became worse than Teacher after Student 11. RMSEs of Students up to the 22nd fluctuated for both SSD and Control-I. The SSD models subsequent to Student 22 approached the lowest RMSE and reached the minimum at Student 35. In contrast, the accuracies of Control-I models became worse than the SSD models after Student 22.

Moreover, the box plot in Fig. 3A demonstrates that executing SSD for up to 35 cycles is beneficial to obtain an optimal model. To analyze the box plot, we chose Students 13, 29, and 35 which resulted in the local minima of MAEs and RMSEs during the SSD (Fig. 2A), in addition to Teacher. For the test set of **Exp-DB**, Student 13 shows more significant outliers (gray dots) with higher errors than the Teacher, although the MAE is lower (blue line). The error of the first outlier becomes even higher in Student 29 than in Student 13. However, such errors of outliers become lowest in Student 35, which is another indication of the mitigation of overfitting through SSD. The outlying behavior is remedied in Student 35 while maintaining a lower MAE and similar interquartile range (yellow box) compared to the Teacher. In contrast, the accuracy of Student 35 from Control-I is even worse than Teacher, and their outliers also

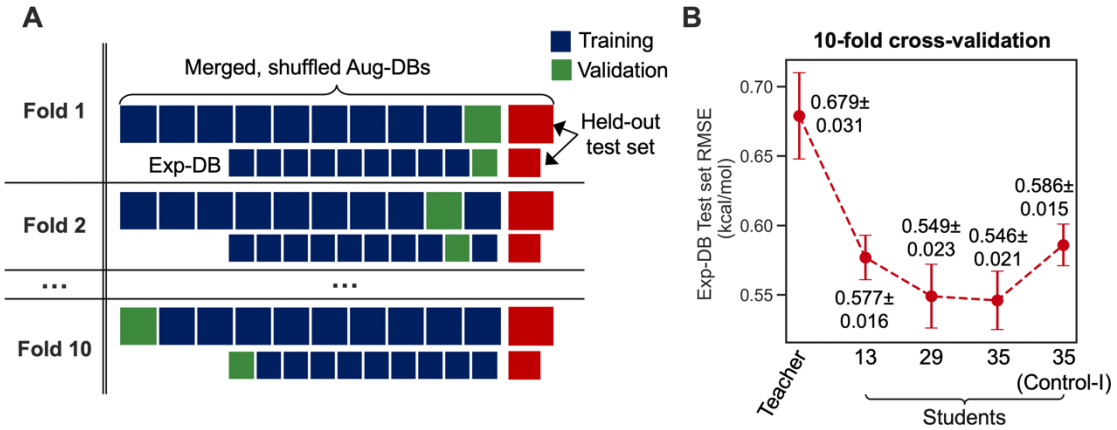


Figure 4. (A) Schematic description of the 10-fold cross-validation with considering the data splits of **Exp-DB** and **Aug-DB**s together. (B) Model accuracies from the 10-fold cross-validation. RMSEs of the **Exp-DB** test set were evaluated for each of the 10 folds. The data points and error bars indicate the mean and standard deviation of 10 RMSEs, respectively.

show higher errors. Fig. 3B illustrates the parity plot of the solubility values from the databases vs. those from the predictions of Student 35. For the whole integrated database (**Exp-DB** + 35 **Aug-DB**s), Student 35 achieved balanced accuracies among the training, validation, and test sets, with MAEs of 0.18, 0.19, and 0.19 kcal/mol, respectively.

The above results (Figs. 2 and 3) were obtained from only one certain random split of training, validation, and test sets. Therefore, it should be verified that the model is not prone to be biased or overfitted to a specific split of data sets. To this end, we carried out the 10-fold cross-validation as depicted in Fig. 4. The test sets were first held out, and the remaining data points were randomly split into ten parts. At each fold, one and the rest parts correspond to the validation and training set, respectively. This data set split was performed separately for **Aug-DB**s and **Exp-DB** to balance the ratio of the data points from data augmentation and experiments (Fig. 4A). Fig. 4B displays the **Exp-DB** test set RMSEs of the models shown in Fig. 3A. Means and standard deviations of 10 RMSEs were evaluated for each model. The mean of RMSEs decreases significantly from 0.679 to 0.577 kcal/mol when SSD proceeds from Teacher to Student 13 and is reduced further at Student 29 (0.549 kcal/mol). The accuracy of Student 35 (Mean of RMSEs: 0.546 kcal/mol) is slightly higher than Student 29, with a lower standard deviation among ten folds (0.023 and 0.021 kcal/mol for Student 29 and 35, respectively). Of note, the held-out test set from **Aug-DB**s was not considered for the evaluation since the sizes of **Aug-DB**s are different for all Student models.

All the above results demonstrate the effectiveness of SSD in improving the accuracy of the ground-truth experimental solubilities while augmenting the database. Our results are consistent with the recent studies which mathematically proved that a few rounds of SSD often improve the accuracy for the held-out data and reduce overfitting.[73, 74] It was verified that the self-distillation amplifies the regularization of the space of trainable parameters if the model architectures for Teacher and Students are identical, as shown in Fig. 1. When Students are trained using an extensive distilled database with a limited parameter space, their variance is reduced without significantly increasing its bias. In other words, the models' trainable parameters neither change too sensitively to different training set batches nor are biased to specific batches, therefore overfitting is reduced. Meanwhile, too many rounds of SSD may over-regularize the model, leading to underfitting. These mathematical findings are in line with our SSD results shown in Fig. 2.

Of note, we also 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

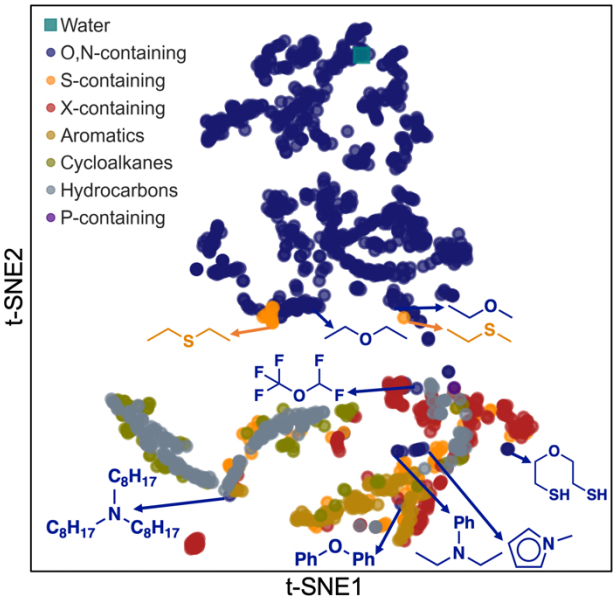


Figure 5. 2D plot of t-distributed stochastic neighbor embeddings (t-SNEs) for the latent vectors of 1,447 solvents obtained from Student 35 model.

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.[65, 66] 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 was 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 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 35 model. 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. 5. 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 identified certain clustering patterns among several categories: O,N-containing (upper side), halogen (X)-containing (mainly lower right), and hydrocarbon solvents (mainly lower left). O,N-containing solvents exclusively occupy a specific 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 in the vicinity of other molecular groups, such as aromatics, hydrocarbons, and X-containing ones. These solvents contain oxygen or nitrogen with the other atoms corresponding to the molecular groups they are close to. For instance, trioctylamine is in the cluster of Hydrocarbons, since it has three alkyl chains having eight carbons per each. Pentafluorodimethyl ether was found adjacent to the X-containing cluster. Ethers, amines, and pyrroles with aromatic rings are placed around the group of aromatic solvents. Meanwhile, an ether with two thiol groups (2-mercaptoethyl ether) was found near S-containing solvents rather than O,N-containing ones, indicating that presumably, their behavior as a solvent is close to S-containing solvents rather than O,N-containing ones. Conversely, some sulfides (diethyl sulfide, ethyl methyl sulfide) are near their ether analogs, and we can assume that their chemical behavior could be analogous to that of ethers.

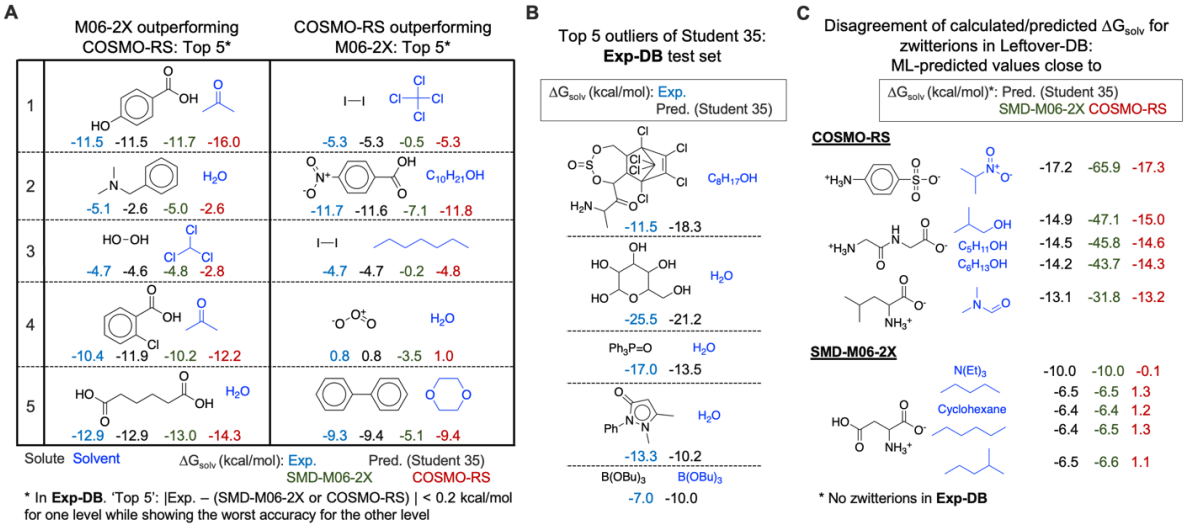


Figure 6. (A) Top 5 solute-solvent pairs in **Exp-DB** where the SMD-M06-2X/Def2-TZVP level outperforms COSMO-RS in calculating ΔG_{solv}, and vice versa. (B) Top 5 outliers of the Student 35 model when comparing the predicted ΔG_{solv} with those in the test set of **Exp-DB**. (C) In **Leftover-DB**, the disagreement of ΔG_{solv} among Student 35, SMD-M06-2X and COSMO-RS mainly occurs for zwitterion solutes which do not exist in **Exp-DB**.

Error analysis.

As introduced in Fig. 1F, the error analysis was performed by comparing the **QM-DB** solubilities calculated in the SMD-M06-2X/Def2-TZVP level of theory with those from COSMO-RS, experiments, and

the Student 35 model. First, we analyzed **Exp-DB** solute-solvent pairs where SMD-M06-2X outperforms COSMO-RS and vice versa to identify the advantages and disadvantages of each theoretical method (Fig. 6A). The left side of Fig. 6A illustrates the five cases whose absolute error between ΔG_{solv} from experiment and SMD-M06-2X does not exceed 0.2 kcal/mol, whereas COSMO-RS shows the worst performance. All these five cases correspond to polar solutes and solvents with halogen atoms, hydrogen bond donors and acceptors. SMD-M06-2X better reproduces the experimental solubilities of these molecules than COSMO-RS, which may be in part attributed to the halogenicity, hydrogen bond acidity, and basicity parameters used by SMD. Three out of five predictions (**1**, **3**, and **5**) from Student 35 are also close to the **Exp-DB** solubilities rather than those from COSMO-RS, indicating that the distillation process (Fig. 1D) effectively corrected the discrepancy between experiment and theory except for some cases.

There are the other five solute-solvent pairs for which COSMO-RS outperforms SMD-M06-2X (Right side of Fig. 6A). In contrast to the former case discussed above, they are solutes and solvents with low or no polarity or molecules with special moieties such as ozone. COSMO-RS accurately evaluates the ΔG_{solv} of these molecules. Investigating the two extreme cases shown in Fig. 6A implies the importance of accounting for multiple theoretical methods in assessing the results from SSD. The analysis on SMD-M06-2X and COSMO-RS was then followed by the outlier analysis of Student 35 (Fig. 6B). The outliers correspond to the gray dots in the box plot shown in Fig. 3B; it can be deduced from their extraordinary chemical structures that they do not strikingly deteriorate the model's accuracy. The top five outliers of Student 35's prediction against **Exp-DB** include the solutes with multiple complex rings, five hydroxy groups, and heteroatoms (P and B) that rarely appear in the whole database (932,509 data points). For example, the solutes with a P=O double bond and aromatic substituents appear only in 808 data points, and only 69 data points have solutes/solvents with B-O single bonds.

Further analysis was performed for **Leftover-DB** consisting of 57,721 solute-solvent pairs in **CombiSolv-QM** that were not included in the **Aug-DB**s but remained after iterating the SSD cycles 35 times. Since **Leftover-DB** does not have experimental values, we compared their ΔG_{solv} values from Student 35, SMD-M06-2X and COSMO-RS for 14,053 out of 57,721 data points whose ΔG_{solv} from all the three models or methods were available. As a result, significant discrepancies among these three computational protocols were observed for the 1,381 data points having zwitterionic solutes. Ten extreme cases are shown in Fig. 6C. SMD-M06-2X relatively overestimates ΔG_{solv} compared to the other two for the above five cases, whereas the ΔG_{solv} from COSMO-RS shows disagreement with Student 35 and SMD-M06-2X for the below five ones. It should be emphasized that no zwitterions are available in **Exp-DB**; although 5,446 zwitterions were already included in **Aug-DB**s during the SSD, there are no experimental ground-truth ΔG_{solv} values for these species. Such a lack of data availability for zwitterions necessitates experimental measurements for their solubility values or the incorporation of additional reliable theoretical methods, possibly leading to a more extensive database from SSD, including zwitterions.

Although the above error analysis suggests room for improving our model, it is sufficiently reliable to be utilized in the practical design of solvent systems in various chemical processes such as catalysis and

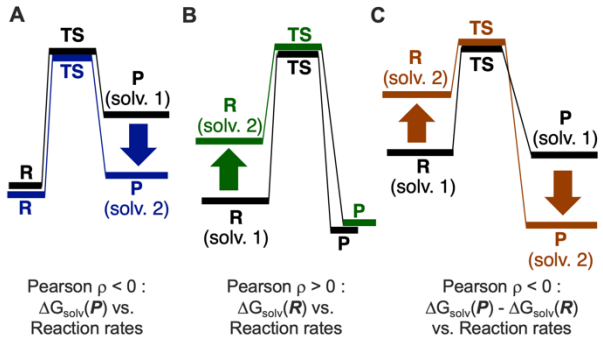
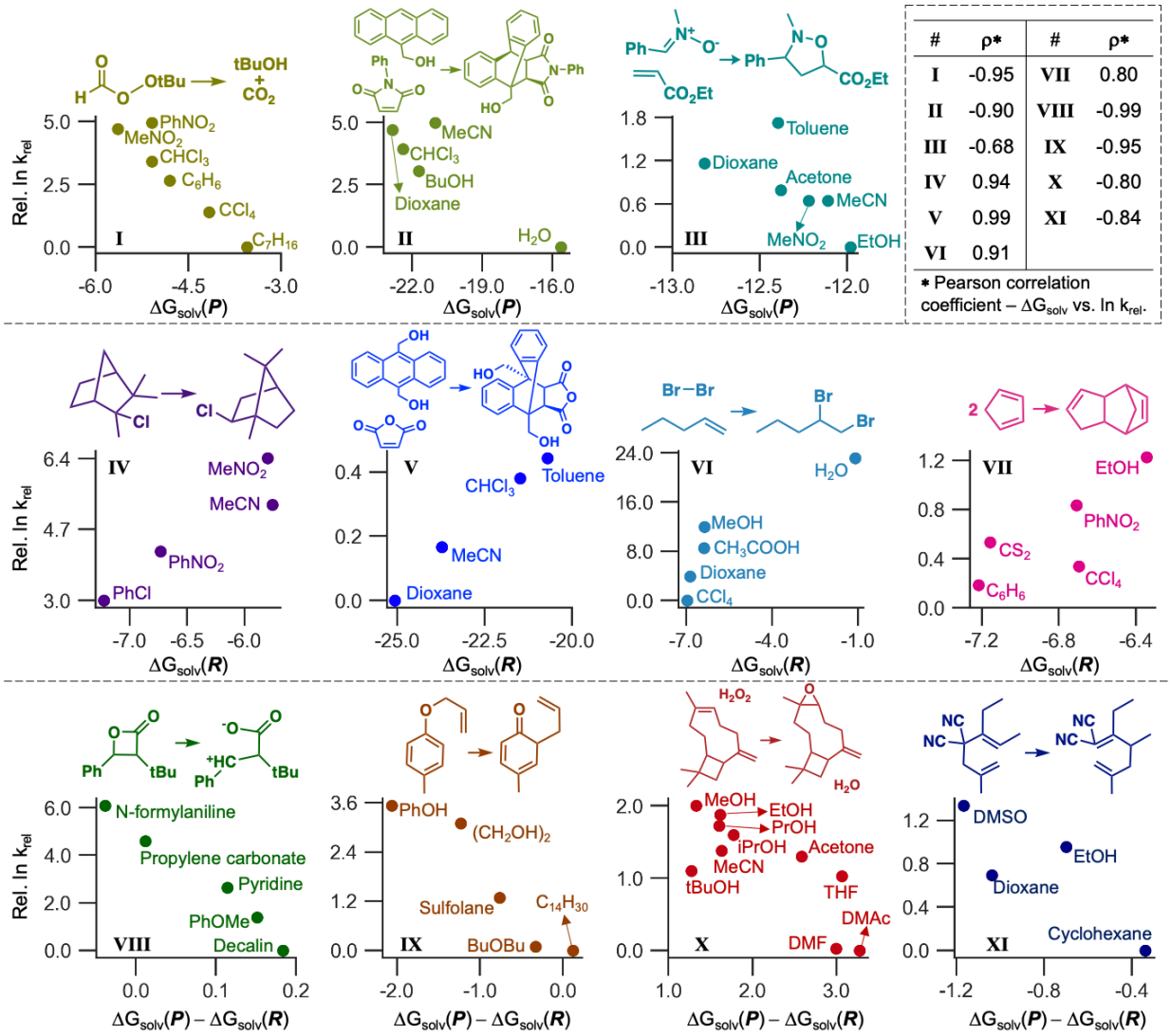


Figure 7. Schematic energy diagrams of the reactions where (A) higher solvation stabilization of the product leads to more product formation, and (B) higher solvation destabilization of the reactant leads to faster reaction. (C) Designing new solvents can also affect both relative free energies of both reactant and product and thus the reaction rates.

separation. The next sections demonstrate the application of our model to two examples: (1) Elucidation of linear relationships between reaction rates and solvation free energies, and (2) prediction of partition coefficients for biomass-derived chemicals and drug-like molecules.

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

It is crucial to find linear solvation free energy relationships (LSERs) 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 of organic reactions through ML. For 11 organic reactions, Gibbs solvation free energies of the product(s) and reactant(s) ($\Delta G_{solv}(\boldsymbol{P})$ and $\Delta G_{solv}(\boldsymbol{R})$, respectively) were predicted by using our GNN model (Student 35). If a reaction has two reactants or products, the sum of their ΔG_{solv} values was used as $\Delta G_{solv}(\boldsymbol{R})$ or $\Delta G_{solv}(\boldsymbol{P})$. These values were used as the descriptors to find highly positive or



#	ρ*	#	ρ*
I	-0.95	VII	0.80
II	-0.90	VIII	-0.99
III	-0.68	IX	-0.95
IV	0.94	X	-0.80
V	0.99	XI	-0.84
VI	0.91		

* Pearson correlation coefficient – ΔG_{solv} vs. ln k_{rel}.

Figure 8. The linear relationships between ΔG_{solv} of reactants/products predicted by our GNN model versus experimental reaction rates for 11 organic reactions from the literature. The reactant and product of each chemical reaction are shown above the graph depicting the linear relationship. The solvents used in the reactions are written next to each data point. The Pearson correlation coefficients (ρ) between ΔG_{solv} and logarithms of relative reaction rates are listed in the upper-right table.

negative Pearson correlation coefficients ρ (i.e., close to 1 or -1) between ΔG_{solv} of reactants/products and experimental reaction rates in different solvents. The reaction rates were collected from the literature.[2, 75-77]

In principle, a negative correlation should be found for $\Delta G_{solv}(\boldsymbol{P})$ vs. reaction rates for the reactions where the solvation stabilization of the product(s) plays a key role in accelerating product formation. Fig. 7A illustrates a schematic energy diagram of such reactions. A lower $\Delta G_{solv}(\boldsymbol{P})$ in one solvent than in the other indicates more product stabilization, and thus, more formation, leading to a negative Pearson ρ between $\Delta G_{solv}(\boldsymbol{P})$ and reaction rates. Meanwhile, for some reactions, higher reaction rates can be achieved by more destabilization of reactants by a solvent [i.e., higher $\Delta G_{solv}(\boldsymbol{R})$] compared to the other (Fig. 7B). Such a correlation leads to a positive Pearson ρ between $\Delta G_{solv}(\boldsymbol{R})$ and reaction rates. The cases in Figs. 7A and 7B mainly occur when the structure of a transition state is analogous to that of reactant(s) and product(s), respectively, according to the Hammond Postulate. In addition, product stabilization and reactant destabilization can be considered together by using $\Delta G_{solv}(\boldsymbol{P}) - \Delta G_{solv}(\boldsymbol{R})$ as a descriptor (Fig. 7C). Changing the sign of $\Delta G_{solv}(\boldsymbol{R})$ and adding to $\Delta G_{solv}(\boldsymbol{P})$ enables the quantification of the influences on the reaction rates by both reactants and products, resulting in a negative Pearson ρ with reaction rates.

Fig. 8 depicts the results of investigating the above three descriptors (Fig. 7) on 11 organic reactions with varying solvents. For each reaction, we chose one descriptor that shows the best correlation. $\Delta G_{solv}(\boldsymbol{P})$ is the best descriptor for the reactions **I**, **II**, and **III**, with Pearson ρ values of -0.95, -0.90, and -0.68, respectively. The reaction **I** is the dissociation of tert-butylperoxyaldehyde into tert-butyl alcohol and CO_2. Our results suggest that solvents with higher polarity (nitrobenzene, nitromethane, and chloroform) better stabilize the tert-butyl alcohol product than non-polar solvents (benzene, tetrachloromethane and heptane) and thus show higher rates. Lower Pearson ρ values were obtained from the other two reactions compared to **I**, but they display good negative correlations except for one solvent (MeCN and toluene for **II** and **III**, respectively).

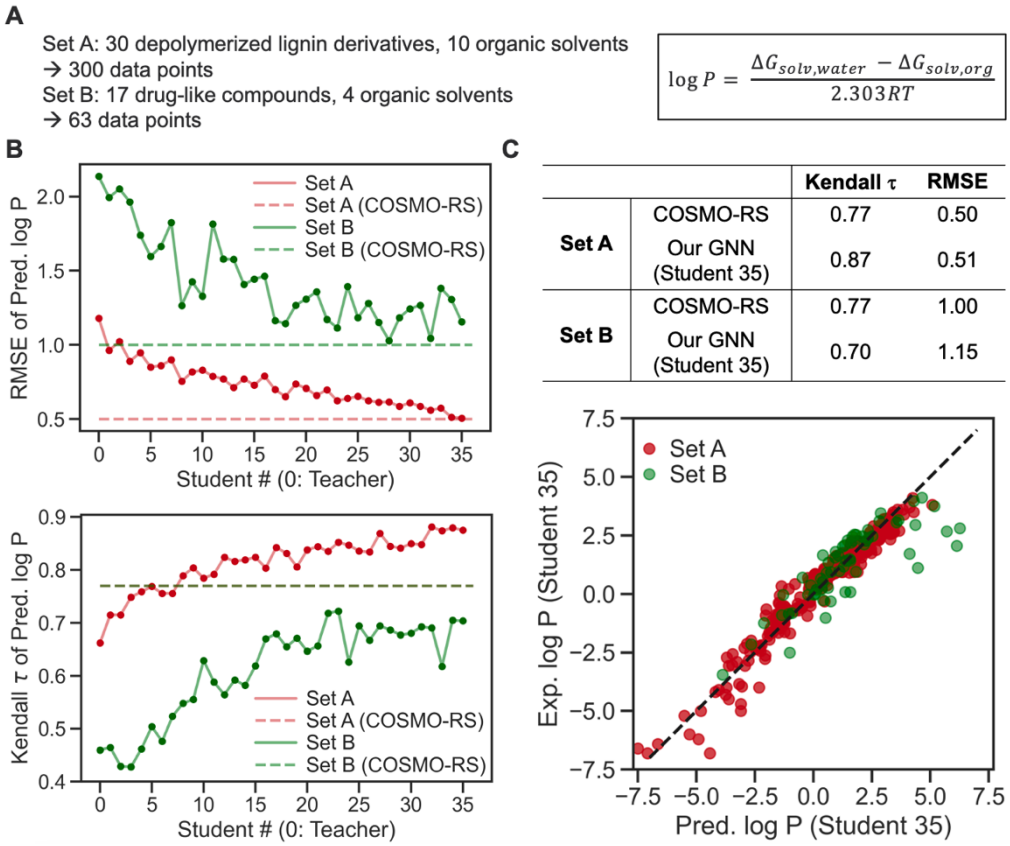
On the other hand, the solvent effects of reactions **IV~VII** are accurately described when the positive correlations between $\Delta G_{solv}(\boldsymbol{R})$ and reaction rates are evaluated. Their Pearson ρ values range from 0.80 to 0.99. Of note, reaction **V** is analogous to **II** except for having more polar reactants than **II**. In this case, using non-polar solvents such as toluene show the most reactant destabilization and the highest reaction rate. The effect of different functional groups for the same reaction was captured by our GNN model, leading to the identification of a strong positive correlation (ρ=0.99). In contrast, high reaction rates were achieved when polar solvents such as water or ethanol were used with non-polar reactants (Br_2, pentene, and cyclopentadiene) for reactions **VI** and **VII**, respectively.

The rest four reactions (**VIII~XI**) can be explained by $\Delta G_{solv}(\boldsymbol{P}) - \Delta G_{solv}(\boldsymbol{R})$ as a descriptor (ρ = -0.99 ~ -0.80). Reaction **VIII** is a ring opening to decarboxylate the reactant and form an alkene whose reaction rates were measured in five solvents. A non-polar 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. Next, the Cope rearrangement (**IX**), 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 (ρ=-0.80) in the complex reaction example, such as the epoxidation of β-caryophyllene investigated in 10 different solvents (**X**).

One can gain insights into the solvent design for maximizing reaction rates by accurately predicting $\Delta G_{solv}(\boldsymbol{R})$ and $\Delta G_{solv}(\boldsymbol{P})$ using the fast GNN model. 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. Although ΔG_{solv} of transition states are not considered here, our model enables rapid solvent screening before the investigation of the transition states. 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. 8 already show decent correlations.

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



		Kendall τ	RMSE
Set A	COSMO-RS	0.77	0.50
	Our GNN (Student 35)	0.87	0.51
Set B	COSMO-RS	0.77	1.00
	Our GNN (Student 35)	0.70	1.15

Figure 9. (A) Description of experimental log P datasets (Set A and Set B) for the application of the GNN solubility model to the prediction of log P, and the formula to evaluate log P from ΔG_{solv}. (B) Prediction accuracies of log P for Teacher and Student models when RMSE and Kendall rank correlation coefficient (τ) are used as metrics. (C) The table comparing the accuracies of Student 35 and COSMO-RS models, and a parity plot showing the experimental vs. predicted log P values for a total of 363 data points.

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.[78] The datasets of log P values consist of two sets (Set A and Set B, Fig. 9A). 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. Here, we predicted the ΔG_{solv} values in water and organic solvents for the solute-solvent pairs in Set A and Set B and evaluated log P values using the formula shown in Fig. 9A. To further verify the feasibility of SSD, we repeated these ML predictions of log P for the Teacher and 35 Student models. Then, two metrics were used to assess their accuracies: RMSE and Kendall rank correlation coefficient (τ). A τ value closer to 1 indicates a stronger rank correlation when ranks of experimental and predicted log P are labeled from the lowest to highest, and thus, higher accuracy of the model.

These metrics were chosen because the literature[78] used the same metrics when the accuracy of COSMO-RS was assessed. We compared the accuracies of our Teacher and Student models from SSD with the COSMO-RS method (Fig. 9B). In terms of RMSEs for Set A, the latter Student models show lower RMSEs than the former ones and Teacher. Eventually, Student 35 achieves an RMSE as low as

that from COSMO-RS, which manifests the effectiveness of the SSD scheme in predicting log P values. The RMSEs for Set B fluctuate among Student models more than Set A, presumably due to the complexity of drug-like molecules in Set B. Nonetheless, the accuracy comparable to COSMO-RS was achieved after undergoing 35 SSD cycles. It should be emphasized that our GNN models exceed the accuracy of COSMO-RS if the accuracy for Set A is evaluated in terms of Kendall τ. After Student 7, the τ values of Student models are already higher than that from COSMO-RS. The τ values of our GNN models are slightly lower than COSMO-RS for Set B (0.70 and 0.77 for Student 35 and COSMO-RS, respectively). However, overall, it displays an increasing trend as SSD proceeds, indicating the strength of SSD.

We compared the accuracy of our Student 35 model with log P calculated using COSMO-RS. The table in Fig. 9C summarizes Kendall tau rank coefficients and RMSEs for COSMO-RS and our ML model. The GNN showed rank coefficients of 0.87 and 0.70 for Set A and Set B, respectively, whereas those for COSMO-RS are 0.77 for both sets.[78] Our GNN resulted in a better correlation for Set A than COSMO-RS, while COSMO-RS performed slightly better in Set B. In terms of RMSE, our model achieved the RMSE almost identical to that from COSMO-RS for Set A. COSMO-RS showed better accuracy in Set B. Fig. 9C depicts the parity plot of ML-predicted log P vs. experimental ones. Overall, the model shows predictions close to experimental ones. Log P of some cases in Set B are overestimated, but similar outliers were also found from the COSMO-RS results.[78] For future work, accuracies for these data points can be improved by the consideration of the ΔG_{solv} for ionic species for predicting distribution coefficients (log D) for acidic or basic solute molecules.

All these results indicate that our ML model reliably captures solubility trends and accurately predicts log P values in different organic solvents. 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. Meanwhile, 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 potentially improved by employing multiple QM methods during the data augmentation process. Considering temperature effects on solubility in ML models should also be pursued in the future 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.

Methods

Computational details for calculating ΔG_{solv} using SMD-DFT.

The AQME Python package[79] was used throughout the overall process for calculating ΔG_{solv} values of given solute-solvent pairs. First, the canonicalized SMILES strings of solutes were converted into 3D geometries, and conformational searches were carried out by employing the MMFF94s force field[80] implemented in the RDKit cheminformatics library.[81] The number of generated conformers was determined based on the number of rotatable bonds. The lowest-energy conformer was then chosen and subject to further geometry optimizations using DFT with the SMD implicit solvation model. The recent study reported that considering only the most stable conformer is sufficient to obtain the energy value

close to the Boltzmann-weighted ensemble average of multiple conformers for organic molecules.[59] The subsequent geometry optimizations were performed using the M06-2X/Def2-TZVP method with the SMD. Of note, only 3D structures of solutes were optimized, and solvents were specified by their name in the input file. While the SMD is available for any solvents whose descriptor values are available (dielectric constant, refractive index, surface tension, etc.), calculations were performed for only the solvents available in the Gaussian 16 package.[82]

The optimized structures were confirmed as valid if there are no imaginary frequencies and they did not undergo the decomposition into disconnected molecules. If the structure is not valid or it was not fully converged, we assumed that the SMD-DFT cannot properly simulate the corresponding solute-solvent pair, and it was discarded. To calculate ΔG_{solv} from the optimized geometry, the external iteration method in Gaussian 16 was utilized, which considers the self-consistent solvent reaction field to calculate solute's electrostatic potential. These calculations were carried out in the same level of theory, with specifying the keywords 'Externaliteration' and '1stVac' in the Gaussian 16 input file.

Development of graph neural networks with SSD.

The GNN models were developed using Python 3.7[83] with TensorFlow 2.4,[84] Keras 2.9,[85] and Neural Fingerprint (NFP)[86] 0.3.0 libraries. The NFP library provides the framework for deep learning using message-passing GNN with the atom, bond, and global features (Fig. 1A) generated through the RDKit cheminformatics package.[81] The stochastic depth method was implemented by employing TensorFlow-Addons 0.14 to examine the effect of introducing noises to message-passing layers, although the SSD without noises showed the best prediction accuracy. The optimal GNN structure shown in Fig. 1A was determined by the hyperparameter tuning. We carried out an iterative grid search of possible combinations of different hyperparameters. These hyperparameters are the number of message-passing layers (3-6), dimension of hidden layer vectors (64, 128, and 256), learning rate ($a \cdot 10^{-b}$; a=1, 5, and b=3-5), batch size(2^n, n=7-10), and activation functions (Rectified linear unit – ReLU, and LeakyReLU). We trained the models against **Exp-DB** with different hyperparameters and identified the one that shows the best compromise between accuracy and computational cost, resulting in the model shown in Fig. 1A. During the SSD process, all Teacher and Student models were trained for 1,000 epochs with a learning rate of $1 \cdot 10^{-4}$, followed by 200 epochs with a learning rate of $5 \cdot 10^{-5}$, using a batch size of 1,024. The ADAM optimizer with the MAE loss function was employed.

Exp-DB and all **Aug-DB**s were split into the training, validation, and test sets with a ratio of 72:8:9. We adopted this ratio instead of the typical 8:1:1 ratio to perform 10-fold cross-validation with varying the training and validation sets. The training/validation set and training/test set ratios are 9:1 and 8:1, respectively, enabling the 10-fold partitioning while maintaining the held-out test set. The validation loss value was monitored at each epoch throughout the training to archive the best model with the lowest validation set error. It was sufficient to identify the best model when the model was trained for 1,200 epochs with two different learning rates mentioned above. Due to the high computational costs of cross-validation, only one of the 10 folds was utilized for the model training and data augmentation (Fig. 1D). However, the full 10-fold cross-validation was performed for Teacher, Students 13, 29, and 35 models (Fig. 3). This is to verify that the models are not prone to overfitting and the SSD scheme effectively reduces the deviation of prediction errors among different data splits. The model was trained using one GV100 GPU; the time taken for training ranges from 50 minutes (**Exp-DB**, 11,637 data points) to 1.7 days (**Exp-DB** + **Aug-DB**s, 932,509 data points).

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. Vassileiou, A. D.; Robertson, M. N.; Wareham, B. G.; Soundaranathan, M.; Ottoboni, S.; Florence, A. J.; Hartwig, T.; Johnston, B. F., A unified ML framework for solubility prediction across organic solvents. *Digital Discovery* **2023,** *2* (2), 356-367.
57. Lee, S.; Lee, M.; Gyak, K.-W.; Kim, S. D.; Kim, M.-J.; Min, K., Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* **2022,** *7* (14), 12268-12277.
58. 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. *Digital Discovery* **2023,** *2* (2), 409-421.
59. 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.
60. 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**.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.

66. 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.
67. He, J.; Gu, J.; Shen, J.; Ranzato, M. A., Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788* **2019**.
68. Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. A., Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chemical Science* **2022,** *13* (5), 1446-1458.
69. 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.
70. 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.
71. 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**.
72. 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**.
73. Mobahi, H.; Farajtabar, M.; Bartlett, P., Self-distillation amplifies regularization in hilbert space. *Adv. Neural Inf. Process Syst.* **2020,** *33*, 3351-3361.
74. Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; Ma, K. In *Be your own teacher: Improve the performance of convolutional neural networks via self distillation*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019; pp 3713-3722.
75. Welton, T.; Reichardt, C., *Solvents and solvent effects in organic chemistry*. John Wiley & Sons: 2011.
76. 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.
77. Kiselev, V. D.; Kornilov, D. A.; Sedov, I. A.; Konovalov, A. I., Solvent Influence on the Diels-Alder Reaction Rates of 9-(Hydroxymethyl)anthracene and 9,10-Bis(hydroxymethyl)anthracene with Two Maleimides. *Int. J. Chem. Kinet.* **2017,** *49* (1), 61-68.
78. 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.
79. Alegre-Requena, J. V.; Sowndarya S. V, S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S., AQME: Automated quantum mechanical environments for researchers and educators. *WIREs Comput. Mol. Sci.* **2023**, e1663.
80. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996,** *17* (5-6), 490-519.
81. Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015,** *55* (12), 2562-2574.
82. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

83. Van Rossum, G. In *Python Programming Language*, USENIX annual technical conference, Santa Clara, CA: 2007; pp 1-36.
84. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. In *Tensorflow: a system for large-scale machine learning*, Osdi, Savannah, GA, USA: 2016; pp 265-283.
85. Gulli, A.; Pal, S., *Deep learning with Keras*. Packt Publishing Ltd: 2017.
86. St John, P. *NFP (Neural Fingerprint) 0.3.0.* https://github.com/NREL/nfp; National Renewable Energy Lab.(NREL), Golden, CO (United States): 2019.