Organic Solubility Prediction at the Limit of Aleatoric Uncertainty

Lucas Attia[1†], Jackson W. Burns[1†], Patrick S. Doyle[1], William H. Green[1*]

[1]Chemical Engineering, MIT, Cambridge, MA.

*Corresponding author(s). E-mail(s): whgreen@mit.edu;
†These authors contributed equally to this work.

Abstract

Small molecule solubility is a critically important property which affects the efficiency, environmental impact, and phase behavior of synthetic processes. Experimental determination of solubility is a time- and resource-intensive process and existing methods for *in silico* estimation of solubility are limited by their generality, speed, and accuracy. This work presents two models derived from the FASTPROP and CHEMPROP architectures and trained on BigSolDB which are capable of predicting solubility at arbitrary temperatures for any small molecule in organic solvent. Both extrapolate to unseen solutes 2-3 times more accurately than the current state-of-the-art model and we demonstrate that they are approaching the aleatoric limit (0.5-1 $\mathbf{log}\,\boldsymbol{S}$), suggesting that further improvements in prediction accuracy require more accurate datasets. These models, collectively referred to as FASTSOLV, are open source, freely accessible via a Python package and web interface, highly reproducible, and up to 50 times faster than the next best alternative.

Keywords: solubility, deep learning, aleatoric uncertainty, QSPR

Small molecule solubility solubility refers to the extent to which a chemical species will dissolve into a surrounding solvent. The solubility of organic solids is an essential molecular property that impacts the efficiency,[1] environmental impact,[2, 3] and phase behavior[4] of synthetic processes. Solubility is crucial in wide-ranging chemical processes spanning length and time scales including membrane-based chemical separations,[5, 6] pharmaceutical design and discovery,[7] drug delivery and

formulation,[8] the environmental fate of per-and polyfluoroalkyl substances (PFAS)[9] and geological-scale dissolved organic carbon flux.[10] By convention, solubility S in $mol\ L^{-1}$ is expressed as $\log_{10} S$ since values can range over several orders of magnitude.

Experimental methods for determining solubility are notoriously time- and resource-intensive,[11] and are regarded as highly inaccurate with reported inter-laboratory experimental variability ranging between 0.5-1 $\log S$.[12–17] For variable-solvent datasets scraped from literature this variability defines the aleatoric limit - the 'irreducible error' which models cannot surpass without memorization. Given that solubility as a function of temperature - almost always positive monotonic - is often desired, experimental determination becomes even more onerous. The challenges of measuring solubility are particularly painful in pharmaceutical development, where organic solubility complicates synthesis and purification,[18] aqueous solubility limits *in vivo* efficacy,[19] and the solid state form of the drug confounds accurate measurement.[20] For these reasons *a priori* estimation of $\log S$ has long been of immense interest to the chemical sciences.

Methods have evolved from empirical group additivity correlations,[21, 22], to *ab initio* conductor-like screening model (COSMSO) and its extension to realistic solvation (COSMO-RS),[23] to bespoke-solvent machine learning (ML) models with random forest regressors.[24] Given the specific importance of aqueous solubility in drug discovery, most effort has focused on predicting aqueous solubility,[25] and relatively fewer works have explored organic solvents, which are particularly crucial in synthetic processes. The aforementioned experimental variability limit of 0.5-1 $\log S$ represents a bound on the performance of any data-driven prediction method, since this variability is irreducible. This variability limit has primarily been explored in the context of aqueous solubility in today's literature. [14, 15, 26] However, there is little reason to believe that the experimental uncertainty should be any lower in organic solvents, and may actually be higher due to increased variability in the experimental methodologies used across laboratories.[27]

State-of-the-art methods focus on applying deep learning to organic solubility prediction, including graph-based neural networks and descriptor-based models [24, 27–30]. Existing models, however, suffer from a lack of generalizability for a variety of reasons. Boobier et al. trained solvent-specific models on only commonly available solvents at room temperature due to a lack of sufficient data to do otherwise, thus rendering the model non-generalizable by construction. Other works like those of Ye and Ouyang and Lee et al. fail to evaluate model performance when extrapolating to new, unseen solutes, a task which mirrors the real task where solubility prediction would be applied in a synthetic pipeline. The state-of-the-art model in literature by Vermeire et al. overcame some of these limitations by training a composition of deep learning models on compiled thermochemical data and using a thermocycle to predict solubility in arbitrary solvents for a wide range of temperatures. In this case, though, data was leaked from training to testing, resulting in overly optimistic performance reports as we demonstrate in this work. Reliance on a collection of machine learning models makes inference times with this model relatively slow, as well.

Here, we combine advances in cheminformatics software and a recently compiled database of organic solubility, BigSolDB, [31] and develop a new state-of-the-art general organic solubility prediction model and validate it under rigorous extrapolation conditions. By adapting the FASTPROP[32] and CHEMPROP [33, 34] architecures to ingest two molecular structures and a temperature, we can train models on BigSolDB to regress logS directly. Our optimized models yield a factor of three improvement over the existing state-of-the-art organic solubility prediction models, with rapid inference times suitable for use in high-throughput workflows. We further demonstrate that our optimized FASTPROP and CHEMPROP models, which use fundamentally different molecular representations, both reach the irreducible error, or aleatoric limit, of accuracy with only a small fraction of the total available data. This indicates that further improvements on organic solubility predictions must be achieved through higher quality datasets rather than larger datasets or more expressive models. Our fastest model, termed FASTSOLV, is open source and can be downloaded as a python package, accessed online via fastsolv.mit.edu, and is integrated in ASKCOS (askcos.mit.edu) and Reaction Mechanism Generator (RMG, rmg.mit.edu).[35]

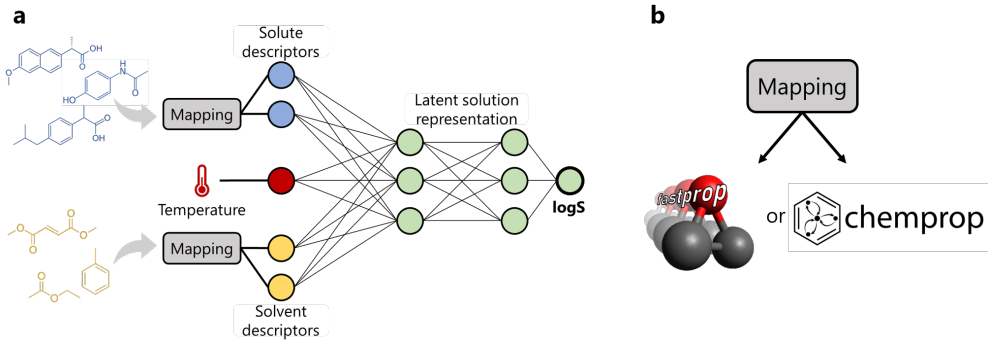

Fig. 1 Machine learning representation of solutions. (a) In our modeling approach, solute and solvent structures are mapped to feature vectors. These feature vectors are concatenated to the solution temperature to arrive at a solution representation, which is passed into a fully-connected neural network and regressed to the $\log S$. (b) Structures are mapped to feature vectors using a fixed representation of Mordred descriptors as implemented in FASTPROP, or a learned representation derived from message passing on a graph representation from CHEMPROP. We compare the performance of models trained on these fundamentally different solution representations.

1 Results

1.1 Datasets and model training approach

In a real discovery context, solubility prediction is usually applied towards a solute extrapolation task, wherein it is desirable to know the solubility of novel candidate compound in a variety of standard solvents for a given temperature. Thus, we stringently trained and evaluated our model performance with this task in mind. However,

attention to the ability of a model to extrapolate to new solutes is not typically heeded in the literature, which makes benchmarking our model performance challenging.

We selected Vermeire et al. as the current literature state-of-the-art model against which to benchmark our results, given that it is widely regarded as a highly performant solvent and temperature-general model.[36, 37] This model contains multiple individual models trained on multiple thermochemical datasets, with the model as a whole being tested on the experimental solubility data compiled in the SolProp dataset. Unfortunately a significant proportion of the testing solutes appeared in the training dataset, demonstrating the need to rigorously evaluate model extrapolation (Figure 1a).

We trained our models on the organic solubility data in BigSolDB, removing solutes that overlap with the SolProp testing set in order to preserve the comparatively small SolProp set. The overlap of solutes in BigSolDB with the Leeds testing set was in contrast removed from the Leeds set, since the Leeds set already contains a diverse variety of solutes. In sum, we ensured our training data had no overlapping solutes with testing sets, providing a stringent test of solute extrapolation (Figure 1b). The distribution of the label $\log S$ across these three datasets are similar, centered around -1 with tails in the limit of solubility (Figure 1c).

We trained our models using 95% of the remaining data in BigSolDB, reserving 5% for validation and model selection. To avoid data leaks, we split our data based on individual solubility experiments, one of which is visualized in Figure 2d, which ensure no solutes appear in both training and validation sets. Since we split our dataset by experiment and solute (Figure 2e), we ensure that we rigorously test extrapolation. The performance of the trained solution FASTPROP model on the training and validation performance is shown in the parity plot in Figure 2f. Performance is quantified via the Root Mean Squared Error (RMSE) and the Percentage of Predictions within 1 logS unit - a metric based on the upper reported limit of experimental reproducibility - referred to as $\%\log S \pm 1$.[24] The model achieves excellent interpolation accuracy, with RMSE = 0.22 and $\%\log S \pm 1 = 99.3\%$. Similarly, the optimized solution CHEMPROP model achieved an RMSE = 0.28 and $\%\log S \pm 1 = 99.2\%$

1.2 Model performance on solute extrapolation

After training solution FASTPROP and CHEMPROP models, their performance on extrapolation was evaluated on the Leeds and SolProp test sets and benchmarked against the Vermiere model as made available via a Python package in the original publication. We observe that the Vermiere model performs poorly on the Leeds dataset, with RMSE = 2.16 and $\%\log S \pm 1 = 41.2\%$ (Figure 3a). On inspection there is observable systematic bias, with the model often severely over-predicting the solubility. In contrast, both the solution FASTPROP and CHEMPROP models perform similarly well, with RMSE = 0.95 and $\%\log S \pm 1 = 73.8\%$ for FASTPROP and RMSE = 0.99 and $\%\log S \pm 1 = 70.9\%$ for CHEMPROP (Figure 3b-c). The systematic bias is greatly reduced in both models.

On the SolProp dataset, the Vermiere model performs slightly better (RMSE = 1.43 and $\%\log S \pm 1 = 66.9\%$), but still exhibits severe systematic bias, with several specific experiments appearing with severaly overpredicted temperatures gradients

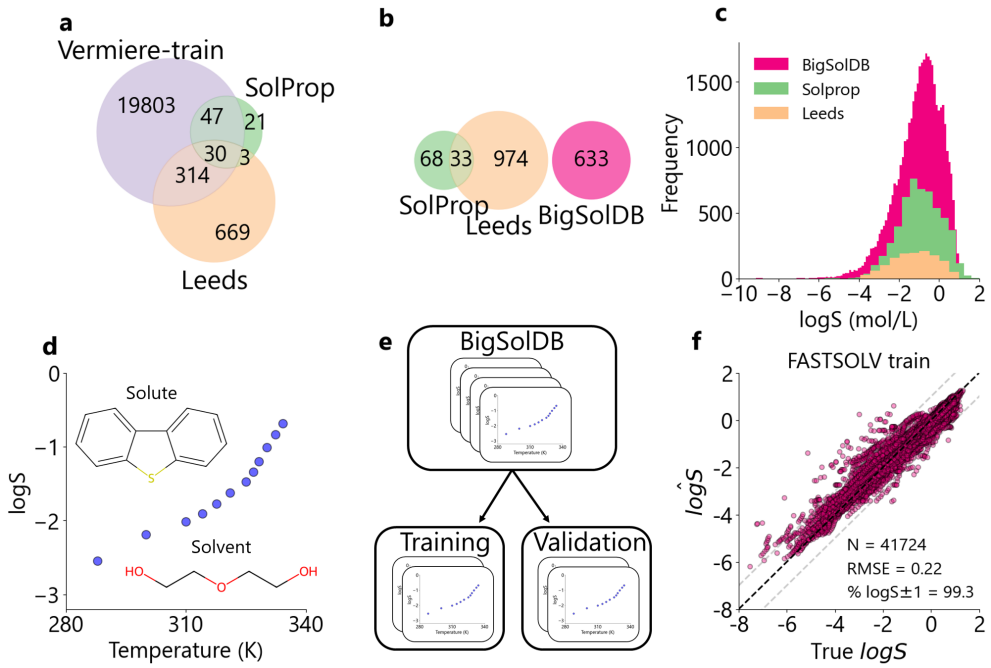

Fig. 2 Datasets and model training approach. (a) The literature best Vermiere et al. trained on a thermocycle with s >20k solutes. The solutes seen during training have overlaps with the SolProp and Leeds testing datasets. (b) In this study, we rigorously test solute extrapolation performance of our models by dropping overlap between the training dataset of BigSolDB, and the SolProp and Leeds testing sets. (c) Distribution of the label, $\log S$ across the training and external testing sets. (d) Demonstration of a single solubility experiment, which contains the measured solubility of a solute (dibenzothiophene) in a solvent (diethylene glycol) across a range of temperatures.[38] (e) Rigorous data splitting strategy splits training and validation data by experiment, ensuring data is not leaked during model selection. (f) Parity plot demonstrating combined training and validation predicitons on BigSolDB of optimized model with FASTPROP mapping. RMSE = 0.22, while $\%\log S \pm 1 = 99.3\%$. See Appendix A for training details.

compared to parity (Figure 3d). In contrast, the solution FASTPROP and CHEMPROP models perform significantly better, with RMSE = 0.83 and $\%\log S \pm 1 = 78.1\%$ for FASTPROP and RMSE = 0.83 and $\%\log S \pm 1 = 76.1\%$ for CHEMPROP (Figure 3e-f). To go beyond these aggregate performance metrics, we also analyzed predictions on specific solutions in the SolProp test set, and observe that our models can correctly rank order solubility in different solvents, and can distinguish solubility in extremely similar solvents when the Vermiere model cannot (Section B).

1.3 Model performance is capped by the aleatoric limit

Given the range of hypothesized experimental limits, ranging from $0.5 < \text{RMSE} < 1.0$, we sought to establish if our models are as performing as accurately as possible given this dataset. To do so, we downsample the training dataset to some smaller size, train

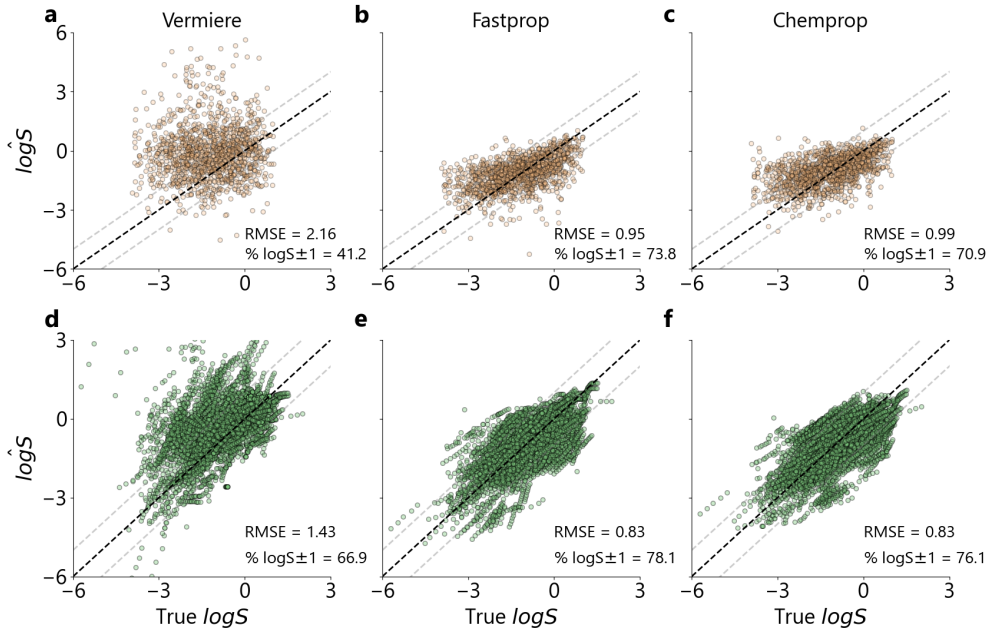

Fig. 3 Performance of literature best Vermiere model and our optimized solution FASTPROP and solution CHEMPROP models on test sets. Parity plots highlighting model predictions of (a) Vermiere (RMSE = 2.16, % log $S\pm1$ = 41.2%), (b) solution FASTPROP (RMSE = 0.95, % log $S\pm1$ = 73.8%), and (c) solution CHEMPROP (RMSE = 0.99, % log $S\pm1$ = 70.9%) on Leeds dataset. Parity plots highlighting model predictions of (d) Vermiere (RMSE = 1.43, % log $S\pm1$ = 66.9%), (e) solution FASTPROP (RMSE = 0.83, % log $S \pm 1$ = 78.1%), and (f) solution CHEMPROP (RMSE = 0.83, % log $S \pm 1$ = 76.1%) on SolProp dataset.

three models and ensemble their performance on the test sets. Repeating this a different sizes of downsampled training sets generates a performance trajectory as a function of training set size (Figure 4a). We observe that the performance trajectories for both the FASTPROP and CHEMPROP models are similar, despite representing fundamentally different modeling approaches. We also observe that the model performance on the SolProp test set plateaus after only 500 experiments (∼5000 data points) are included in training for both the CHEMPROP and FASTPROP models. Similarly, the performance on the Leeds test set plateaus for the CHEMPROP model after only 2000 experiments (∼20,000 data points), although performance of the simpler FASTPROP model takes slightly longer to plateau.

Interestingly, the FASTPROP and CHEMPROP predictions are highly correlated, with a Pearson's $R = 0.86$ (Figure 4b). This correlation is actually stronger than the correlation of either model predictions to the dataset (0.66 for FASTPROP and 0.65 for CHEMPROP). Additionally, comparing the cumulative distribution function (CDF) of predicted gradients of $\log S$ with respect to temperature demonstrates the strong correlation between the FASTPROP and CHEMPROP model predictions (Figure 4c). We obsereve that the Vermiere model has severe systematic error in $\frac{d \log S}{dT}$, achieving an

Earth Mover's Distance (EMD) of 0.06. In contrast, our two models exhibit similar and highly accurate gradient distributions, with EMD of 0.03 and 0.02.

To further examine whether the models are predicting similar results, we plot the cumulative residuals of each model against important features in the SolProp test set. The cumulative residuals against solute molecular weight (Figure 4d) and solvent Wildman-Crippen log P (Figure 4e) both demonstrate that the FASTPROP and CHEMPROP models are making similar predictions across the SolProp test set.

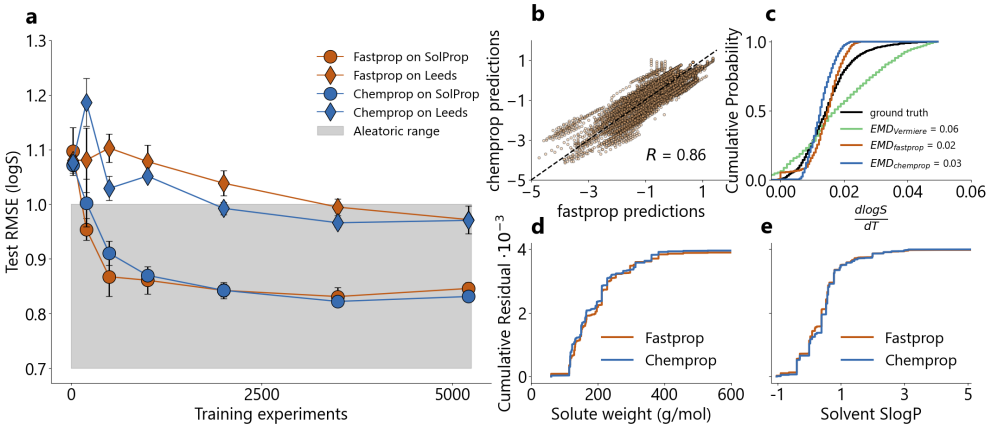

Fig. 4 Model performances reaches aleatoric limit. (a) Test RMSE against the number of experiments used in the training dataset for solution FASTPROP and CHEMPROP on both the SolProp and Leeds datasets. Orange-colored curves show results for solution FASTPROP models, while blue-colored curves show test RMSE for solution CHEMPROP models. For both colors, diamond markers indicate test RMSE on the Leeds dataset, while circular markers indicate test RMSE on the SolProp dataset. The shaded area indicates the range of aleatoric limit reported in literature ($0.5 < \log S < 1.0$). The plot cuts off at 0.7 for ease in visualizing the data. Error bars indicate standard deviation across three randomized trainign trials. (b) Correlation of FASTPROP and CHEMPROP model predictions on the SolProp test set. Pearson's $R = 0.86$. Black dotted line shows parity. (c) CDF of the gradients of $\log S$ with respect to temperature T in the SolProp test set. Vermeire predicted gradient CDF (green) achieves an earth movers distance (EMD) of 0.06 compared to the SolProp ground truth gradient CDF (black). FASTPROP predicted gradient CDF (orange, EMD = 0.03) and CHEMPROP predicted gradient CDF (blue, EMD = 0.02) compared to the SolProp ground truth gradient CDF (black). (d-e) Cumulative residual of FASTPROP (orange) and CHEMPROP (blue) model predictions of $\log S$ in the SolProp test set against (d) solute molecular weight (g/mol) and (e) solvent Wildman-Crippen $\log P$. Cumulative residual is multiplied by 10^{-3} for concise axis labels.

2 Discussion

Here, we leveraged state-of-the-art cheminformatics software and large compiled solubility datasets to develop accurate and generalizable solubility models. In stringent extrapolation, we observe a 2-3 fold decrease in RMSE over the literature best model from Vermeire et al.. In addition, due to the relative simplicity of the architecture and the more modern code, inference times with this approach are up to fifty times

faster. To the best of our knowledge, our models are the best performing solvent- and temperature general models in the literature on solute extrapolation.

The methodological errors in Vermeire et al. model training lead to overly optimistic performance results reported in this study as shown in Figure 3. As shown in the solute Venn diagrams in Figure 2, many entries in the SolProp testing set are present in the training set. In evaluating Vermeire et al.'s performance on the rigorous solute extrapolation test posed by the Leeds set, the performance drops to RMSE = 2.16, demonstrating a more realistic test performance. Additionally, we observe non-physical gradients with respect to temperature, indicating the model has not learned a functional approximation of the temperature dependence of $\log S$.

In contrast, our optimized solution FASTPROP and solution CHEMPROP models exhibit much more consistent performance between the Leeds and SolProp test sets, highlighting the strong performance of our models under rigorous solute extrapolation (Figure 2). The slightly decreased accuracy of our models on the Leeds test set is perhaps attributable to the increased solute diversity in the Leeds dataset, which is to be expected given the difference in its construction. Both models also exhibit accurate and physically-realistic predictions of the gradient on the solubility with respect to temperature, indicating the models learned physically-realistic temperature dependence - critically important in specific process chemistry applications.[39, 40]

The training data size study and the error analysis presented in Figure 3 shows that our optimized models, which rely on distinct molecular mappings, both converge to the same performance limits with similar distributions of predictions and errors. The highly correlated predictions and cumulative residuals show both models achieve not only the same average performance, but predict similarly across the tests sets. CHEMPROP should be able to continuously learn a better representation as the size of the training set increases, as demonstrated by Heid et al. on several molecular properties.[41] However, our optimized solution CHEMPROP model performance stops improving after a relatively small amount of training data, indicating that variability in the training data limits the model's ability to learn an improved representation and predict more accurately.

In the context of aqueous solubility, Palmer and Mitchell compiled a curated set of highly accurate experimental measurements, and concluded experimental uncertainty was not limiting model performance, but rather QSPR methods themselves were limiting. Since then, innovations in cheminformatics software has led to highly accurate and expressive model architectures which have been shown to perform excellently on molecular property prediction tasks.[33, 42] The inability of these models to improve with increasing training set size on organic solubility predictions challenges this paradigm. This instead suggests that experimental uncertainty that limits model performance - the aleatoric limit has been reached - and that a new approach is needed to further improve models.

Currently, there are substantial efforts on compiling larger databases of published experimental datasets from the literature. However, as we demonstrated, only a small subset, $\sim$ 5000 points, of our training set are needed to achieve near-optimal performance. Compiling larger databases will not continue to improve model performance beyond the aleatoric limits demonstrated here. Instead, accurate datasets of

organic solubility are needed, analogous to the CheqSol dataset compiled by Palmer and Mitchell. While it is possible that further innovations in model architecture can improve predictions, the irreducible error imposed by the experimental uncertainty must be reduced in order to impart a noticeable improvement in organic solid solubility predictions.

In conclusion, we present organic solubility prediction models using deep learning on fixed and learned molecular representations, then test them under rigorous solute extrapolation. Our models outperform comparable literature models by a factor of 2-3 on publicly-available test sets. We demonstrate that our models are predicting near the aleatoric limit of the experimental training data, motivating the assembly of highly accurate datasets for the field to notice further improvements in organic solubility prediction. Given the importance of solid solubility prediction, we have termed the fixed representation model FASTSOLV and taken extensive steps to make it available. FASTSOLV can be accessed via any web browser for free at fastsolv.mit.edu, downloaded as a python package for use in scripting, and is integrated directly within the ASKCOS (askcos.mit.edu) and RMG (rmg.mit.edu) platforms for retrosynthesis and reaction mechanism generation, respectively.

3 Methods

3.1 Training Procedure

Rigorous evaluation of extrapolation requires careful preparation of the data for training, validation, and testing. Overlapping solutes structures in the testing and training data are dropped from the training data to avoid data leaks. The ASTARTES software package [43] is used to randomly partition entire experiments into these sets, again ensuring that no solute structures are seen by the model in multiple sets. A single 'model' contains four individual networks trained on a different random training set to account of the affect of random sampling, and a prediction is then the average across these four models.

3.1.1 Network Architecture

Prior to arriving at the simple architecture shown in Figure 1, more complex models inspired by the work of Pathak et al. were tested. Additional separate latent layers were added to solute and solvent which allowed the network to learn a solute- and solvent-specific representation before performing a configurable 'interaction' operation such as concatenation or element-wise multiplication. The same optimization framework mentioned above was used to choose the best network with these options, and it consistently disabled these additional complexities. Simply relying on Universal Approximation theorem is sufficient - additional inductive bias does not improve performance. See B for further details.

3.2 Aleatoric Error Study

The 'performance trajectory' shown in Figure 4 is generated by gradually increasing the amount of training data available to the model. The model is trained in the same

manner as described previously, actually containing four separate networks trained on different random selections of the downsampled data. At first the model sees only a small number of experiments during training before subsequent testing on the holdout sets. The amount is gradually increased to the full size of the dataset, analogous to performing more solubility experiments to gather more samples in hopes of improving model performance.

Acknowledgments

The authors thank Connor W. Coley for his helpful comments and suggestions during the formulation of this concept. The authors thank Florence Vermiere for helpful feedback and suggestions during the writing of this manuscript. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this article.

- Funding: This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Numbers DE-SC0023112 (J. B.) and DE-SC0022158 (L. A.)
- Conflict of interest: The authors report no conflicts of interest.
- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable
- Data availability: We present no original data, but the source code contains step-by-step instructions to retrieve and munge the data referenced in the body.
- Materials availability: Not applicable
- Code availability: The source code for model training, testing, and analysis is available on GitHub (https://github.com/JacksonBurns/fastsolv). FASTSOLV is also packaged through PyPI and installable in Python via pip (https://pypi.org/project/fastsolv/). Model checkpoints are deposited on Zenodo (https://zenodo.org/records/13943074). FASTSOLV is also directly accesisble via a web interface (http://fastsolv.mit.edu/).
- Author contribution: Lucas Attia: conceptualization (equal); methodology (equal); formal analysis (equal); software (supporting); writing – original draft (lead); review and editing (equal). Jackson Burns: conceptualization (equal); methodology (equal); formal analysis (equal); software (lead); writing – original draft (supporting); review and editing (equal). Patrick S. Doyle: supervision (supporting); formal analysis (supporting); review and editing (supporting). William H. Green: supervision (lead); formal analysis (supporting); review and editing (supporting).

Appendix A Training Details

The PyTorch [45] library is used to implement networks via the PyTorch Lightning [46] framework, which delivers excellent reproducibility and reusability. Network hyperparameters are optimized automatically using the Optuna software package [47]. Network hyperparameter optimization and training took place on the MIT SuperCloud High

Performance Computing cluster [48] GPU nodes containing 2 x Nvidia Volta V100s. All reported metrics are defined according to their usual formulae and are implemented via standard Python machine learning packages. Early stopping was used during training, which allows the network to continue training until the error on the validation set increased, indicating overfitting. At that point, the network is reverted to the previous weights prior to the increase and carried forward for testing.

Extensive efforts were made to identify a physics-informed neural network architecture which would infuse inductive bias into the network and improve predictions. Each additional facet was enabled during the automated hyperparameter optimization such that the algorithm would automatically deduce which was the most effective. To allow the network to learn a unique per-solvent and per-solute representation rather than a single 'solution' representation, distinct linear layers were added after the initial Mapping. This reflects the intuitive understanding that each solute and solvent should have a unique contribution to the resulting solubility which is independent of the exact solution. The manner in which the latent solute and solvent representations was also configurable - in analogy to existing solubility prediction models, the network could choose to perform element-wise multiplication, subtraction, or addition instead of simple concatenation. Across many repeated hyperparameter optimization instances none of the physics-infused models were able to outperform the simple architecture presented in the main text. Inductive bias was unable to surpass reliance on universal approximator theorem, at least given the current aleatoric limit of available data.

Sobolev training [49] was implemented for both the Chemprop- and FASTPROP-based FASTSOLV models. This approach penalizes the network during training for both the error in the prediction of the solubility and the gradient of the predicted solubility with respect to the input temperature. The latter is approximated from the input data using finite differences, a reasonable approximation for the typically monotonic and locally-linear solubility curves. During training the gradient is found by continuing backpropagation through all network layers, as is usually done, and additionally the input temperature. The effect of Sobolev training is that networks generally converge in fewer epochs and are stronger interpolators. The latter is of specific interest in some process applications of FASTSOLV.

All code implementing the above is open source, permissively licensed, and available online at github.com/JacksonBurns/fastsolv.

Appendix B Predictions on Specific Solutions

We further validated the performance of our models by investigating the predictions on specific solutions. We selected two structurally distinct solutes from the held out SolProp test set, risperidone and L-prolinamide, and compared model predictions from FASTSOLV and Vermiere against the experimental solubility data as a function of temperature for different solvents. Risperidone is a water-insoluble benzisoxazole derivative and antipyschotic, [50] while L-prolinamide is an amino acid amide used in peptide synthesis and as an organic catalyst.[51]

For risperidone, we observe that the FASTSOLV predictions in acetone and isopropanol are highly accurate, with the model correctly predicting the relative order

(higher solubility in acetone), and accurately predicting the gradient of solubility with respect to temperature (Figure B1a). In contrast, the Vermiere model overpredicts the solubility in both acetone and isopropanol, exhibits much greater model uncertainty, and overpredicts the slope of the curve with respect to temperature. For L-prolinamide, we impose a challenging task for the model by predicting solubility in extremely similar solvents, hexane and heptane. We observe that FASTSOLV is able to discriminate solubility between these similar solvents, correctly predicting higher solubility in heptane, while the Vermiere model predicts identical soubility in both solvents (Figure B1b). Again, we also observe accurate absolute predictions, gradients, and lower model uncertainty.

In both cases, we observe FASTSOLV correctly predicts the rank ordering of solubility in the solvents, which is an important task for high-throughput virtual screening workflows. Additionally, the ability to discriminate between highly similar solvents is a difficult task for ML models, and the ability of FASTSOLV to do so indicates it is accurately representing solvent molecules. Overall, we these results further validate model performance under solute extrapolation beyond the aggregate performance metrics reported in Figure 3.

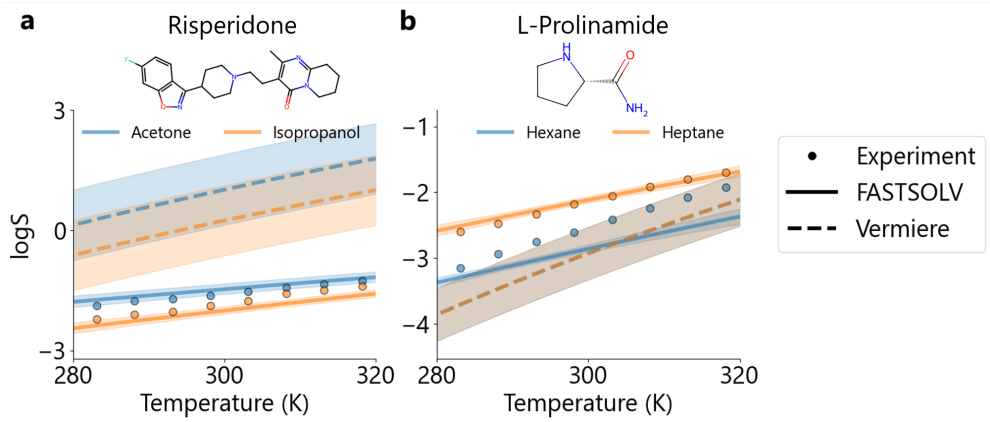

Fig. B1 Performance on example solutions. Solubility predictions and experimental data as a function of temperature for solutions of (a) risperidone in acetone (blue) and isopropanol (orange) and (b) L-prolinamide in hexane (blue) and heptane (orange). Experimental data are plotted as circles, predictions from the Vermiere model are plotted as a dotted line, and FASTSOLV predictions are plotted as solid lines. The shaded areas indicate model uncertainty. Experimental risperidone solubility data is compiled in SolProp from Mealey et al.. Experimental L-prolinamide solubility data is compiled in SolProp from Cui et al.

References

[1] Blakemore, D.C., Castro, L., Churcher, I., Rees, D.C., Thomas, A.W., Wilson, D.M., Wood, A.: Organic synthesis provides opportunities to transform drug discovery. Nature chemistry **10**(4), 383–394 (2018)

[2] Simon, M.-O., Li, C.-J.: Green chemistry oriented organic synthesis in water. Chemical Society Reviews **41**(4), 1415–1427 (2012)

[3] Chanda, A., Fokin, V.V.: Organic synthesis "on water". Chemical reviews **109**(2), 725–748 (2009)

[4] Coquerel, G.: Crystallization of molecular systems from solution: phase diagrams, supersaturation and other basic concepts. Chemical Society Reviews **43**(7), 2286–2300 (2014)

[5] Jenekhe, S.A., Chen, X.L.: Self-assembled aggregates of rod-coil block copolymers and their solubilization and encapsulation of fullerenes. Science **279**(5358), 1903–1907 (1998)

[6] Alhazmi, B., Ignacz, G., Di Vincenzo, M., Hedhili, M.N., Szekely, G., Nunes, S.P.: Ultraselective macrocycle membranes for pharmaceutical ingredients separation in organic solvents. Nature Communications **15**(1), 7151 (2024)

[7] Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews **23**(1-3), 3–25 (1997)

[8] Rabinow, B.E.: Nanosuspensions in drug delivery. Nature reviews Drug discovery **3**(9), 785–796 (2004)

[9] Evich, M.G., Davis, M.J., McCord, J.P., Acrey, B., Awkerman, J.A., Knappe, D.R., Lindstrom, A.B., Speth, T.F., Tebes-Stevens, C., Strynar, M.J., *et al.*: Per-and polyfluoroalkyl substances in the environment. Science **375**(6580), 9065 (2022)

[10] Monteith, D.T., Henrys, P.A., Hruška, J., Wit, H.A., Krám, P., Moldan, F., Posch, M., Räike, A., Stoddard, J.L., Shilland, E.M., *et al.*: Long-term rise in riverine dissolved organic carbon concentration is predicted by electrolyte solubility theory. Science Advances **9**(3), 3491 (2023)

[11] Murdande, S.B., Pikal, M.J., Shanker, R.M., Bogner, R.H.: Aqueous solubility of crystalline and amorphous drugs: challenges in measurement. Pharmaceutical development and technology **16**(3), 187–200 (2011)

[12] Ali, J., Camilleri, P., Brown, M.B., Hutt, A.J., Kirton, S.B.: Revisiting the general solubility equation: in silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. Journal of chemical information and

modeling **52**(2), 420–428 (2012)

[13] Hughes, L.D., Palmer, D.S., Nigsch, F., Mitchell, J.B.: Why are some properties more difficult to predict than others? a study of qspr models of solubility, melting point, and log p. Journal of chemical information and modeling **48**(1), 220–232 (2008)

[14] Palmer, D.S., Mitchell, J.B.: Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? Molecular Pharmaceutics **11**(8), 2962–2972 (2014)

[15] Llinas, A., Oprisiu, I., Avdeef, A.: Findings of the second challenge to predict aqueous solubility. Journal of chemical information and modeling **60**(10), 4791–4803 (2020)

[16] Katritzky, A.R., Wang, Y., Sild, S., Tamm, T., Karelson, M.: Qspr studies on vapor pressure, aqueous solubility, and the prediction of water- air partition coefficients. Journal of Chemical Information and Computer Sciences **38**(4), 720–725 (1998)

[17] Jorgensen, W.L., Duffy, E.M.: Prediction of drug solubility from structure. Advanced drug delivery reviews **54**(3), 355–366 (2002)

[18] Tzschucke, C.C., Markert, C., Bannwarth, W., Roller, S., Hebel, A., Haag, R.: Modern separation techniques for the efficient workup in organic synthesis. Angewandte Chemie International Edition **41**(21), 3964–4000 (2002)

[19] Barrett, J.A., Yang, W., Skolnik, S.M., Belliveau, L.M., Patros, K.M.: Discovery solubility measurement and assessment of small molecules with drug development in mind. Drug discovery today **27**(5), 1315–1325 (2022)

[20] Leuner, C., Dressman, J.: Improving drug solubility for oral delivery using solid dispersions. European journal of Pharmaceutics and Biopharmaceutics **50**(1), 47–60 (2000)

[21] Abraham, M.H.: Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. Chemical Society Reviews **22**(2), 73–83 (1993)

[22] Acree Jr, W.E., Abraham, M.H.: Solubility predictions for crystalline polycyclic aromatic hydrocarbons (pahs) dissolved in organic solvents based upon the abraham general solvation model. Fluid phase equilibria **201**(2), 245–258 (2002)

[23] Klamt, A.: The cosmo and cosmo-rs solvation models. Wiley Interdisciplinary Reviews: Computational Molecular Science **1**(5), 699–709 (2011)

[24] Boobier, S., Hose, D.R., Blacker, A.J., Nguyen, B.N.: Machine learning with

physicochemical relationships: solubility prediction in organic solvents and water. Nature communications **11**(1), 5753 (2020)

[25] Delaney, J.S.: Predicting aqueous solubility from structure. Drug discovery today **10**(4), 289–295 (2005)

[26] Llinàs, A., Glen, R.C., Goodman, J.M.: Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? Journal of chemical information and modeling **48**(7), 1289–1303 (2008)

[27] 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 **2**(2), 356–367 (2023)

[28] 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 **7**(14), 12268–12277 (2022)

[29] Ye, Z., Ouyang, D.: Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. Journal of cheminformatics **13**(1), 98 (2021)

[30] Vermeire, F.H., Chung, Y., Green, W.H.: Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. Journal of the American Chemical Society **144**(24), 10785–10797 (2022) https://doi.org/10.1021/jacs.2c01768

[31] Krasnov, L., Mikhaylov, S., Fedorov, M., Sosnin, S.: Bigsoldb: Solubility dataset of compounds in organic solvents and water in a wide range of temperatures. ChemRxiv (2023) https://doi.org/10.26434/chemrxiv-2023-qqslt

[32] Burns, J., Green, W.: Generalizable, fast, and accurate deepqspr with fastprop part 1: Framework and benchmarks. arXiv preprint arXiv:2404.02058 (2024)

[33] 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. Journal of Chemical Information and Modeling **59**(8), 3370–3388 (2019) https://doi.org/10.1021/acs.jcim.9b00237 https://doi.org/10.1021/acs.jcim.9b00237. PMID: 31361484

[34] Heid, E., Greenman, K.P., Chung, Y., Li, S.-C., Graff, D.E., Vermeire, F.H., Wu, H., Green, W.H., McGill, C.J.: Chemprop: A machine learning package for chemical property prediction. Journal of Chemical Information and Modeling **64**(1), 9–17 (2024) https://doi.org/10.1021/acs.jcim.3c01250 https://doi.org/10.1021/acs.jcim.3c01250. PMID: 38147829

[35] Liu, M., Grinberg Dana, A., Johnson, M.S., Goldman, M.J., Jocher, A., Payne, A.M., Grambow, C.A., Han, K., Yee, N.W., Mazeau, E.J., *et al.*: Reaction mechanism generator v3. 0: advances in automatic mechanism generation. Journal of Chemical Information and Modeling **61**(6), 2686–2696 (2021)

[36] Tuttle, M.R., Brackman, E.M., Sorourifar, F., Paulson, J., Zhang, S.: Predicting the solubility of organic energy storage materials based on functional group identity and substitution pattern. The Journal of Physical Chemistry Letters **14**(5), 1318–1325 (2023)

[37] Kim, Y., Jung, H., Kumar, S., Paton, R.S., Kim, S.: Designing solvent systems using self-evolving solubility databases and graph neural networks. Chemical Science **15**(3), 923–939 (2024)

[38] Tao, B., Li, X., Yan, M., Luo, W.: Solubility of dibenzothiophene in nine organic solvents: Experimental measurement and thermodynamic modelling. The Journal of Chemical Thermodynamics **129**, 73–82 (2019)

[39] Avdeef, A.: Solubility temperature dependence predicted from 2d structure. ADMET and DMPK **3**(4.), 298–344 (2015)

[40] Khoshraftar, Z., Ghaemi, A.: Prediction of co2 solubility in water at high pressure and temperature via deep learning and response surface methodology. Case Studies in Chemical and Environmental Engineering **7**, 100338 (2023)

[41] Heid, E., McGill, C.J., Vermeire, F.H., Green, W.H.: Characterizing uncertainty in machine learning for chemistry. Journal of Chemical Information and Modeling **63**(13), 4012–4029 (2023)

[42] Burns, J., Green, W.: Generalizable, fast, and accurate deepqspr with fastprop part 1: Framework and benchmarks. arXiv preprint arXiv:2404.02058 (2024) https://doi.org/10.48550/arXiv.2404.02058

[43] Burns, J.W., Spiekermann, K.A., Bhattacharjee, H., Vlachos, D.G., Green, W.H.: Machine Learning Validation via Rational Dataset Sampling with astartes. Journal of Open Source Software **8**(91), 5996 (2023) https://doi.org/10.21105/joss.05996

[44] Pathak, Y., Mehta, S., Priyakumar, U.D.: Learning atomic interactions through solvation free energy prediction using graph neural networks. Journal of Chemical Information and Modeling **61**(2), 689–698 (2021) https://doi.org/10.1021/acs.jcim.0c01413

[45] Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y.,

Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM, ??? (2024). https://doi.org/10.1145/3620665.3640366 . https://pytorch.org/assets/pytorch2-2.pdf

[46] Falcon, W., The PyTorch Lightning team: PyTorch Lightning. https://doi.org/10.5281/zenodo.3828935 . https://github.com/Lightning-AI/lightning

[47] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)

[48] Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., Jones, M., Klein, A., Milechin, L., Mullen, J., Prout, A., Rosa, A., Yee, C., Michaleas, P.: Interactive supercomputing on 40,000 cores for machine learning and data analysis. In: 2018 IEEE High Performance Extreme Computing Conference (HPEC), pp. 1–6 (2018). IEEE

[49] Czarnecki, W.M., Osindero, S., Jaderberg, M., Świrszcz, G., Pascanu, R.: Sobolev training for neural networks (2017) https://doi.org/10.48550/ARXIV.1706.04859

[50] Cohen, L.J.: Risperidone. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy **14**(3), 253–265 (1994)

[51] Xu, J., Fu, X., Wu, C., Hu, X.: Simple, inexpensive, and facile l-prolinamide used as a recyclable organocatalyst for highly efficient large-scale asymmetric direct aldol reactions. Tetrahedron: Asymmetry **22**(8), 840–850 (2011)

[52] Mealey, D., Svärd, M., Rasmuson, Å.C.: Thermodynamics of risperidone and solubility in pure organic solvents. Fluid Phase Equilibria **375**, 73–79 (2014)

[53] Cui, Z., Ye, J., Yao, L., Wang, Z., Wang, T., Hu, Y.: Determination and analysis of solubility of l-prolinamide in ten pure solvents and three binary solvent mixtures at different temperatures (t= 278.15–323.15 k). Journal of Chemical & Engineering Data **66**(2), 1172–1184 (2021)