Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3031

Predicting aqueous and organic solubilities with machine learning: a workflow for identifying organic cosolvents

Maurycy Krzyżanowski, Sirazam Munira Aishee, Nirala Singh * and Bryan R. Goldsmith *

Developing predictive models of solubility is useful for accelerating solvent selection for applications ranging from electrochemical conversion of organics to pharmaceutical drug development. Herein, we report on the development of a machine learning (ML) workflow for identifying organic cosolvents to increase the concentration of hydrophobic molecules in aqueous mixtures. This task is of particular interest for the electrocatalytic conversion of biomass and bio-oils into sustainable fuels, which faces challenges due to the low aqueous solubility of the feedstock. First, we predict the miscibility of potential cosolvents in water, and we only consider cosolvents that are miscible. Second, we rank cosolvents based on the predicted solubility of the molecule of interest in them. To achieve this, we train two separate ML models: one using the AqSolDB dataset to predict aqueous solubility, and another using the BigSolDB dataset to predict solubility in organic solvents. We select the Light Gradient Boosting Machine (LGBM) model architecture for aqueous solubility (test R^2 = 0.864, RMSE = 0.851 for log(S (mol^{-1} dm^{-3}))) and organic solubility (test R^2 = 0.805, RMSE = 0.511 for log(x)) predictions based on comparing different ML models and features. We examine the generalizability of the organic solubility model on unseen solutes both quantitatively and qualitatively. We evaluate the utility of this ML workflow by identifying cosolvents for benzaldehyde and limonene—two hydrophobic molecules that are relevant for sustainable fuel production—and validate our predictions *via* experimental solubility measurements.

Received 7th May 2025
Accepted 11th September 2025

DOI: 10.1039/d5dd00134j

rsc.li/digitaldiscovery

Introduction

Solubility is a key property of interest in numerous fields, including oil and gas,[1–3] biomass conversion,[4–6] and pharmaceutical science.[7–11] However, the experimental determination of solubility is often challenging[12,13] as well as time-consuming. Therefore, developing predictive models for estimating solubility is highly beneficial. Applications of predictive solubility models include screening pharmaceuticals based on their aqueous solubility and selecting suitable solvents for organic molecules of interest. The importance of predicting the solubility of molecules in water (aqueous solubility) has driven the development of many models, starting with simple linear equations like the General Solubility Equation[14] and Estimated SOLubility.[15]

Recently, machine learning (ML) models have emerged as the preferred approach for predicting aqueous solubility because of their capacity for high accuracy and generalizability.[16–22] ML models for predicting the solubility of organic molecules in organic solvents (organic solubility) have also been developed,[23–26] although they are less common than those for aqueous solubility. Curated datasets of solubility are becoming available for developing robust ML models. The AqSolDB dataset[27] for aqueous solubility of mostly organic molecules is the most prominent among these datasets. However, the recently created BigSolDB dataset[28] makes it now possible to develop ML models for organic solubility predictions.

One of the many possible applications of solubility models is the identification of organic cosolvents, which are important in areas such as pharmaceutical molecule development[29–31] and electrochemical hydrogenation of hydrophobic molecules such as those found in essential oils and bio-oils.[32–35] In the case of the latter, it is often desirable that the system contains water to serve as an abundant proton source for the reaction. However, the low solubility of hydrophobic molecules in pure water limits the range of concentrations in which the reaction can be performed. One of the solutions to this problem is having an organic serving as a cosolvent with water.

Developing an ML workflow that aids in the identification of organic cosolvents could greatly increase the effectiveness and usefulness of reactions involving hydrophobic molecules in aqueous systems. We consider a good cosolvent to be an organic solvent that (1) is miscible with water and (2) forms a water/organic cosolvent mixture that solubilizes hydrophobic

Department of Chemical Engineering, University of Michigan, Ann Arbor, MI 48109-2136, USA. E-mail: snirala@umich.edu; bgoldsm@umich.edu

organic molecules. We assume that the enhanced solubility of a hydrophobic organic molecule in a water/organic cosolvent mixture correlates with how soluble the hydrophobic molecule is in the pure organic cosolvent. Such an assumption implies that the optimal cosolvent can be identified by developing two ML models: one for aqueous solubility, which identifies organic solvents miscible with water, and another for organic solubility, which determines which of these water-miscible organic solvents best dissolves the hydrophobic molecule. Building on this approach, we introduce an ML workflow that integrates these two ML models to systematically identify suitable organic cosolvents for target molecules.

We opted for such an approach—as opposed to developing an ML model that predicts solubility for ternary mixtures at any given concentration of a cosolvent—due to the limited availability of data for mixed-solvent systems compared to the freely available AqSolDB and BigSolDB datasets. Moreover, we avoided DFT-derived features, which are resource-intensive to calculate, and instead utilized group contribution methods[36] and molecular fingerprints—both derived directly from SMILES (Simplified Molecular Input Line Entry System) strings and computationally efficient for high-throughput screening.

We use existing machine learning architectures—Light Gradient Boosting Machine (LGBM) and Random Forest (RF)—to train ML models, and we demonstrate their applicability in a workflow for identifying organic cosolvents. We compare the performance of LGBM and RF on different sets of features to select the best-performing approach for the aqueous and organic solvent solubility predictions. Given the complexity of predicting organic solvent solubility, we evaluate the model to assess its ability to distinguish between good and poor organic solvents, predict solubility trends across organic solvents, exactly rank cosolvents, and provide quantitatively accurate solubility predictions. We evaluate the workflow's performance using our experimentally derived data and demonstrate its potential to discover organic cosolvents for two hydrophobic molecules of relevance to sustainable fuel production[37–40]—limonene and benzaldehyde. Table 1 summarizes the key findings of the organic solubility model and the cosolvent identification workflow's performance in terms of its reliability for different use cases. We also evaluate the aqueous solubility model for predicting the miscibility of organic solvents in water.

Methods

Data preprocessing

AqSolDB. The AqSolDB dataset contains 9982 standardized solubility measurements at room temperature (25 ± 5 °C). The AqSolDB dataset includes solutes that are either liquid or solid at room temperature.[27] Although the AqSolDB dataset contains mostly solubility values for organic compounds, it has inorganic salts as well. The AqSolDB dataset was filtered to contain only organic compounds. Multiple SMILES strings associated with the same InChIKey (textual identifier for chemical substances) were identified and replaced with the first SMILES string from the list. The aqueous solubility is expressed as $\log(S\ mol^{-1}\ dm^{-3})$, where S represents the molar concentration of the solute in water ($mol\ L^{-1}$). After preprocessing, the AqSolDB dataset contained 8549 data points.

BigSolDB. The BigSolDB dataset includes 54 273 measurements spanning a temperature range of −30.0 °C to 130.0 °C. Unlike the AqSolDB dataset, this dataset consists almost exclusively of data for solutes that are solid at room temperature. The organic solubility is expressed as $\log(x)$, where x denotes the mole fraction of the solute in the organic solvent. The BigSolDB dataset contains multiple solubility values for an array of temperatures, in contrast to the AqSolDB dataset, which contains a singular solubility value for measurements taken within the range of 25 ± 5 °C. Therefore, we excluded organic solubility measurements for temperatures outside this range for our BigSolDB dataset. If multiple solubility values existed within that range, their average value was used. This approach avoids bias in the performance evaluation as previously observed for multiple solute–solvent pairs at different temperatures.[23] Only organic solvents for which there are more than nine solute–solvent pairs were considered, because predictions made for solvents with fewer solute data points were assumed to be untrustworthy due to data scarcity. For example, if there is a solvent for which there are only five data points, such a solvent is removed from the dataset. After preprocessing, the BigSolDB dataset contained 4557 data points.

Features generation

We obtained the group contribution (GC) features using the RDKit[41] and Thermo[42] packages. For aqueous solubility, only the features of the solutes were generated, whereas for organic solubility, features were generated for both the solute and the solvent. Some GC features were obtained through the UNIFAC method.[43] We applied the Molecular ACCess System (MACCS)[44] as molecular fingerprints. MACCS features represent molecules as binary vectors, each bit indicating the presence or absence of specific chemical substructures. We extract both GC and MACCS features from the SMILES notation, which is a text-based format for representing molecular structures. The total number of GC features was 30 for aqueous solubility and 60 for

Table 1 Reliability of the organic solubility model and cosolvent identification workflow

Strengths	Caveats	Weaknesses
Distinguishes between good and poor organic solvents and predicts solubility trends for the organic solvent well. Test $R^2 = 0.805$, RMSE = 0.511 for $\log(x)$ on BigSolDB dataset.	Has moderate accuracy in unseen solvent ranking. Kendall's tau < 0.3 for most solutes.	Not suitable for quantitatively predicting solubility for unseen solutes. RMSE of $\log(x)$ > 0.3 for most solutes.

organic solubility. The total number of MACCS features was 167 for aqueous solubility and 334 for organic solubility. The group contribution features for the model training are listed in Table S1 along with the names of the Python packages used.

Model training, cross-validation, and testing

We evaluated two model architectures to predict aqueous solubility and organic solubility: LGBM and RF. Our model selection is justified by the fact that tree-based models often outperform deep learning models on medium-sized tabular datasets such as BigSolDB and AqSolDB.[45] For example, Lee and coworkers[46] demonstrated that the LGBM model trained on a combined organic and aqueous solubility dataset predicted solubility just as well as a graph convolutional network (GCN). The GCN did perform better than the LGBM model trained on just molecular fingerprints (*data01*; 17 536 data points), with R^2 values of 0.80 *versus* 0.74. However, in the case of a smaller dataset enriched with physicochemical features—akin to the datasets used here (*data03*; 6945 data points)—the LGBM actually performed better than the GCN model (R^2 0.85 *vs.* 0.81). Ye and Ouyang[23] reported that an LGBM model trained on organic solubility data comprising 5081 data points performed better than a deep neural network model (DNN) in generalization capability: for unseen solutes, LGBM achieved an R^2 of 0.49, significantly outperforming the DNN model ($R^2 = 0.22$). In the case of aqueous solubility, Boobier and coworkers[24] benchmarked various ML architectures against the AquaSol model[22] (an undirected recursive neural network) and found that their Extra Trees model ($R^2 = 0.93$) outperformed AquaSol ($R^2 = 0.86$) on the *water_set_wide* dataset. It was also shown that a RF model outperformed ChemProp,[47] a well-established graph neural network.[48] The authors compared both models trained on the AqSolDB dataset, with the RF achieving an RMSE of 0.86 *versus* 0.89 for ChemProp.[48]

The features were normalized using min–max scaling before training an ML model. A train/test split was performed with a training set size of 80%. The split was stratified based on the target value (solubility). Nested five-fold cross-validation (CV), also stratified based on solubility, was carried out on the training set to identify the optimal hyperparameters. The score metrics used for the evaluation of the model on the training set, during nested cross-validation, and on the test set were the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The AqSolDB training set contained 6839 data points, and the test set contained 1710 data points. The BigSolDB training set contained 3645 data points, and the test set contained 912 data points.

Feature importance

We employed two methods for feature importance analysis. The first method is split feature importance analysis for LGBM, which estimates the importance of a given feature by measuring how often it is used to split nodes in the decision trees that make up the model. The other one is SHapley Additive exPlanations (SHAP).[49] Both methods were applied to the LGBM model trained on the BigSolDB dataset. We compared the feature importance originating from all the GC features *vs.* the ones from all the MACCS features for the GC-MACCS model. We also analyzed the feature importance of the GC features within the GC-MACCS model to see which GC features contribute the most to its performance (Fig. S1).

Workflow for organic cosolvent identification

The main components of the organic cosolvent identification workflow are summarized in Fig. 1. The MACCS and GC fingerprints used as input features are summarized in Fig. 1a. The workflow involves creating two ML models based on the LGBM architecture. One model is for aqueous solubility trained on the AqSolDB dataset, and the other for organic solubility trained on the BigSolDB dataset (Fig. 1b). The organic cosolvents are identified *via* a two-step process (Fig. 1c) with these two models. The first step involves using the aqueous solubility model to determine the miscibility of solvents in water so that water-immiscible solvents can be removed. Having identified a pool of water-miscible solvents, the second step involves using an organic solubility model to predict the solubility of the hydrophobic molecule in these solvents.

Results and discussion

Performance of ML models

We systematically assess the performance of RF and LGBM architectures and the GC, MACCS, and GC-MACCS feature sets for predicting aqueous and organic solubility. The performance of the models is evaluated using three metrics: R^2, RMSE, and MAE, to determine which model architecture and features are superior for aqueous solubility, $\log(S\ mol^{-1}\ dm^{-3})$, and organic solubility, $\log(x)$. Parity plots (Fig. 2) illustrate the performance of the LGBM models on the aqueous and organic solubility datasets, with the values of the performance metrics (R^2, RMSE, and MAE) for the test set displayed in the inset. The same kind of parity plots are provided for the RF model in the SI (Fig. S2). The RF models had slightly worse score metrics than the LGBM models on the test set and in the nested five-fold CV for both datasets. Therefore, the focus of the model performance evaluation across different sets of features—MACCS, GC, and GC-MACCS—is on the LGBM model. All numerical data for training, nested five-fold cross-validation (CV), and testing are provided in Tables S2 and S3, with Table S2 containing results for the aqueous solubility model and Table S3 for the organic solubility model. The hyperparameters used for training the models are reported in Table S4 and Table S5. Overall, the LGBM model trained on GC-MACCS features is our best-performing model for both aqueous solubility and organic solubility predictions, as shown in the parity plots in Fig. 2c and f. The aqueous solubility model achieved a five-fold CV R^2 score of 0.863 and a test score of 0.864 (Fig. 2c). This test score of 0.864 is comparable to the previous studies that trained ML models on the AqSolDB dataset.[20,48,50,51] Namely, the model outperformed the SolTranNet model reported by Francoeur and Koes,[20] who validated their model on the same aqueous solubility dataset, AqSolDB. The base architecture of the model is the Molecule Attention Transformer. The authors reported an

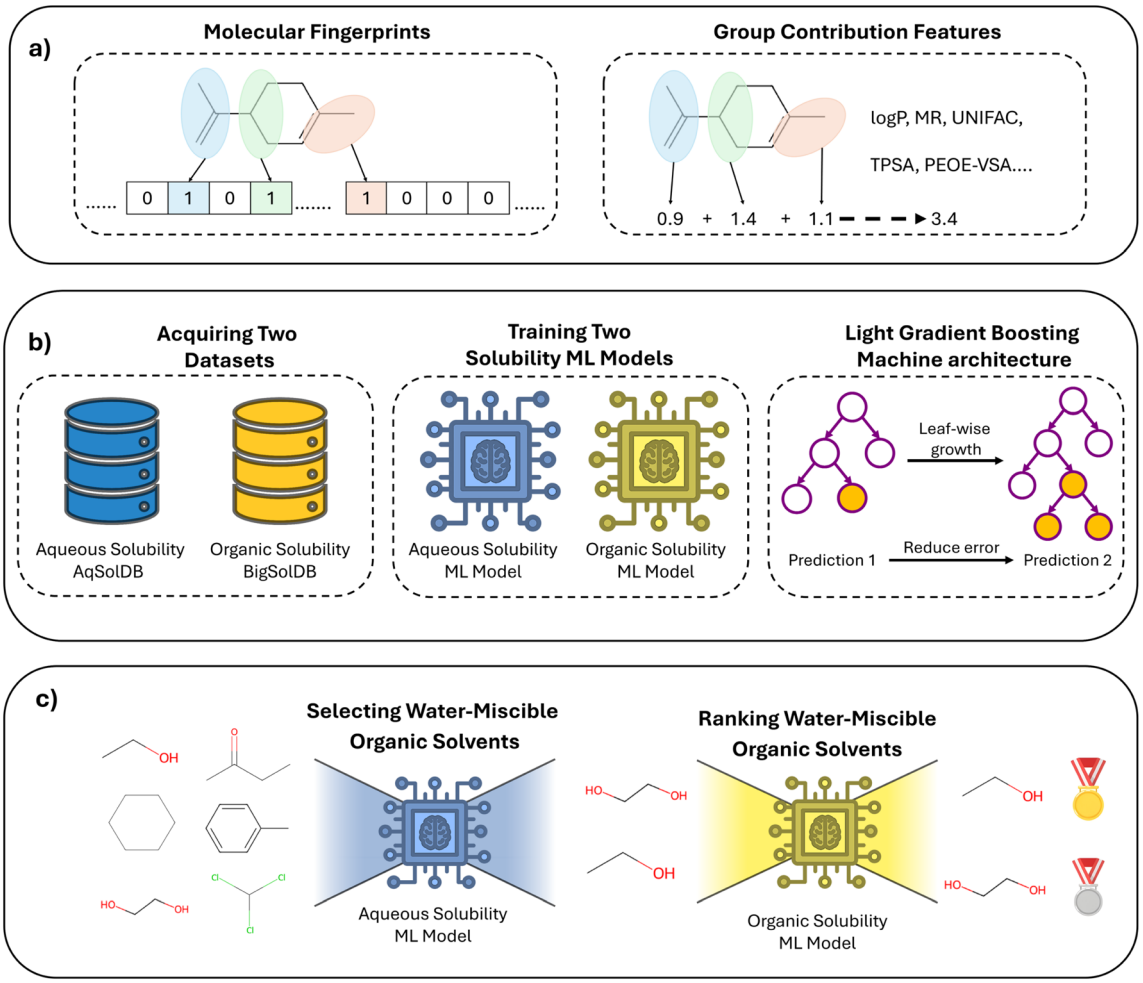

Fig. 1 Workflow for the organic cosolvent identification. (a) Representation of the MACCS and GC features used for training ML models. (b) Two ML models were used: one trained on the AqSolDB dataset to predict aqueous solubility, and the second trained on the BigSolDB dataset to predict organic solubility. This section also includes the schematics of the Light Gradient Boosting Machine architecture. (c) The workflow for organic cosolvent identification. The aqueous solubility model removes water-immiscible solvents, while the organic solubility model ranks the remaining solvents from best to worst based on the solubility of the molecule of interest.

R^2 score of 0.68 for the 3-fold scaffold split cross-validation. Llompart and coworkers[48] provided a comprehensive summary of solubility models reported in the literature. In this report, we can compare the performance results of other ML models trained on the AqSolDB dataset. For example, Sluga and coworkers reported a testing R^2 score of 0.93 and RMSE of 0.59 for their neural network model[50] while Falcón-Cano and coworkers reported a R^2 score of 0.72 and RMSE of 0.73 for their RF model.[52] In the case of the organic solubility model, a direct comparison with the previous studies is not possible because of differences in datasets and data pre-processing. Nonetheless, the herein reported nested five-fold CV R^2 score of 0.787 and a test R^2 score of 0.805 is comparable to previous studies.[23,25]

Establishing a water-miscibility threshold using predicted aqueous solubility

While it is not always necessary for the organic solvent to be fully miscible in water, having a miscible organic solvent is preferred as it offers greater flexibility, as any proportion of water to solvent can be used. Because our aqueous solubility model does not directly predict whether a solvent is miscible with water, we instead sought to identify an aqueous solubility threshold above which solvents can be classified as miscible. To create a classification model to identify if an organic solvent is miscible or immiscible in water, we obtained a small water miscibility/immiscibility dataset of 26 solvents obtained from the Sigma-Aldrich solvent miscibility table.[53] We predict the aqueous solubility for each of these solvents (Table S6) using the model trained on the AqSolDB dataset, ensuring that the 26 solvents were excluded from the training set to enable proper validation. To establish the water-miscibility cutoff value, we train a support vector machine classifier with a linear kernel, which enables straightforward determination of the miscibility threshold as the intercept divided by the coefficient of the linear equation. To estimate the 90% confidence interval of the threshold, we performed bootstrapping, where the data were sampled with replacement 5000 times and each time the model

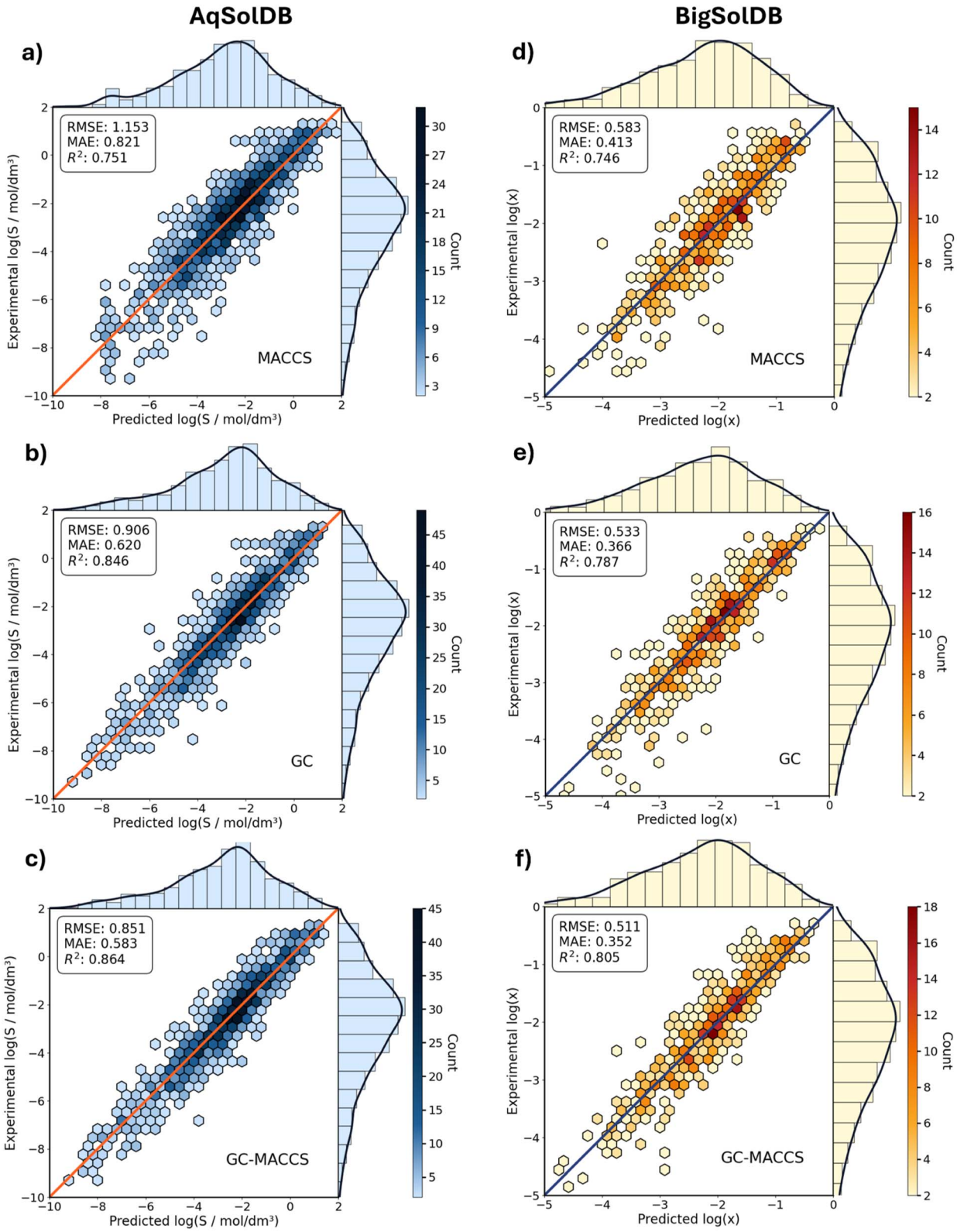

Fig. 2 Parity plots showing the performance of the LGBM models on the test set. The model was trained on (a) AqSolDB dataset with MACCS features, (b) AqSolDB dataset with GC features, (c) AqSolDB dataset with GC-MACCS features, (d) BigSolDB dataset with MACCS features, (e) BigSolDB dataset with GC features, and (f) BigSolDB dataset with GC-MACCS features. The axes are the experimental *vs.* predicted values of solubility of an organic solvent in water (left column, log(S mol^{-1} dm^{-3})) or solubility of a molecule of interest in an organic solvent (right column, log(*x*)) along with the distribution of the values next to the respective axes. The color of the hexagons corresponds to the number of counts within each of the hexagonal bins. The diagonal solid line denotes a perfect prediction. Shown RMSE, MAE and R^2 values inset correspond to scores on the test set.

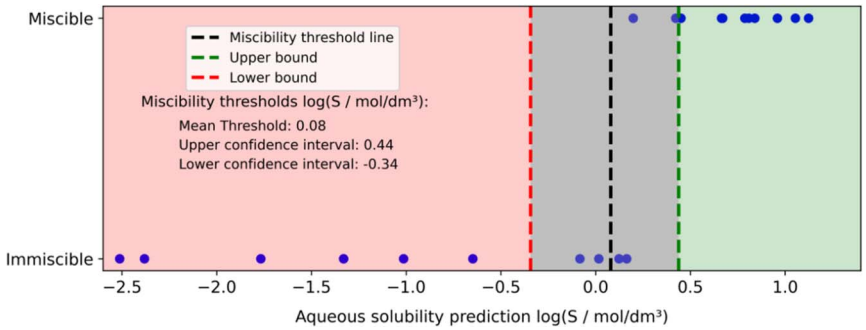

Fig. 3 Determination of the miscibility threshold value based on aqueous solubility. Each data point represents a solvent that is either miscible or immiscible in water, which is indicated on the *y*-axis, whereas the *x*-axis shows the aqueous solubility prediction values from our model. The black vertical line indicates the miscibility threshold. The red and green vertical lines represent the lower and upper bounds, respectively. The exact values of the miscibility threshold, along with the lower and upper bounds, are displayed in the figure. For the sake of visual clarity, data points with predicted solubility lower than $\log(x) = -2.6$ were excluded from the plot but are given in Table S6.

was fitted to subsampled data and the threshold value was obtained. Each subsample contained the same number of data points (26) as the original dataset. The mean of these values is the final aqueous miscibility threshold and has a value of log(S mol^{-1} dm^{-3}) equal to 0.08, corresponding to S of 1.2 mol dm^{-3}. The upper and lower bounds of the 90% confidence interval were obtained by calculating the 5th and 95th percentiles of the threshold values, respectively. The lower bound of log(S mol^{-1} dm^{-3}) is −0.34 and the upper bound is to 0.44. Fig. 3 shows the predicted aqueous solubility values of the organic solvents and their miscibility in water.

Using the identified miscibility threshold, we achieved a classification accuracy of 92%. As such, solvents with solubilities below the lower bound are classified as immiscible, while the solvents above the upper bound are miscible. Within those bounds (marked as the gray area) are solvents that are highly soluble in water but immiscible, as well as miscible solvents that were misclassified.

Evaluation of organic solubility trends and model generalizability

Although the models achieved satisfactory results based on quantitative performance metrics, it is important to evaluate directly if the models can make appropriate predictions for previously unseen solutes. Vassileiou and coworkers[25] observed a performance drop from $R^2 = 0.78$ for ten-fold CV for single solute–solvent pairs to $R^2 = 0.56$ for the leave-one-solute-out CV when predicting organic solubility. Such an observation highlights the potential issues with the generalizability of the models for organic solubility. As previously stated, some of these issues can be attributed to the experimental uncertainty of solubility measurements, where the standard deviation of log(S mol^{-1} dm^{-3}) can reach 0.5.[54] Another potential issue is the uneven distribution of data points across solvents in BigSolDB as both the number of measurements per solvent and the diversity of solutes in each solvent vary substantially. Consequently, the development of quantitatively accurate organic solubility models is challenging. However, we hypothesize that an ML model should be capable of capturing the solubility trends, that is, being able to rank the solvents from best to worst.

Thirty molecules were selected for a case study analysis from BigSolDB. These molecules were chosen through stratified sampling, which was based on the log(*P*) (logarithm of the octanol–water partition coefficient) and aliphatic ratio of the molecules to ensure a representative distribution of lipophilicity and aromaticity among the molecules. Because our experimental focus is on small molecules, before sampling we filter large molecules with the number of non-hydrogen atoms exceeding fifteen. Additionally, only molecules with at least seven solvents were included to ensure our case study molecules span a diverse set of solvents. For example, if a molecule has solubility data for only four solvents, the predictions lack both chemical diversity and reliable ranking assessment by the model. Stratified sampling was performed using a total of 16 two-dimensional bins, formed by combining four bins for log(*P*) and four bins for the aliphatic ratio. The number of bins was determined based on Sturges' rule.[55] We note that the number of solvents available for each molecule varies, as this is inherent to the BigSolDB dataset. For each case study molecule, a new LGBM model is trained on GC-MACCS features to test its generalization performance. Two metrics are used to assess the performance of the model. RMSE is used to determine how far the solubility prediction is from experimental values. We used Kendall's Tau[56] to quantify the ability of the model to accurately rank the solvents from best to worst. The value range of this metric is from 0 to 1, where the value of zero means that ranking lists are identical, and the value of one means that ranking lists are complete opposites.

As an example of this analysis, we predict the five best solvents for two molecules—maleic anhydride (Fig. 4a) and oxindole (Fig. 4b). While only the five best solvents are shown, the reported RMSE values and the values of Kendall's Tau were obtained for the set of all organic solvents. The ranked solvents are shown for all thirty molecules in Fig. S3 and the distribution of RMSE values and Kendall's Tau for those molecules are shown in Fig. 4c. The analysis reveals that while the RMSE values can be non-ideal, the model is able to predict the

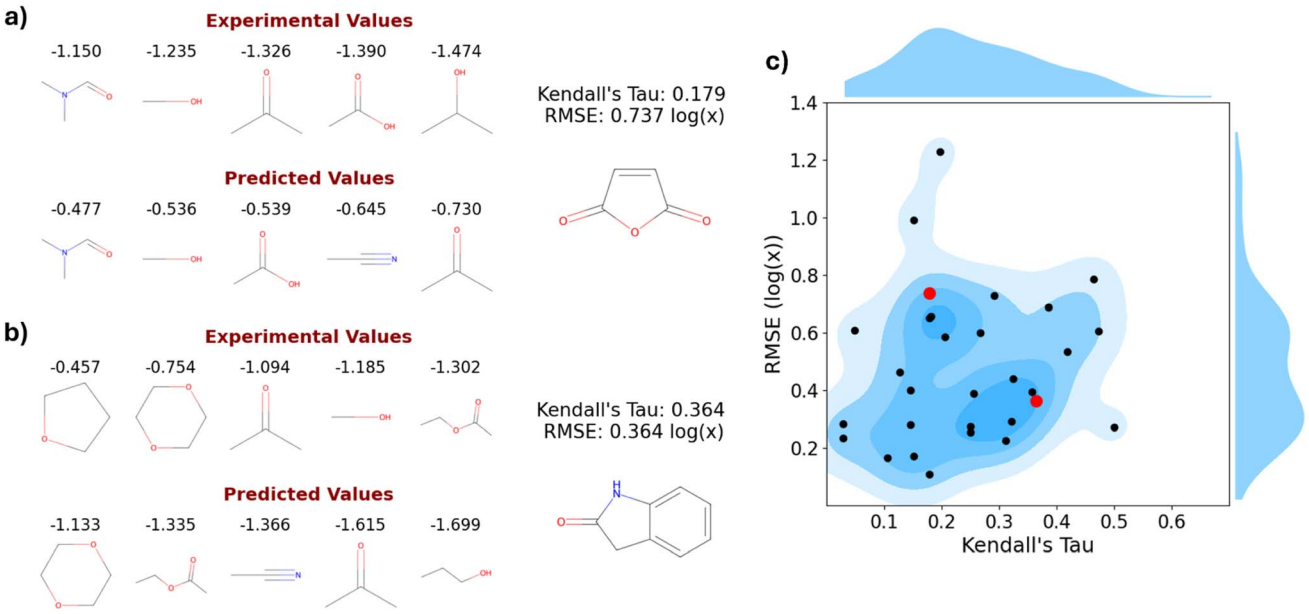

Fig. 4 Case study of organic solubility predictions for maleic anhydride and oxindole. The five solvents with the highest solubility values are shown for (a) maleic anhydride and (b) oxindole along with their predicted solubility values, while (c) shows the distribution of the Kendall's Tau and RMSE values obtained for studied molecules, with maleic anhydride and oxindole emphasized as red dots. Data points were overlaid onto kernel density estimates, which portray the two-dimensional distribution of data points. On the top and on the right of the plot, the distributions of Kendall's Tau and RMSE values, respectively, are shown.

solubility trends across solvents well, with most of the predictions yielding a Kendall's Tau lower than 0.3 and only one prediction slightly exceeding 0.5. This metric highlights the ability of the model to rank the solvents from best to worst and demonstrates that such models are capable of qualitatively screening organic solvents. Maleic anhydride is an example where the model ranked the solvents well (Kendall's Tau of 0.179), but the solubility predictions themselves are not ideal (RMSE of $\log(x) = 0.737$). Oxindole, on the other hand, is an example where the model performed worse at ranking the solvents (Kendall's Tau of 0.364), but the predicted solubility values were closer to the actual values (RMSE of $\log(x) = 0.364$).

Identifying cosolvents for molecules of interest

Having established that the miscibility threshold obtained from the aqueous solubility predictions can filter solvents that are immiscible in water, the next step to identify organic cosolvents is based on the solubility of the molecules of interest in those solvents. As case studies, here we focus on limonene and benzaldehyde because of their relevance as precursors to valuable fuels and chemicals. We note that both of these molecules are liquid at room temperature, whereas the BigSolDB dataset contains mostly compounds that are solid at room temperature. This contrast gives the opportunity to explore how well a model trained on the solubility of solid-state molecules translates to the solubility of liquid molecules. A pool of 46 organic solvents was obtained from the BigSolDB dataset for screening. The data in Fig. 5 shows the aqueous solubility of these organic solvents and the solubility of limonene (Fig. 5a) and benzaldehyde (Fig. 5b) in those organic solvents. The previously determined aqueous miscibility threshold, along with the confidence intervals, is also displayed. Out of all the organic solvents, 15 were found to be above the upper bound of miscibility. The water-miscible solvents are ranked from best to worst based on the predicted solubility of the limonene or benzaldehyde in them. Because we do not expect the solubility in the pure organic to quantitatively match the solubility in the organic-water mixture, we use a ranking system rather than the predicted solubility value. Our approach also assumes that organic solubility is proportional to solubility in organic solvent/water mixtures, which as we discuss below is not always the case due to non-idealities in mixing. The organic solubility predictions of limonene and benzaldehyde in the water-miscible organic solvents are given in Table S7. To test our predictions, we experimentally investigated the solubility of limonene and benzaldehyde in organic solvent/water mixture systems. Data points marked in blue (Fig. 5) are the organic solvents that we selected for experimental validation. We selected the solvents that are within the miscible area, commercially available, and to represent a distribution of their predicted solubility (*i.e.*, not only the highest ranked organic solvents). We chose the volume percent of the organic solvent/water mixture systems for experimental testing as described in the methods section of the SI. To account for potential composition-dependent effects, we conducted solubility measurements at two different solvent compositions. Briefly, for a given organic volume percent (V%) in water the concentration of limonene or benzaldehyde was increased until the solution became cloudy (*i.e.*, the cloud-point method).[57] Examples of experimental solubility determination by the cloud-point method are shown in Fig. S4. We found that 80 mM limonene in 75 V% acetic acid/water (Fig. S4a) and

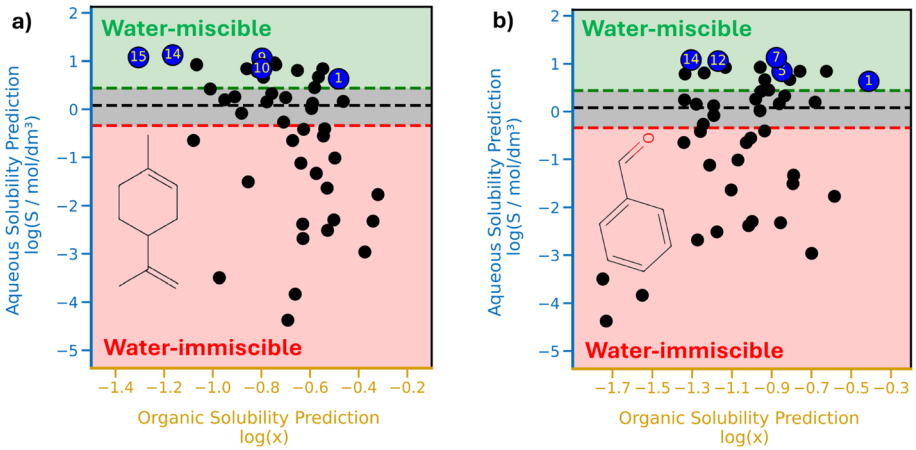

Fig. 5 Organic solvent selection. Each data point represents a single organic solvent, with the *x*-axis corresponding to the predicted solubility value of the (a) limonene and (b) benzaldehyde in the organic solvent, whereas the *y*-axis corresponds to the aqueous solubility prediction of the organic solvent. The background color is related to the miscibility of the solvents in water—red indicates immiscible, whereas green indicates miscible. The small area colored gray represents classification uncertainty and contains solvents that are immiscible in water but have high solubility in water. The area also contains solvents that were potentially misclassified as immiscible. Solvents that were selected for the experimental validation study are marked by the large, blue dots, with numbers that correspond to their rank, which is based on the solubility of limonene or benzaldehyde in those solvents.

140 mM benzaldehyde in 15 V% acetic acid/water (Fig. S4c) produced clear solutions, whereas a 20 mM increment in both limonene (Fig. S4b) and benzaldehyde (Fig. S4d) concentration gave cloudy solutions. Consequently, we noted the highest concentration of both limonene and benzaldehyde without any cloudiness for maximum solubility in the acetic acid/water mixture as 80 mM and 140 mM, respectively. To account for the potential variability of the experiment, we conducted a total of three measurements for each case of molecule of interest and each V% of organic solvent/water mixture. While for some mixtures we observed minor variability, for most cases the observed solubility was the same. The recorded solubility values are given in Table S8.

The experimental solubility measurements for different organic volume percents in water for limonene and benzaldehyde are shown in Fig. 6. The order of the studied solvents, from top to bottom, corresponds to their predicted organic solubility rank from Fig. 5. For limonene, the model accurately predicted the solubility trends for ethanol, acetic acid, methanol, and ethylene glycol mixtures in water (Fig. 6a). However, the solubility in the DMA/water mixture is lower than that of the other organic cosolvents (except for ethylene glycol) for 75 V% mixtures, despite the model predicting DMA to be the second best overall organic solvent for limonene. Interestingly, for 90 V% mixtures, the solubility in the DMA mixture is higher than that in acetic acid, methanol, and ethylene glycol. As discussed in the methods, we did not observe the solubilization of 20 mM limonene in 25 V% and 50 V% organic solutions for any of the solvents tested.

For benzaldehyde, the model correctly predicted the solubility trends for acetic acid, methanol, ethanol and ethylene glycol mixtures in water (Fig. 6b). However, just like for limonene, DMA was predicted to be the best organic cosolvent for benzaldehyde, but acetic acid was found to be a slightly better cosolvent for the 25 V% mixture, while the solubility was the same for the 15 V% mixture. The severity of the error in the prediction for DMA is much less than in the case of limonene, as the model correctly predicted that DMA and acetic acid are better solvents than methanol, ethanol, and ethylene glycol. However, it is worth noting that the solubility of benzaldehyde in methanol/water is the same as in ethanol/water, despite the model predicting methanol to rank 7 and ethanol to rank 12. As such, a better solubility of benzaldehyde in the methanol/water mixture would be anticipated.

A possible cause for the deviation of the prediction and the experimental results, aside from potential inaccuracies in the model and visual inspection error in experiments, comes from non-idealities of mixing. A common method to estimate the solubility of a species in a mixture of two solvents is to use a weighted average of the solubility in either individual solvent, such that the solubility in a mixture of any composition of cosolvents is a linear interpolation between the two pure solvent's solubilities. However, in many instances,[58] this will not be the case and solubilities in the mixtures may either be higher or lower than that predicted by the weighted average. The observations regarding the solubility of benzaldehyde in acetic acid/water and DMA/water mixtures highlight the non-linear relationship between the volume percentage of the cosolvent and the solubility. The non-linear behavior of the solubility of limonene in the DMA/water mixture provides additional evidence of the limitations of our approach. We also note that the model was trained primarily on the solubility of solid organic compounds due to availability of data, while both benzaldehyde and limonene are liquid organic compounds. Not only does the experimental methodology for determining solubility differ between solids and liquids, but the nature of their solubility also varies. In contrast to solid solubility, the solute phase in liquids often also contains diffused solvent.

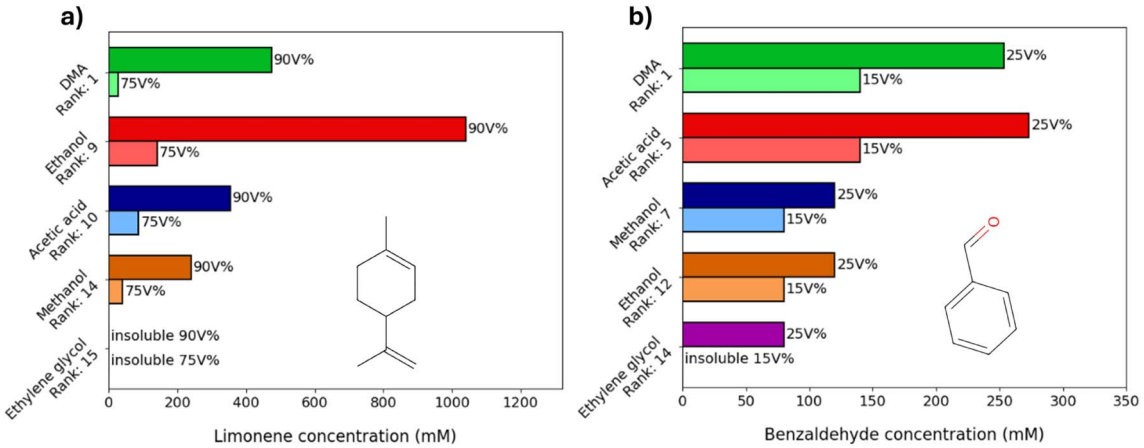

Fig. 6 Semi-quantitative measurements of maximum solubility of molecules of interest. The measurements were done for (a) limonene in 75 V% and 90 V% organic solvent/water and (b) benzaldehyde in 15 V% and 25 V% organic solvent/water mixtures. The height of the bar corresponds to the mean value of the three solubility measurements recorded for the organic solvent/water mixture as indicated on the *x*-axis. The percentage of the organic solvent in the mixture is displayed on top of the bar. The *y*-axis contains the name of the organic solvent. Measurements were made at room temperature. Insoluble refers to the solubility by the cloud-point method being below 20 mM for limonene and below 80 mM for benzaldehyde, as described in the methods section. The variance of the measurements is relatively low, and, as such, the error bars are not shown in the figure, but all the solubility measurements are given in Table S8. The low variance is due to the measurements being made in relatively large increments of 20 mM *via* the cloud-point method.

Although the organic solubility model was shown to perform well in terms of distinguishing between good and poor solvents, it is important to acknowledge that, while having most of the Kendall's Tau values below 0.3 indicates good screening abilities, it does not mean that the model is able to perfectly predict the rank of the solvent, as that would correspond to Kendall's Tau values closer to zero.

Conclusions

We developed an ML workflow for organic cosolvent identification. Creating the ML workflow involved training two independent ML models, one for each of the two solubility datasets: AqSolDB (aqueous solubility) and BigSolDB (organic solubility). We demonstrated strong cross-validation and testing performance for both solubility models. The generalizability of the organic solubility model was further assessed on unseen solutes. We demonstrated that, while the organic solubility model struggles with quantitative solubility predictions, it qualitatively predicts solubility trends. As such, the organic solubility model can distinguish between good and poor organic solvents.

Comparison between our ML workflow predictions and experimental results showed that while the ranking of solvents varied slightly between compositions, the differences were not substantial. The workflow accurately predicted solubility trends for mixtures of water with ethanol, acetic acid, and ethylene glycol. However, it overestimated the solubilities of limonene in *N*,*N*-dimethylacetamide (DMA) and benzaldehyde in methanol. These results demonstrate that while the approach of using two stand-alone ML models effectively screens for miscible organic solvents and distinguishes between good and poor cosolvents, it struggles with accurately ranking them. Beyond model errors, this limitation also stems from our assumption of a linear correlation between solubility in cosolvent-water mixtures and solubility in pure cosolvents, which does not account for non-idealities in solubility behavior within organic/water mixtures. Additionally, since the organic solubility model was trained on mostly solid solutes, its performance may be less reliable for liquid solutes like limonene and benzaldehyde, as the physicochemical interactions governing solubility can differ significantly between solid and liquid phases.

Author contributions

M. K. designed, implemented, and tested the ML models and performed all data analysis. M. K. wrote the initial draft of the manuscript. B. R. G provided project guidance and advising to M. K. since the project inception and helped write the manuscript. S. M. A. performed the experiments and helped write the experimental section and methods. N. S. advised S. M. A. and helped write the manuscript.

Conflicts of interest

The authors declare no competing interests.

Data availability

The source code of the ML models and the datasets used in this study are available on GitHub at **https://github.com/MKrzyzan2070/Predicting-aqueous-and-organic-solubilities-with-machine-learning** and archived on Zenodo at **https://doi.org/10.5281/zenodo.17074657**.

Supplementary information: experimental solubility measurements; ML model benchmarking metrics and

hyperparameters; and feature importance analysis. See DOI: https://doi.org/10.1039/d5dd00134j.

Acknowledgements

This work was supported by the Office of Naval Research, project # N00014-23-1-2439.

References

1 H. Chen, S. Chen, X. Quan, Y. Zhao and H. Zhao, Solubility and Sorption of Petroleum Hydrocarbons in Water and Cosolvent Systems, *J. Environ. Sci.*, 2008, **20**(10), 1177–1182, DOI: 10.1016/S1001-0742(08)62206-8.

2 J. Tomasek and J. Schatz, Olefin Metathesis in Aqueous Media, *Green Chem.*, 2013, **15**(9), 2317–2338, DOI: 10.1039/C3GC41042K.

3 X. Zhang, B. Li, F. Pan, X. Su and Y. Feng, Enhancing Oil Recovery from Low-Permeability Reservoirs with a Thermoviscosifying Water-Soluble Polymer, *Molecules*, 2021, **26**(24), 7468, DOI: 10.3390/molecules26247468.

4 L. Shuai and J. Luterbacher, Organic Solvent Effects in Biomass Conversion Reactions, *ChemSusChem*, 2016, **9**(2), 133–155, DOI: 10.1002/cssc.201501148.

5 G. R. Dick, A. O. Komarova and J. S. Luterbacher, Controlling Lignin Solubility and Hydrogenolysis Selectivity by Acetal-Mediated Functionalization, *Green Chem.*, 2022, **24**(3), 1285–1293, DOI: 10.1039/D1GC02575A.

6 P. Weerachanchai, S. K. Kwak and J.-M. Lee, Effects of Solubility Properties of Solvents and Biomass on Biomass Pretreatment, *Bioresour. Technol.*, 2014, **170**, 160–166, DOI: 10.1016/j.biortech.2014.07.057.

7 K. T. Savjani, A. K. Gajjar and J. K. Savjani, Drug Solubility: Importance and Enhancement Techniques, *Int. Scholarly Res. Not.*, 2012, **2012**, e195727, DOI: 10.5402/2012/195727.

8 S. J. Franklin, U. S. Younis and P. B. Myrdal, Estimating the Aqueous Solubility of Pharmaceutical Hydrates, *J. Pharm. Sci.*, 2016, **105**(6), 1914–1919, DOI: 10.1016/j.xphs.2016.03.040.

9 M. A. Filippa and E. I. Gasull, Experimental Determination of Naproxen Solubility in Organic Solvents and Aqueous Binary Mixtures: Interactions and Thermodynamic Parameters Relating to the Solvation Process, *J. Mol. Liq.*, 2014, **198**, 78–83, DOI: 10.1016/j.molliq.2014.06.031.

10 A. M. Vargason, A. C. Anselmo and S. Mitragotri, The Evolution of Commercial Drug Delivery Technologies, *Nat. Biomed. Eng.*, 2021, **5**(9), 951–967, DOI: 10.1038/s41551-021-00698-w.

11 D. M. Mudie, G. L. Amidon and G. E. Amidon, Physiological Parameters for Oral Delivery and *In Vitro* Testing, *Mol. Pharm.*, 2010, **7**(5), 1388–1405, DOI: 10.1021/mp100149j.

12 S. B. Murdande, M. J. Pikal, R. M. Shanker and R. H. Bogner, Aqueous Solubility of Crystalline and Amorphous Drugs: Challenges in Measurement, *Pharm. Dev. Technol.*, 2011, **16**(3), 187–200, DOI: 10.3109/10837451003774377.

13 A. Veseli, S. Žakelj and A. Kristl, A Review of Methods for Solubility Determination in Biopharmaceutical Drug Characterization, *Drug Dev. Ind. Pharm.*, 2019, **45**(11), 1717–1724, DOI: 10.1080/03639045.2019.1665062.

14 N. Jain and S. H. Yalkowsky, Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes, *J. Pharm. Sci.*, 2001, **90**(2), 234–252, DOI: 10.1002/1520-6017(200102)90:2<234::AID-JPS14>3.0.CO;2-V.

15 J. S. Delaney, ESOL: Estimating Aqueous Solubility Directly from Molecular Structure, *J. Chem. Inf. Comput. Sci.*, 2004, **44**(3), 1000–1005, DOI: 10.1021/ci034243x.

16 Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh and T. Hou, Chemistry-Intuitive Explanation of Graph Neural Networks for Molecular Property Prediction with Substructure Masking, *Nat. Commun.*, 2023, **14**(1), 2585, DOI: 10.1038/s41467-023-38192-3.

17 W. Ahmad, H. Tayara, H. Shim and K. T. Chong, SolPredictor: Predicting Solubility with Residual Gated Graph Neural Network, *Int. J. Mol. Sci.*, 2024, **25**(2), 715, DOI: 10.3390/ijms25020715.

18 C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction, *J. Chem. Inf. Model.*, 2017, **57**(8), 1757–1772, DOI: 10.1021/acs.jcim.6b00601.

19 S. Lee, H. Park, C. Choi, W. Kim, K. K. Kim, Y.-K. Han, J. Kang, C.-J. Kang and Y. Son, Multi-Order Graph Attention Network for Water Solubility Prediction and Interpretation, *Sci. Rep.*, 2023, **13**(1), 957, DOI: 10.1038/s41598-022-25701-5.

20 P. G. Francoeur and D. R. Koes, SolTranNet–A Machine Learning Tool for Fast Aqueous Solubility Prediction, *J. Chem. Inf. Model.*, 2021, **61**(6), 2530–2536, DOI: 10.1021/acs.jcim.1c00331.

21 Q. Chen, Y. Zhang, P. Gao and J. Zhang, An Interpretable Graph Representation Learning Model for Accurate Predictions of Drugs Aqueous Solubility, *Artif. Intell. Chem.*, 2023, **1**(2), 100010, DOI: 10.1016/j.aichem.2023.100010.

22 A. Lusci, G. Pollastri and P. Baldi, Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules, *J. Chem. Inf. Model.*, 2013, **53**(7), 1563–1575, DOI: 10.1021/ci400187y.

23 Z. Ye and D. Ouyang, Prediction of Small-Molecule Compound Solubility in Organic Solvents by Machine Learning Algorithms, *J. Cheminf.*, 2021, **13**(1), 98, DOI: 10.1186/s13321-021-00575-3.

24 S. Boobier, D. R. J. Hose, A. J. Blacker and B. N. Nguyen, Machine Learning with Physicochemical Relationships: Solubility Prediction in Organic Solvents and Water, *Nat. Commun.*, 2020, **11**(1), 5753, DOI: 10.1038/s41467-020-19594-z.

25 A. D. Vassileiou, M. N. Robertson, B. G. Wareham, M. Soundaranathan, S. Ottoboni, J. A. Florence, T. Hartwig and B. F. Johnston, A Unified ML Framework for Solubility Prediction across Organic Solvents, *Digital Discovery*, 2023, **2**(2), 356–367, DOI: 10.1039/D2DD00024E.

26 F. H. Vermeire, Y. Chung and W. H. Green, Predicting Solubility Limits of Organic Solutes for a Wide Range of

Solvents and Temperatures, *J. Am. Chem. Soc.*, 2022, **144**(24), 10785–10797, DOI: 10.1021/jacs.2c01768.

27 M. C. Sorkun, A. Khetan and S. Er, AqSolDB, a Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds, *Sci. Data*, 2019, **6**(1), 143, DOI: 10.1038/s41597-019-0151-1.

28 L. Krasnov, S. Mikhaylov, M. Fedorov, and S. Sosnin, BigSolDB: Solubility Dataset of Compounds in Organic Solvents and Water in a Wide Range of Temperatures, *ChemRxiv*, 2023, preprint, DOI: 10.26434/chemrxiv-2023-qqslt.

29 J. W. Millard, F. A. Alvarez-Núñez and S. H. Yalkowsky, Solubilization by Cosolvents: Establishing Useful Constants for the Log–Linear Model, *Int. J. Pharm.*, 2002, **245**(1), 153–166, DOI: 10.1016/S0378-5173(02)00334-4.

30 P. L. Gould, M. Goodman and P. A. Hanson, Investigation of the Solubility Relationships of Polar, Semi-Polar and Non-Polar Drugs in Mixed Co-Solvent Systems, *Int. J. Pharm.*, 1984, **19**(2), 149–159, DOI: 10.1016/0378-5173(84)90157-1.

31 S. S. Solanki, L. K. Soni and R. K. Maheshwari, Study on Mixed Solvency Concept in Formulation Development of Aqueous Injection of Poorly Water Soluble Drug, *J. Pharm.*, 2013, **2013**, 678132, DOI: 10.1155/2013/678132.

32 C. Han, J. Zenner, J. Johny, N. Kaeffer, A. Bordet and W. Leitner, Electrocatalytic Hydrogenation of Alkenes with Pd/Carbon Nanotubes at an Oil–Water Interface, *Nat. Catal.*, 2022, **5**(12), 1110–1119, DOI: 10.1038/s41929-022-00882-4.

33 D. S. Santana, G. O. Melo, M. V. F. Lima, J. R. R. Daniel, M. C. C. Areias and M. Navarro, Electrocatalytic Hydrogenation of Organic Compounds Using a Nickel Sacrificial Anode, *J. Electroanal. Chem.*, 2004, **569**(1), 71–78, DOI: 10.1016/j.jelechem.2004.02.015.

34 S. Wu, J. Cheng, Y. Xiang, Y. Tu, X. Huang and Z. Wei, Electrochemical Semi-Hydrogenation of Adiponitrile over Copper Nanowires as a Key Step for the Green Synthesis of Nylon-6, *Chem. Sci.*, 2024, **15**(29), 11521–11527, DOI: 10.1039/D4SC02280G.

35 H. Chen, T. Peng, B. Liang, D. Zhang, G. Lian, C. Yang, Y. Zhang and W. Zhao, Efficient Electrocatalytic Hydrogenation of Cinnamaldehyde to Value-Added Chemicals, *Green Chem.*, 2022, **24**(9), 3655–3661, DOI: 10.1039/D1GC04777A.

36 S. Wildman and G. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 868–873, DOI: 10.1021/ci990307l.

37 U. Sanyal, S. F. Yuk, K. Koh, M.-S. Lee, K. Stoerzinger, D. Zhang, L. C. Meyer, J. A. Lopez-Ruiz, A. Karkamkar, J. D. Holladay, D. M. Camaioni, M.-T. Nguyen, V.-A. Glezakou, R. Rousseau, O. Y. Gutiérrez and J. A. Lercher, Hydrogen Bonding Enhances the Electrochemical Hydrogenation of Benzaldehyde in the Aqueous Phase, *Angew. Chem., Int. Ed.*, 2021, **60**(1), 290–296, DOI: 10.1002/anie.202008178.

38 L. Gong, C. Y. Zhang, J. Li, G. Montaña-Mora, M. Botifoll, T. Guo, J. Arbiol, J. Y. Zhou, T. Kallio, P. R. Martínez-Alanis and A. Cabot, Enhanced Electrochemical Hydrogenation of Benzaldehyde to Benzyl Alcohol on Pd@Ni-MOF by Modifying the Adsorption Configuration, *ACS Appl. Mater. Interfaces*, 2024, **16**(6), 6948–6957, DOI: 10.1021/acsami.3c13920.

39 G. Rubulotta, K. L. Luska, C. A. Urbina-Blanco, T. Eifert, R. Palkovits, E. A. Quadrelli, C. Thieuleux and W. Leitner, Highly Selective Hydrogenation of R-(+)-Limonene to (+)-p-1-Menthene in Batch and Continuous Flow Reactors, *ACS Sustainable Chem. Eng.*, 2017, **5**(5), 3762–3767, DOI: 10.1021/acssuschemeng.6b02381.

40 D. Valencia, E. Martinez-Hernandez, A. García and J. Aburto, Sustainable Hydrogenation of Limonene to Value-Added Products Using Cu–Ni Catalysts Supported on KIT-5, *J. Cleaner Prod.*, 2024, **434**, 140356, DOI: 10.1016/j.jclepro.2023.140356.

41 G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, R. Vianello, P. Sriniker, P. Gedeck, G. Jones, E. Kawashima, N. Schneider, D. Nealschneider, A. Dalke, T. Hurst, M. Swain, B. Cole, S. Turk, A. Savelev, A. Vaucher, M. Wójcikowski, I. Take, H. Faara, R. Walker, V. F. Scalfani, D. Probst, K. Ujihara, N. Maeder, A. Pahl, G. Godin, J. Lehtivarjo, F. Berenger, J. Strets and Biggs, RDKit 2024_03_1 (Q1 2024) Release Beta, *Zenodo*, 2024, DOI: 10.5281/zenodo.591637.

42 C. Bell CalebBell/Thermo, 2024, https://github.com/CalebBell/thermo accessed 2024-03-31.

43 A. Fredenslund, R. L. Jones and J. M. Prausnitz, Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures, *AIChE J.*, 1975, **21**(6), 1086–1099, DOI: 10.1002/aic.690210607.

44 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, Reoptimization of MDL Keys for Use in Drug Discovery, *J. Chem. Inf. Comput. Sci.*, 2002, **42**(6), 1273–1280, DOI: 10.1021/ci010132r.

45 L. Grinsztajn, E. Oyallon and G. Varoquaux, Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?, *arXiv*, 2022, preprint, arXiv:2207.08815, DOI: 10.48550/arXiv.2207.08815.

46 S. Lee, M. Lee, K.-W. Gyak, S. D. Kim, M.-J. Kim and K. Min, Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features *vs* Graph Convolutional Neural Networks, *ACS Omega*, 2022, **7**(14), 12268–12277, DOI: 10.1021/acsomega.2c00697.

47 E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, Chemprop: A Machine Learning Package for Chemical Property Prediction, *J. Chem. Inf. Model.*, 2024, **64**(1), 9–17, DOI: 10.1021/acs.jcim.3c01250.

48 P. Llompart, C. Minoletti, S. Baybekov, D. Horvath, G. Marcou and A. Varnek, Will We Ever Be Able to Accurately Predict Solubility?, *Sci. Data*, 2024, **11**(1), 303, DOI: 10.1038/s41597-024-03105-6.

49 S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, From Local Explanations to Global Understanding with Explainable AI for Trees, *Nat. Mach. Intell.*, 2020, **2**(1), 56–67, DOI: 10.1038/s42256-019-0138-9.

50 J. Sluga, K. Venko, V. Drgan and M. Novic, QSPR Models for Prediction of Aqueous Solubility: Exploring the Potency of Randic-Type Indices, *Croat. Chem. Acta*, 2020, **93**(4), 1d.
51 N. Ulrich, K. Voigt, A. Kudria, A. Böhme and R.-U. Ebert, Prediction of the water solubility by a graph convolutional-based neural network on a highly curated dataset, *J. Cheminf.*, 2025, **17**(1), DOI: **10.1186/s13321-025-01000-9**.
52 G. Falcón-Cano, C. Molina and M. Á. Cabrera-Pérez, ADME Prediction with KNIME: *In Silico* Aqueous Solubility Consensus Model Based on Supervised Recursive Random Forest Approaches, *ADMET & DMPK*, 2020, **8**(3), 251–273, DOI: **10.5599/admet.852**.
53 Solvent Miscibility Table, **https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/analytical-chemistry/purification/solvent-miscibility-table**, accessed 2025-03-17.
54 K. V. Balakin, N. P. Savchuk and I. V. Tetko, *In Silico* Approaches to Prediction of Aqueous and DMSO Solubility of Drug-like Compounds: Trends, Problems and Solutions, *Curr. Med. Chem.*, 2006, **13**(2), 223–241, DOI: **10.2174/092986706775197917**.
55 H. A. Sturges, The Choice of a Class Interval, *J. Am. Stat. Assoc.*, 1926, **21**(153), 65–66, DOI: **10.1080/01621459.1926.10502161**.
56 R. Fagin, R. Kumar and D. Sivakumar, Comparing Top k Lists, *SIAM J. Discrete Math.*, 2003, **17**(1), 134–160, DOI: **10.1137/S0895480102412856**.
57 D. F. Othmer, R. E. White and E. Trueger, Liquid-Liquid Extraction Data, *Ind. Eng. Chem.*, 1941, **33**(10), 1240–1248, DOI: **10.1021/ie50382a007**.
58 W. E. Acree and J. H. Rytting, Solubilities in Binary Solvent Systems II. The Importance of Non-Specific Interactions, *Int. J. Pharm.*, 1982, **10**(3), 231–238, DOI: **10.1016/0378-5173(82)90073-4**.