Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy

Yunsie Chung,[†] Florence H. Vermeire,[†] Haoyang Wu,[†] Pierre J. Walker,[†,‡] Michael H. Abraham,[¶] and William H. Green[*,†]

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

[‡]*Department of Chemical Engineering, Imperial College London, SW7 2AZ, United Kingdom*

[¶]*Department of Chemistry, University College London, 20 Gordon Street, London WC1H OAJ, United Kingdom*

E-mail: whgreen@mit.edu

Abstract

We present a group contribution method (SoluteGC) and a machine learning model (SoluteML) to predict the Abraham solute parameters, as well as a machine learning model (DirectML) to predict solvation free energy and enthalpy at 298 K. The proposed group contribution method uses atom-centered functional groups with corrections for ring and polycyclic strain whilst the machine learning models adopt a directed message passing neural network. The solute parameters predicted from SoluteGC and SoluteML are used to calculate solvation energy and enthalpy via linear free energy relationships. Extensive data sets containing 8366 solute parameters, 20253 solvation free energies, and 6322 solvation enthalpies are compiled in this work to train the models. The three

models are each evaluated on the same test sets using both random and substructure-based solute splits for solvation energy and enthalpy predictions. The results show that the DirectML model is superior to the SoluteML and SoluteGC models for both predictions and can provide accuracy comparable to that of advanced quantum chemistry methods. Yet, even though the DirectML model performs better in general, all three models are useful for various purposes. Uncertain predicted values can be identified by comparing the 3 models, and when the 3 models are combined together, they can provide even more accurate predictions than any one of them individually. Finally, we present our compiled solute parameter, solvation energy, and solvation enthalpy databases (SoluteDB, dGsolvDBx, dHsolvDB) and provide public access to our final prediction models through a simple web-based tool, software package, and source code.

1 Introduction

Information on solvation free energy aids in the selection of viable solvents in chemical processes such as the synthesis of organic molecules,[1,2] optimization of purification processes,[3] and pollutant level management.[4] The solvation Gibbs free energy (ΔG_{solv}) of a solute in a solvent is directly related to that solute's partition coefficient between the gas and solvent phase. This property is typically reported at room temperature and can be a valuable feature for the prediction of the solute's liquid-liquid partition coefficient and solid solubility in organic solvents. For process optimization, ΔG_{solv} is required at the specified process temperature and in a variety of solvents. Recently, we reported a strategy to calculate ΔG_{solv} of a dilute neutral solute in organic solvents at different temperatures.[5] Using only the solvation free energy and solvation enthalpy at 298 K and solvent's temperature-dependent density, ΔG_{solv} for temperatures between 298 K and the solvent's critical temperature can be calculated along the solvent's saturation curve in a fast and automated manner. For this work, we aim to provide improved predictions of $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$, which can be used for the calculation of ΔG_{solv} at elevated temperatures, an easy-access tool for our predictive models, and new databases for $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$.

The interest in solvation free energies dates back many years[6] and has led to the development of numerous predictive methods. These range from molecular dynamics and quantum chemistry methods to empirical or data-driven approaches. Quantum chemistry methods distinguish themselves between explicit and implicit solvent representations. Commonly used quantum chemistry methods are based on the implicit polarizable continuum model for solvent representation, such as the SMx methods developed by Cramer, Truhlar, and coworkers[7–9] and the COSMO(-RS) models proposed by Klamt and coworkers.[10–12] These first-principle methods are useful for the calculation of solvation properties of new solvent-solute combinations, but they are computationally expensive and require comprehensive and challenging searches for all relevant conformers of the considered solvent and solute molecules. Empirical or data-driven approaches, on the other hand, allow for the fast prediction of solvation properties. The main bottleneck for these methods is the scarcity and quality of

available experimental data. As more data become available, more studies focus on the application of empirical and data-driven models to the prediction of solvation-related properties such as ΔG_{solv}. In this study, we focus on improving predictions for $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ with several empirical and data-driven models, more specifically using the commonly employed linear solvation energy relationship (LSER), a group contribution method, and state-of-the-art graph-convolutional message passing neural networks.

The LSER equation used in this work is the one developed by Abraham et al.[13,14] The equation is built based on earlier attempts to relate the free energy of solvation to molecular descriptors by quantitative structure-property relationships (QSPR), for example the solvatochromic comparison method by Kamlet and Taft.[15] The relationship developed by Abraham and coworkers is given in Eq. 1. It combines Abraham solute (E, S, A, B, L) and solvent (c, e, s, a, b, l) parameters through a linear equation for the determination of the gas-liquid partition coefficient K. The Gibbs free energy of solvation can be directly calculated from the partition coefficient through Eq. 2.

$$\log_{10} K(298\,\mathrm{K}) = c + eE + sS + aA + bB + lL \tag{1}$$

$$\Delta G_{\mathrm{solv}} = -RT \ln K \tag{2}$$

The LSER is further explored by Mintz et al.[16] for the prediction of $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ as shown in Eq. 3. The solute parameters in Eq. 3 are the same Abraham solute parameters as those used in the Abraham LSER in Eq. 1, while new solvent parameters (c', e', s', a', b', l') are used.

$$\frac{\Delta H_{\mathrm{solv}}(298\,\mathrm{K})}{1\,\mathrm{kJ\,mol^{-1}}} = c' + e'E + s'S + a'A + b'B + l'L \tag{3}$$

Each of the solute parameters used in the LSER is related to the physical property of a solute and can be either determined experimentally or regressed from experimental values of partition coefficients. E is the solute excess molar refractivity, S is the solute dipolarity/polarizability, A and B are the overall hydrogen bond acidicity and basicity, respectively, and L is the logarithm of the gas-hexadecane partition coefficient.[14,17] The solute parameters can be also used to estimate liquid-liquid partition coefficients, solid solubilities, heat capacities, enthalpies of sublimation, and liquid phase hydrogen abstraction reaction rates.[18–21] The solvent parameters, on the other hand, are largely treated as empirical parameters and obtained from fitting to experimental data.

Several methods have been proposed to estimate some or all of the solute parameters from molecular structure,[22–30] among which the group contribution (GC) approach is widely used. Platts et al.[24,25] were the first to devise the group contribution scheme that can predict all solute parameters. They developed 81 functional group fragments for predicting E, S, B, and L and 51 fragments for predicting A. The ACD/Absolv software[31] adopted the Platts-type fragments and further optimized the fragments for their module. ACD/Absolv has shown to give reasonably good estimates for many compounds, but it had large prediction errors for

certain classes of molecular structure, such as highly halogenated compounds, triazoles, and bridged ring structures.[32] Brown et al. developed a different set of fragments or substructures using the iterative fragment selection approach, in which the fragments are selected by using k-fold cross validation.[33,34] They used solute parameter data of around 3700 compounds and built an open-access GC model that is available through the UFZ-LSER database.[29] Their model showed good predictive performance for the L parameter,[34] but the performance on the other solute parameters has not been reported.

As more comprehensive databases become available, researchers have started exploring the use of machine learning (ML) for the prediction of $\Delta G_{\text{solv}}(298\,\text{K})$. Initially, these efforts focused on hydration free energies using the FreeSolv database.[35] This database is often used as a benchmark to compare model architectures in chemical property prediction.[36,37] The Minnesota solvation database (MNSol)[38] contains solvation free energies for a wide range of solvents. Hutchinson and Kobayashi[39] were the first to use this database to account for different solvents in a neural network by using the DeepChem framework[40] with functional class fingerprints that have solvent-specific features. More recent works have been published based on a larger database Solv@TUM.[41] Pathak et al.[42] proposed a chemically interpretable graph interaction network model comprised of a message passing, an interaction, and a prediction phase using 6239 unique solvent-solute combinations extracted from Solv@TUM and FreeSolv data sets. Lim and Jung[43] used the same data sets and developed MLSolvA, a ML architecture that computes pairwise atomic interactions from the solvent and solute atomistic feature vectors and makes prediction by summing up the interactions. As part of our previous work, Vermeire and Green[44] presented a transfer learning approach in which the model was pre-trained on 1 million quantum calculations and further fine-tuned on 10145 solvation free energy experimental data. The main purpose of that work was to demonstrate how transfer learning from quantum chemical data improves performance on small dataset sizes and on out-of-range sample predictions. Overall, the model achieves a mean absolute error of 0.21 kcal/mol on a random test split.

In this work, we adopt three different approaches to predict $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$, two of which predict Abraham solute parameters and then calculate solvation properties using the LSERs. The third approach obtains $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$ directly for the specified solvent-solute pair. For the prediction of $\Delta G_{\text{solv}}(298\,\text{K})$ we chose not to make use of our previously published transfer learning model, since the purpose of this work is to extensively compare different approaches to experimental data-driven methods. Extensive databases of the solute parameters, $\Delta G_{\text{solv}}(298\,\text{K})$, and $\Delta H_{\text{solv}}(298\,\text{K})$ have been compiled for this work and all models are trained on experimental data. The performance of the three approaches is assessed for solvent-solute combinations that are considered out-of-sample with respect to the training data. We specifically use the random and substructure-based solute splits such that none of the solutes or selected solute substructures in the test sets appear in the training sets. One of the novelties of this work is that the performance of all the methods is evaluated on exactly the same test sets for different splits. For the final prediction of solvation properties, we advise to combine all three methods as they provide a different level

of accuracy and interpretation of the contribution of several physical phenomenon to the calculated values. We also provide an easy-accessible tool for our predictive models that can be used to calculate the solute parameters, $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$. Lastly, we provide new databases for the solute parameters, $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ compiled and curated from different sources.

2 Methods – Databases

Three different approaches are used to predict $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$: (1) a GC method for the solute parameters in the Abraham equation **(SoluteGC)**, (2) ML for the prediction of the same solute parameters **(SoluteML)**, and (3) ML for the direct prediction of solvation properties **(DirectML)**. The advantage of using these three different methods for the prediction of solvation properties is that they each provide a different level of information on the physical contributions to the calculated values. Whereas the DirectML method is a black-box, the SoluteML model provides solute parameters with physical meaning, and the SoluteGC model relates those parameters to chemical substructures.

The three approaches start from two different experimental data sets: one with solute molecules and their solute parameters **(SoluteDB)**, and one with solvent-solute pairs and values for $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ **(dGsolvDB*x*, and dHsolvDB)**. The SoluteGC and SoluteML methods predict the solute parameters from SMILES[45] of the solute compounds. The solvation free energy and enthalpy are subsequently computed from the predicted solute parameters and empirically fitted solvent parameters through the LSERs (Eqs. 1 and 3). The DirectML method takes SMILES of the solvents and solutes as input and predicts the solvation free energy and enthalpy. Figure 1 depicts the prediction flowchart of the three models starting from the two data sets. The details on the individual models and information on the construction and splitting of the two data sets are given below.

2.1 Data Collection and Compilation

The experimental data for the solute parameters (E, S, A, B, L), $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ are collected from various sources. All data except those collected from the Minnesota solvation database are open-source and published as a part of this work. The data are limited to neutral solute compounds containing H, C, N, O, S, P, F, Cl, Br, or I atoms and nonionic liquid solvents in this work. The standard state of 1 $\mathrm{mol\,L^{-1}}$ gas phase and 1 $\mathrm{mol\,L^{-1}}$ liquid phase is used for ΔG_{solv} and ΔH_{solv}.

The solute and solvent information in the collected databases are given in various representations, such as CAS numbers, InChI,[46] SMILES, chemical names, and 3D coordinates. The data are standardized by converting the given identifiers to both SMILES and InChI strings

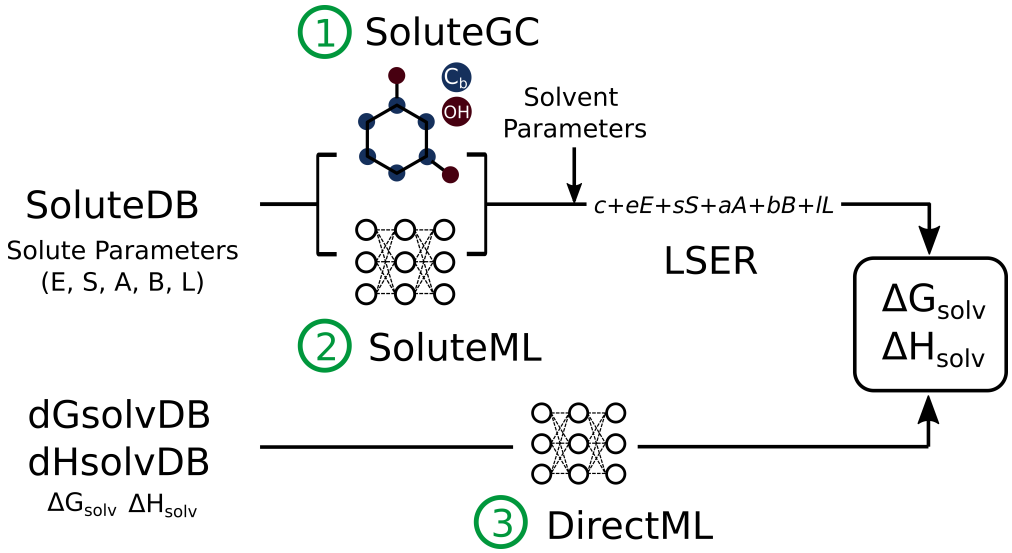

Figure 1: Schematic overview of $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ prediction with the three methods (SoluteGC, SoluteML, and DirectML) starting from the three data sets (SoluteDB, dGsolvDB, and dHsolvDB)

using PubChemPy,[47] CIRPy,[48] RDKit,[49] and ARC.[50] All available chemical identifiers are converted and the resulting InChI strings are compared to eliminate incorrect or ambiguous naming. The compounds with specified stereochemistry are converted to isomeric SMILES. The non-standard InChI containing a fixed-hydrogen layer is used instead of the standard InChI to distinguish tautomers. Those InChIs are used as unique identifiers to identify different solutes or solute-solvent pairs in the data sets. When multiple data points are found for the same solute or solute-solvent pair, the mean values of the Abraham solute parameters, $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, or $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ are used and the standard deviations are calculated. No data are removed based on the standard deviations as most of the data are found to have relatively low standard deviations.

Abraham solute parameters. The data statistics of the in-house Abraham solute parameter database (SoluteDB) are summarized in Table 1. The number of data points per parameter varies as some solutes have missing parameters. The number of Abraham and Mintz solvent parameters used in this study are provided in the table as well. These solvent parameters are empirically fitted based on the collected solute parameters, $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ data. The details on the fitting method, the fitted solvent parameters, and the molar weight distribution of the solutes in the data set can be found in the Supporting Information Sections 1 and 2.1.

Table 1: Data summary for the in-house Abraham solute parameter database (SoluteDB) and the number of fitted solvent parameters. The total number of data points (N total), mean values, standard deviations (std. dev.), and minimum and maximum (min, max) values are presented.

solute parameter	N total	mean	std. dev.	min	max
E	8163	0.97	0.81	-1.51	6.87
S	7654	1.14	0.78	-1.60	10.97
A	8159	0.21	0.35	0.00	6.80
B	7395	0.66	0.61	0.00	7.97
L	7038	6.77	3.79	-1.20	49.98
Total number of solutes				8366	
Number of solvents with fitted Abraham solvent parameters (e, s, a, b, l, c for ΔG_{solv})				195	
Number of solvents with fitted Mintz solvent parameters (e', s', a', b', l', c' for ΔH_{solv})				66	

Solvation free energy and enthalpy. A summary and analysis of the compiled solvation free energy and enthalpy data sets are given in Table 2. The solvation free energy data are acquired from the Minnesota solvation database (MNSol),[38] CompSol database,[51] FreeSolv database,[35] and published work by Abraham, Acree and coworkers[52–83] and compiled together as a dGsolvDB1 data set. FreeSolv database is obtained from MoleculeNet.[36] Additionally, we have nearly 4000 gas-water partition coefficient data ($\log K_{\text{w}}$) from the in-house database. While adding these data significantly increases the size of the data set and the number of solute compounds considered, it also causes the solvation free energy data to be more heavily biased towards water as a solvent. Therefore, a separate data set, dGsolvDB2, is prepared that includes the dGsolvDB1 data and the aqueous solvation free energy data converted from the in-house $\log K_{\text{w}}$ data set.

We also convert octanol-water partition coefficient data ($\log P_{\text{ow}}$) from the in-house database and solvent-water partition coefficient ($\log P$) data from OCHEM,[84] DrugBank,[85] PHYSPROP,[86] and published work by Abraham, Acree, and coworkers[52,53,57–75,78–82,87] to solvation free energies. For the solute species with $\log K_{\text{w}}$ and $\log P$ data available, the gas-solvent partition coefficient ($\log K$) can be calculated from Eq. 4.

$$\log(K) \approx \log(P) + \log(K_{\text{w}}) \tag{4}$$

$\log K$ is subsequently converted to the solvation free energy through Eq. 2. Note that Eq. 4 assumes that $\log P$ is measured in dry solvents, *i.e.* the solvents do not dissolve into one another while in contact. This approximation would introduce an additional uncertainty to the data set, especially for polar solvents in contact with water. Moreover, including the in-house $\log P_{\text{ow}}$ data causes the data set to be biased toward 1-octanol as a solvent. Thus, another separate data set, dGsolvDB3, is prepared that includes the dGsolvDB2 data and

the solvation free energy data converted from the $\log P_{ow}$ and $\log P$ data. As Table 2 shows, the dGsolvDB3 data set has the most data but nearly a half of the solvents correspond to water and 1-octanol. All three solvation energy data sets (dGsolvDB1 - dGsolvDB3) are used separately to build DirectML models, and the test set errors are compared to determine which data set gives the best prediction.

Solvation enthalpy data are collected from Acree Enthalpy of Solvation data set[88] and CompSol database.[51] Self-solvation data (solvation of a compound in itself) are included in both solvation energy and enthalpy data sets. It can be seen from Table 2 that the majority of the solvents in the data sets only have a single entry that corresponds to the self-solvation datum in the final database. As a result, if self-solvation data are excluded, the number of solvents would reduce to 302 and 142 solvents in the solvation free energy and enthalpy data sets, respectively. The molar weight distribution of the solutes in each data set is provided in the Supporting Information Section 2.1.

Table 2: Summary of the data sets for solvation free energy (dGsolvDB) and enthalpy (dHsolvDB), including the total number of data points (N total), the number of solutes (N solutes), the number of solvents with and without solvents that only appear in self-solvation (N solvents), and a list of the most commonly found solvents. (CCl4: carbon tetrachloride, DMF: dimethylformamide.)

data set (included data)	N total	N solutes	N solvents (excl. self-solv)	top 5 solvents (% data)
dGsolvDB1 (ΔG_{solv} [35,38,51–83])	12202	2387	1459 (302)	water (11.6 %) 1-octanol (3.2 %) hexadecane (2.1 %) heptane (1.7 %) hexane (1.7 %)
dGsolvDB2 (ΔG_{solv}, [35,38,51–83] in-house $\log K_w$ data)	16180	5991	1459 (302)	water (33.3 %) 1-octanol (2.4 %) hexadecane (1.6 %) heptane (1.3 %) hexane (1.3 %)
dGsolvDB3 (ΔG_{solv}, [35,38,51–83] in-house $\log K_w$ data, in-house $\log P_{ow}$ data, $\log P$ [52,53,57–75,78–82,84–87])	20253	5991	1459 (303)	water (26.6 %) 1-octanol (21.1 %) hexadecane (1.3 %) ethanol (1.1 %) heptane (1.1 %)
dHsolvDB (ΔH_{solv} [51,88])	6322	1665	1432 (142)	cyclohexane (3.5 %) methanol (3.2 %) benzene (3.1 %) CCl4 (2.8 %) DMF (2.8 %)

Comparison of the LSER estimates and experimental data. Based on the collected data sets, the errors associated with the LSER are evaluated prior to the construction of the prediction models. Solvation free energies and enthalpies are calculated from the Abraham and Mintz LSERs using the experimental solute parameters (SoluteDB) and the fitted solvent parameters, and the calculated values are compared to the $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ data sets listed in Table 2. The parity plots and errors are shown in Figure 2. It can be seen that the LSER has small $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ error and relatively higher $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ error. These errors are due to both the aleatoric uncertainty of the experimental data and the fact that the linearity of the LSER cannot capture the non-linear relationship between the solute parameters and solvation energy/enthalpy. Since the LSER estimates of both solvation free energy and enthalpy in Figure 2 are calculated using the same experimental solute parameters, it is likely that the model error associated with the Mintz LSER (Eq. 3) is higher than that of the Abraham LSER (Eq. 1). It is also possible that the solvation enthalpy data (dHsolvDB) have higher experimental uncertainty than the solvation energy data (dGsolvDBx), but we are unable to assess the experimental uncertainties of the data sets used in this work as the majority of solvent-solute pairs only have a single data point. The SoluteGC and SoluteML models that predict $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ through the LSER will, at their best performance, have these underlying errors when compared to the dGsolvDBx and dHsolvDB data sets.

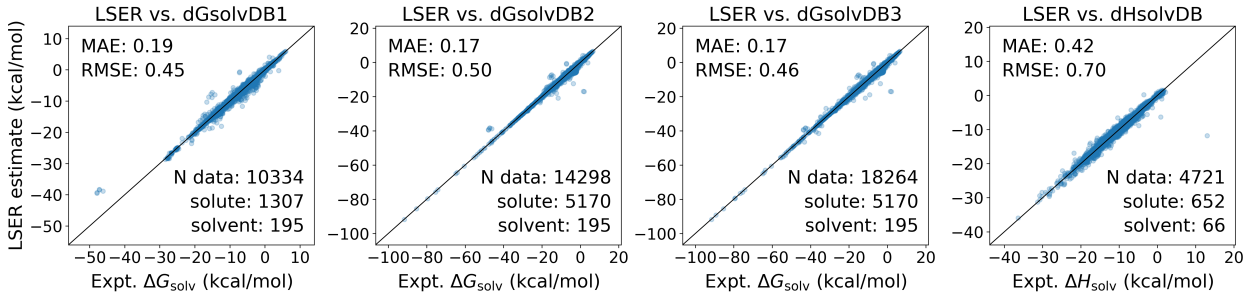

Figure 2: Comparison between the experimental $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ data and the estimations from the LSER. The LSER estimations are calculated using the Abraham solute parameter database and the fitted solvent parameters. The experimental data from dGsolvDB1, dGsolvDB2, dGsolvDB3, and dHsolvDB are compared to the LSER estimates. The mean absolute error (MAE) and root mean square error (RMSE) are reported in kcal/mol.

2.2 Data Split for Model Comparison

As summarized in Tables 1 and 2, a total of four data sets are prepared for the prediction of solvation free energy: (1) SoluteDB, (2) dGsolvDB1, (3) dGsolvDB2, and (4) dGsolvDB3. The final models that are available through the easy-access tool are trained on all of the data available. However, to evaluate and compare the performance of the models, each data set is split into a ∼90 % training/validation and a ∼10 % carefully selected test set. The test set is constructed such that the solute compounds from the test set do not overlap with the training and validation set. For example, if compound A is chosen as a test set

solute, all solvent-solute pairs that have compound A as a solute are included in the test sets and excluded from the training and validation sets. The test set solutes are selected in a (1) random and (2) substructure-based manner while maintaining a ∼90/10 % data split. The substructure-based splits are employed to test the out-of-range performance of the models on new classes of solute molecules. All solute compounds that contain any of the selected functional groups and scaffolds are included in the test set for our substructure splits. Those splits are comparable to the Murcko scaffold-based splits that are commonly used in molecular property predictive models for drug discovery.[36,37] The substructures are manually selected to maintain a ∼90/10 % split for all data sets. The substructures are represented in SMARTS,[89] and substructure search on solute compounds is done using RDKit. Examples of chosen substructures are benzoic acid, adamantane and phenanthrene scaffolds, and a trifluoromethyl group. The list of all substructures used to split the data sets is presented in the Supporting Information Section 2.2. The data sets used for the prediction of solvation enthalpy (SoluteDB and dHsolvDB) are split in the same fashion.

The training and testing sets we have prepared so far are designed for evaluating the model performance on unseen solutes. To investigate the performance of the DirectML models on unseen solute and unseen solvent pairs, 10 and 8 solvents are removed from the training sets of dGsolvDB and dHsolvDB, respectively, and placed in the test sets. The solvents with Abraham and Mintz solvent parameters are selected to allow the solvent-wise error comparison with the SoluteGC and SoluteML models. Moreover, the solvents are selected such that the number of solutes in each training set remains unchanged even after the chosen solvent data are excluded from the training set. The list of excluded solvents is provided in the Supporting Information Section 2.3, and detailed data statistics of all training and test sets used in this study is presented in the Supporting Information Section 2.4.

3 Methods – Models

3.1 SoluteGC: Group Contribution Method for Abraham Solute Parameter Prediction

The SoluteGC model is built as part of Reaction Mechanism Generator (RMG),[90,91] open source chemical kinetic modeling software package. RMG uses the Benson-type[92] group contribution method for gas phase thermochemistry estimations and has over 2000 groups in the database. For the SoluteGC model, RMG's gas phase group contribution scheme is adopted to include the Abraham solute parameters, and missing groups that are important to the solvation data sets are added.

RMG's GC method estimates thermochemistry by dividing a molecule into atom-centered (AC) functional groups and summing the contribution from all groups. Additionally, RMG implements ring strain correction (RSC) and long distance interaction (LDI) groups to ac-

count for more advanced structural effects that cannot be captured by the atom-based approach. The SoluteGC model follows the same scheme to calculate the solute parameters as shown in Eq. 5

$$E, S, A, B, \text{ or } L = \sum_{i=1}^{N_{\text{atom}}} \text{AC}_i + \sum_{j=1} \text{RSC}_j + \sum_{k=1} \text{LDI}_k \tag{5}$$

where N_{atom} is the number of heavy atoms in a molecule and RSC and LDI corrections are applied for each ring cluster and long distance interaction group found in a molecule, respectively. For more details on the GC scheme, the reader is referred to the dedicated work by Gao et al.[90] and Han et al.[93] as well as the RMG documentation.[94]

While most molecules follow Eq. 5, halogenated molecules are treated differently in the SoluteGC model: all halogen atoms are first replaced by hydrogen atoms, then the GC estimate is made on the replaced structure, and halogen corrections are lastly added for each halogen atom to get the final GC prediction as shown in Eq. 6

$$\begin{aligned} E, S, A, B, \text{ or } L &= \sum_{i=1}^{N_{\text{atom}*}} \text{AC}_i + \sum_{j=1} \text{RSC}_j + \sum_{k=1} \text{LDI}_k + \sum_{l=1}^{N_{\text{halogen}}} \text{Halogen}_l \\ &= (E, S, A, B, \text{ or } L)_{\text{replaced compound}} + \sum_{l=1}^{N_{\text{halogen}}} \text{Halogen}_l \end{aligned} \tag{6}$$

where the subscript 'replaced compound' denotes the compound whose halogen atoms are replaced by hydrogen atoms, and $N_{\text{atom}*}$ and N_{halogen} represent the number of non-halogen heavy atoms and the number of halogen atoms in a molecule, respectively. The halogen groups are halogen-centered functional groups and defined by neighboring atoms including other halogens that are bonded to the same atom. We take this unique approach because it allows one to use the experimental solute parameter data of a replaced compound and simply apply halogen corrections to get more accurate estimates for a halogenated compound. While this approach is inspired by various works,[95,96] it is primarily based on the hydrogen bond increment (HBI) method devised by Lay et al.,[95] in which a radical correction is applied to a saturated compound datum to get a thermochemistry estimate for a radical compound.

Once the types of functional groups to be considered and the GC relationships are established (Eqs. 5 and 6), a set of functional groups relevant to our data set needs to be created before fitting the associated group values. The functional groups for the SoluteGC model are constructed largely based on the existing groups within RMG. The groups that do not appear in the Abraham solute parameter database (SoluteDB) are removed, and new AC, RSC, and halogen groups that frequently appear in the database but are missing from RMG, mainly halogen, phosphorus, and heterocyclic groups, are created. The number of groups used for the final SoluteGC model is listed in Table 3.

Table 3: Description and the number of groups used for the SoluteGC model.

Group Category	Number of Groups	Description
AC-regular	729	Atom-centered functional groups. Applied for each non-halogen heavy atom
AC-halogen	193	Halogen correction (halogen-centered functional groups)
RSC-ring	116	Monocyclic ring strain correction
RSC-polycyclic	179	Polycyclic ring strain correction
LDI-cyclic	29	Aromatic ortho, meta, para correction
LDI-noncyclic	16	Gauche interaction correction

Finally, the group values are fitted to the experimental data using the ridge regression method from Scikit-learn package.[97] Ridge regression is chosen as a fitting method because it gave overall the lowest error on our test sets compared to other regression methods such as ordinary least squares, lasso, and elastic net. Ridge regression[98] is a linear least squares regression with L2 regularization and can help prevent overfitting. The hyperparameter of the regression is tuned using 10-fold cross-validation on each training set prepared from SoluteDB in the earlier section. The final SoluteGC model implemented in RMG is fitted using the entire data set, omitting the very small molecules where GC approach is unsuitable. All molecules with more than two heavy atoms are used to fit the group values. The group definitions and the number of data used to fit each group can be found on GitHub as part of the RMG-database at `https://github.com/ReactionMechanismGenerator/RMG-database/tree/master/input/solvation/groups`. Detailed information on how the groups are defined and organized in the RMG-database is given in the Supporting Information Section 3.1.

3.2 SoluteML and DirectML: Machine Learning for Solute Parameters, Solvation Free Energy, and Enthalpy Prediction

Two machine learning models are developed for the prediction of solvation free energy and enthalpy. The first deep neural network ensemble (SoluteML) is trained on the database with the Abraham solute parameters (SoluteDB). The average prediction of this ensemble of neural networks is combined with the solvent parameters through the LSER (Eqs. 1 and 3) to calculate $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$. The second ensemble of deep neural networks (DirectML) is trained on dGsolvDB and dHsolvDB and used to predict the solvation properties directly. For both machine learning models, the final models are trained on the complete data sets. For the purpose of comparing model performance in the remainder of this work, the models are trained on the different splits (see Section 2.2).

The model architecture for both models used in this work are based on the state-of-the-art

chemical property prediction software Chemprop.[37] This software uses a directed message passing neural network (D-MPNN), a type of graph convolutional neural network, to convert atom and bond feature vectors to a molecular latent representation. The embedded molecule is subsequently sent through a second neural network for the property prediction task. For more details on the D-MPNN, the reader is referred to the dedicated work by Yang et al.[37] The software package can be found on GitHub (`https://github.com/fhvermei/chemprop_solvation`). Specifics related to this work are discussed below.

For the SoluteML model, the training set consists of solute SMILES as input and their corresponding solute parameters as output to the neural network. Minor adjustments are made compared to the original version of Chemprop to include atom, bond, and molecular features specific to solvation. A summary of the used atom and bond features is given in Table 4. Different sets of additional molecular descriptors are tested for training the neural network. These are concatenated with the molecular latent representation generated by the D-MPNN. We find that using all the available 2D-RDKit molecular descriptors (a vector of 200 features automatically generated by RDKit[37]) yields overall the best performance for the SoluteML model, and therefore they are chosen as the additional molecular features.

For the DirectML model, the original version of Chemprop is adapted to the application of multiple input molecules, a solvent-solute pair. The solvent and solute SMILES strings are each converted to a latent representation by a separate D-MPNN, after which the embeddings are concatenated prior to the second neural network for property prediction. The same atom and bond features are used for the D-MPNN as for the SoluteML model (see Table 4). Similar to SoluteML, different sets of additional molecular descriptors are tested to improve the performance of the model. In this case, the selected molecular features are the RDKit-calculated octanol-water partitioning coefficient and Van der Waals Surface Area (VSA) combined in RDKit's `SlogP_VSA_` descriptor. The molecular feature vectors for a solute and a solvent are concatenated with the latent representation after the D-MPNN and prior to the second neutral network for property prediction.

Different hyperparameters are used for the SoluteML and the DirectML models. For each model, the hyperparameters are selected based on a search algorithm that includes optimization of the neural network for different hyperparameter combinations. The search procedure combines the algorithms available in the software package Hyperopt[99,100] with manual intervention to balance the size of the neural networks with the gain in accuracy. The hyperparameters that are optimized include the depth and hidden size of the D-MPNN, the number of layers and hidden size of the neural network for property prediction, the learning rates, and the batch size. An overview of the selected hyperparameters for the different models is given in the Supporting Information Section 3.2.

The models are trained with 5-fold cross-validation. For each of the different splits, the test set is pre-defined as detailed in Section 2.2. During the 5-fold cross-validation, a different

Table 4: Atom and bond features used for the directed message passing neural network.

Atom features	**Type**	**Bond features**	**Type**
Atomic number	One-hot	Bond type	One-hot
Total neighbor number	One-hot	Conjugation	One-hot
Formal charge	One-hot	In ring	One-hot
Connected hydrogen atoms	One-hot	Stereo	One-hot
Hybridization	One-hot		
Lone pairs	One-hot		
H-bond donor	One-hot		
H-bond acceptor	One-hot		
Ring size	One-hot		
Aromaticity	One-hot		
Electronegativity	Binary		
Molar mass	Binary		

random validation set of 10 % is generated and distinguished from the training set to determine at which iteration (or epoch) to stop training the model. Within each fold, an ensemble of 5 different models is generated by Glorot initialization of the model parameters[101] with different seeds. All models are used to make predictions for the pre-defined test set. The predictions of each individual model are averaged to calculate the prediction of the model ensemble. Additionally, the variance on the predictions of the individual models is used as a measure for the model uncertainty (*i.e.* epistemic uncertainty) on the individual data points.

3.3 Prediction Using Existing Methods

The performance of our models is compared to the following quantum chemistry (QM), ML, and GC methods from literature: SMD,[9] COSMO-RS,[12] the solvation free energy ML model by Lim and Jung (MLSolvA),[43] the transfer learning model by Vermeire and Green[44] (transfer learning), the solvation enthalpy ML model by Jacquis et al.,[102] and the solute parameter GC method from the UFZ-LSER database (UFZ-LSER).[29] The COSMO-RS calculations are performed in-house at the BP86/TZVPD-FINE level of theory using the software COSMO*therm*.[103] These calculations are done for the solvent-solute pairs whose pre-calculated quantum chemical COSMO data are available in the COSMO*base* database.[104] Note that no additional quantum chemical calculations are done in this work. The solvation enthalpy at 298 K is computed by calculating solvation free energies at 297, 298, 299 K, estimating the temperature gradient at 298 K from the three data points, and using the relationship $\Delta H = \Delta G - T\frac{\mathrm{d}\Delta G}{\mathrm{d}T}$. The GC method from the UFZ-LSER database[29] is used to predict the solute parameters. These predicted parameters are combined with our in-house solvent parameters to compute the prediction errors of the UFZ-LSER GC method. Since the training data used for the regression of the solute parameters in the UFZ-LSER database are unknown, the errors are evaluated using all of our data. Only the solute molecules with more than two heavy atoms are used to evaluate the UFZ-LSER GC method as the GC

method is not suitable for small molecules. For the remaining methods, SMD[9] and the ML models,[43,44,102] the reported errors from literature are used for comparison. Note that these errors are reported on a different test set than the one used in this work.

4 Results and Discussion

In the subsequent section, we evaluate the performance of the three models (SoluteGC, SoluteML, and DirectML) on 10 % test sets for both random and substructure-based solute splits. The comparison is done for the test splits with unseen solutes, and additionally for the DirectML model, on unseen solute and unseen solvent pairs. Because the size of the test set is different for each model, only overlapping test data of the compared models are considered in our comparison. For the comparison of the three models, the test solvents are limited to those with Abraham or Mintz solvent parameters since the SoluteGC and SoluteML models can be evaluated on only those solvents (see Table 1). The results on the entire test set data of each model can be found in the Supporting Information Section 4, which includes the results on the solvents without the solvent parameters for the test sets of the DirectML model. Furthermore, only the prediction of the solvation free energy and enthalpy is discussed in the main text. The prediction results on the individual solute parameters are presented in the Supporting Information Section 4. At last, the performance of our models is compared to that of the existing quantum chemistry, ML, and GC models.

The model performance is analyzed by comparing parity plots, the mean absolute error (MAE), and the root-mean-square error (RMSE), both in kcal/mol. Because the scale of the solvation free energies and their errors differ between some of the compared test sets, we also compare the relative error using the percent MAE (PMAE) defined as:

$$\text{PMAE} = \left| \frac{\text{MAE}}{\text{test mean}} \right| \cdot 100\ \% \tag{7}$$

The test mean represents the average experimental value in the test set. PMAE is used instead of the mean relative error, which is a mean of the individual absolute errors divided by their experimental values, because several solvation free energy and enthalpy have experimental values close to zero. This leads to inflation of the mean relative error, which would not be representative of the model performance.

4.1 Comparison of the DirectML Models Trained on Different Solvation Free Energy Data Sets

Before comparing the performance of the different models (SoluteGC, SoluteML, and DirectML), we start by comparing the performance of the DirectML models that are trained

and validated on three different solvation free energy data sets (dGsolvDB1, dGsolvDB2, dGsolvDB3) to determine which model gives the best prediction. As explained in Section 2.1 and summarized in Table 2, the data sets differ in what kind of solvation data is included (ΔG_{solv}, in-house $\log K_{\text{w}}$ data, and $\log P$ data). The size of the data set increases from dGsolvDB1 to dGsolvDB3, including more solutes, but the solvents become more skewed towards water and 1-octanol as $\log K_{\text{w}}$ and $\log P$ data are added. To select the best performing model, in this section, only overlapping test data between the three DirectML models trained with the three data sets are compared. Because each data set is the subset of the next larger data set (dGsolvDB1 $\subset$ dGsolvDB2 $\subset$ dGsolvDB3), the overlapping test data only include the ΔG_{solv} data from dGsolvDB1 and do not include in-house $\log K_{\text{w}}$ and $\log P$ data. The results on the entire test set of each model are given in the Supporting Information Section 4.1.

The test errors of the three models on the overlapping test set data are shown on the parity plots in Figure 3. For a random split, the performance of the three models is very similar. The model trained on dGsolvDB3 gives the lowest MAE of 0.29 kcal/mol, which is only 0.02 kcal/mol lower than the model trained on dGsolvDB1. The difference between the models is a bit more pronounced in the substructure split. The model trained on dGsolvDB3 gives the lowest MAE of 0.81 kcal/mol, which is 0.06 kcal/mol lower than the model trained on dGsolvDB1. The result indicates that adding both $\log K_{\text{w}}$ and $\log P$ data to the training and validation data set slightly improves solvation energy predictions compared to training and validating the models on only solvation free energy data. Yet, considering that the total number of data and the number of solutes in dGsolvDB3 are nearly twice as many as those in dGsolvDB1, the improvement is not as significant as expected. In the Supporting Information Figure S3, a similar comparison is done using the complete (10 %) test set of each respective data set rather than only the overlapping data. In this case, the model trained, validated, and tested on dGsolvDB1 gives the lowest MAE of 0.31 kcal/mol compared to 0.40 kcal/mol for the model trained, validated, and tested on dGsolvDB3 for a random split. For the substructure split, the difference is less significant with the lowest MAE of 0.87 kcal/mol for dGsolvDB1 closely followed by 0.89 kcal/mol for dGsolvDB3. If the PMAE is compared instead, the dGsolvDB1 and dGsolvDB3 models both have the lowest PMAE of 4.9 % for the random split and the dGsolvDB3 model has the lowest PMAE of 9.2 % for the substructure split compared to the PMAE of 10.1 % for the dGsolvDB1 model.

These observations are in line with the conclusions made in earlier work by Vermeire and Green.[44] The model performance improves with an increasing amount of data in the training and validation set; however, the extent to which the model performance can be assessed is limited by the experimental (or aleatoric) uncertainty in the test set. For the same test set, the performance of the three models on a random split is similar since the aleatoric limit of assessing the performance on that test set is reached. For the substructure split, there is still a slight improvement in performance observed by the addition of more data to the training and validation set. For the comparison against the complete (10 %) test set of each data set in the Supporting Information, it is expected that the test splits of the data

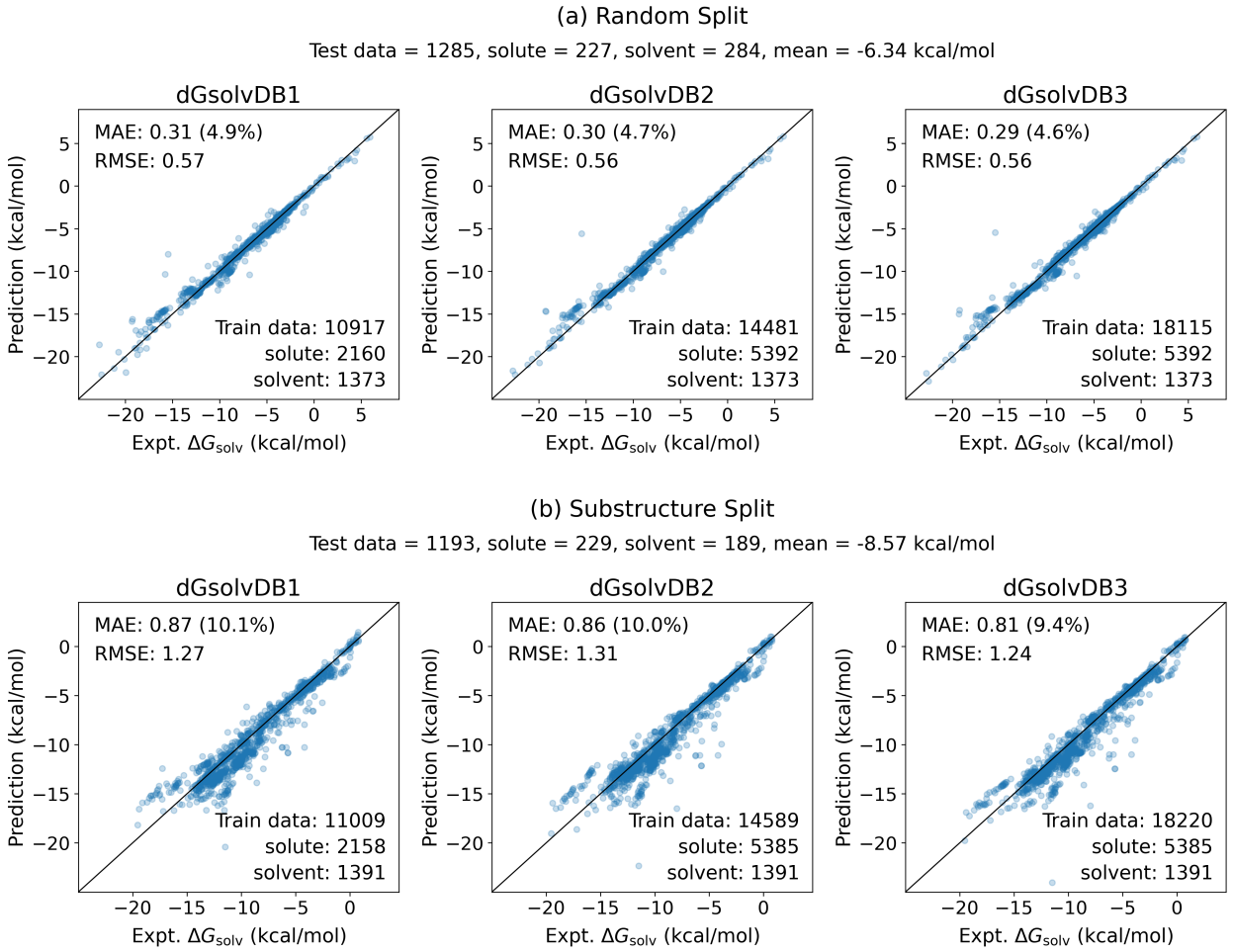

Figure 3: Parity plots of the DirectML models trained and validated using 3 different solvation free energy data sets (dGsolvDB1, dGsolvDB2, and dGsolvDB3) that differ in the type and amount of solvation data considered. The plots only show the overlapping test data in each split type. The MAE and RMSE are in kcal/mol, and the PMAE is given in parenthesis. Information on the training and test data are given on the figures together with the overall errors. Note that the test sets are prepared using random and substructure-based solute splits, and therefore none of the test set solutes appear in the training and validation sets.

sets containing $\log K_{\mathrm{w}}$ and $\log P$ data have a higher associated experimental uncertainty. The experimental uncertainty in solvation free energies of neutral compounds is typically estimated as 0.2 kcal/mol,[105,106] but the solvation free energies measured in water ($\log K_{\mathrm{w}}$) often have higher uncertainties,[107] which we believe are due to larger magnitudes of the values and disagreements between data reported by different sources. The conversion of $\log P$ data to solvation free energies also comes along with an additional uncertainty because of the assumption that solvents are not in contact with one another. As a result, the model trained and validated on dGsolvDB1 seems to have better performance if we compare the MAE evaluated on the complete test set. However, this conclusion cannot be directly drawn

because of (1) the additional uncertainty in the other test sets and (2) the different range of the magnitudes of solvation free energies in the test set data. The latter is compensated for by comparing the PMAE of the different models instead of the MAE or RMSE.

Overall, including more data in the training and validation set is beneficial as it lowers the prediction error when assessed on the overlapping test set data, and thus dGsolvDB3 is chosen as the optimal solvation free energy data set for this work and comparison to other methods. We also compare the solvent-wise test errors of the three DirectML models in the Supporting Information Figure S4, and it is found that having data bias towards two opposing solvents like water and 1-octanol does not cause any particularly high errors in other solvents. However, since the accuracy of the three models is not significantly different, it is possible that different data splits or data sizes can yield contrasting results.

4.2 Comparison of the Three Prediction Approaches

4.2.1 Performance on Unseen Solutes

The test set errors for the solvation free energies $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ predicted by the SoluteGC, SoluteML, and DirectML models are presented in Figure 4. The three different models are tested on unseen solutes, and only overlapping test data of the three models are compared in the figure. All solvents in the overlapping test data appear in the training set of the DirectML model, and the SoluteGC and SoluteML models use the empirical solvent parameters, and therefore, the results shown in Figure 4 can be considered as the predictive performance on pairs of unseen solutes and trained solvents. The DirectML model is trained and validated using the data set dGsolvDB3 (including ΔG_{solv} data, in-house $\log K_{\mathrm{w}}$ data, and $\log P$ data), which is selected based on earlier results from Section 4.1. Note that the overlapping test data in Figure 4 are different from those in Figure 3, and therefore the test errors of the DirectML model from the two figures are slightly different each other.

For both the random and substructure solute splits, the DirectML model achieves the best predictions and the SoluteGC model gives the highest error. While the DirectML model performs better than the SoluteML model, there is no significant difference; the MAE differs less than 0.1 kcal/mol for both splits. The SoluteML and DirectML models have likely reached close to the aleatoric limit of the experimental data in the random split, and we believe that the relative underperformance of the SoluteML model in comparison to the DirectML model is due to the inherent error caused by the linearity the LSER as discussed in Section 2.1. Even though the SoluteML model has about 2100 more solutes in the training set than the DirectML model has, the additional solute data are not able to compensate for the underlying error.

The three models are compared for the prediction of solvation enthalpies $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ on

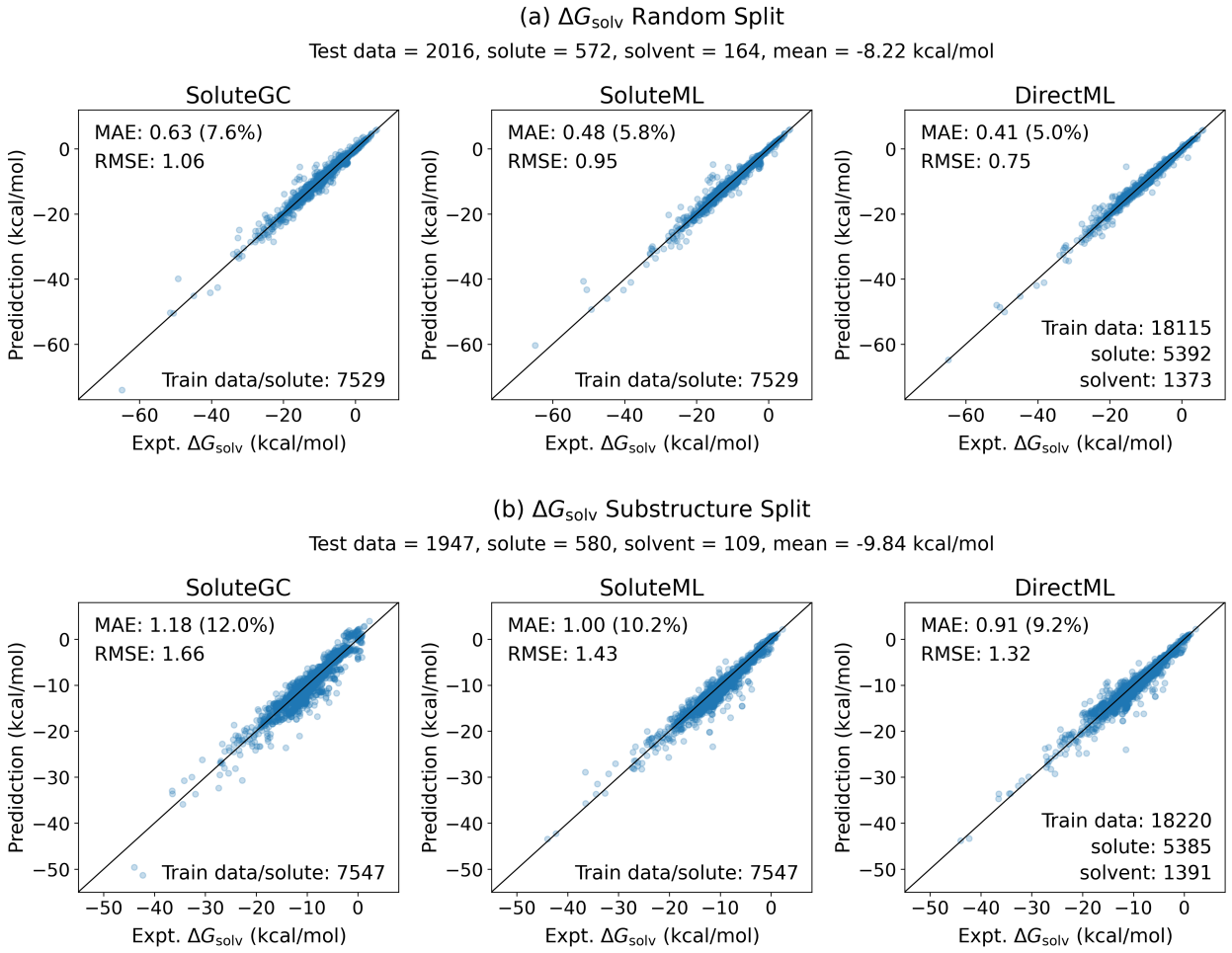

Figure 4: Parity plots for experimental and predicted $\Delta G_{\text{solv}}(298\,\text{K})$ on the 10 % test sets. The plots only show the overlapping test data in each split type. The three different models are tested on pairs of unseen solutes and trained solvents for the random and substructure solute splits. The MAE and RMSE are in kcal/mol, and the PMAE is given in parenthesis. Information on the training and test data are given on the figures together with the overall errors.

unseen solutes in Figure 5. Again, all solvents in the overlapping test data appear in the training set of the DirectML model, and hence, the figure reflects the results on pairs of unseen solutes and trained solvents. The DirectML model is trained using much fewer data that contain only about 1500 solutes due to the limited availability of data on solvation enthalpies. Nevertheless, the DirectML model achieves a similar accuracy as the SoluteML model on a random split and outperforms the SoluteML model on the substructure split.

The MAE of each substructure used for the substructure splits is compared for the three models in Figure 6. The SMARTS strings and drawings of the substructures can be found in Section 2.2 of the Supporting Information. The three models have errors of similar magnitudes

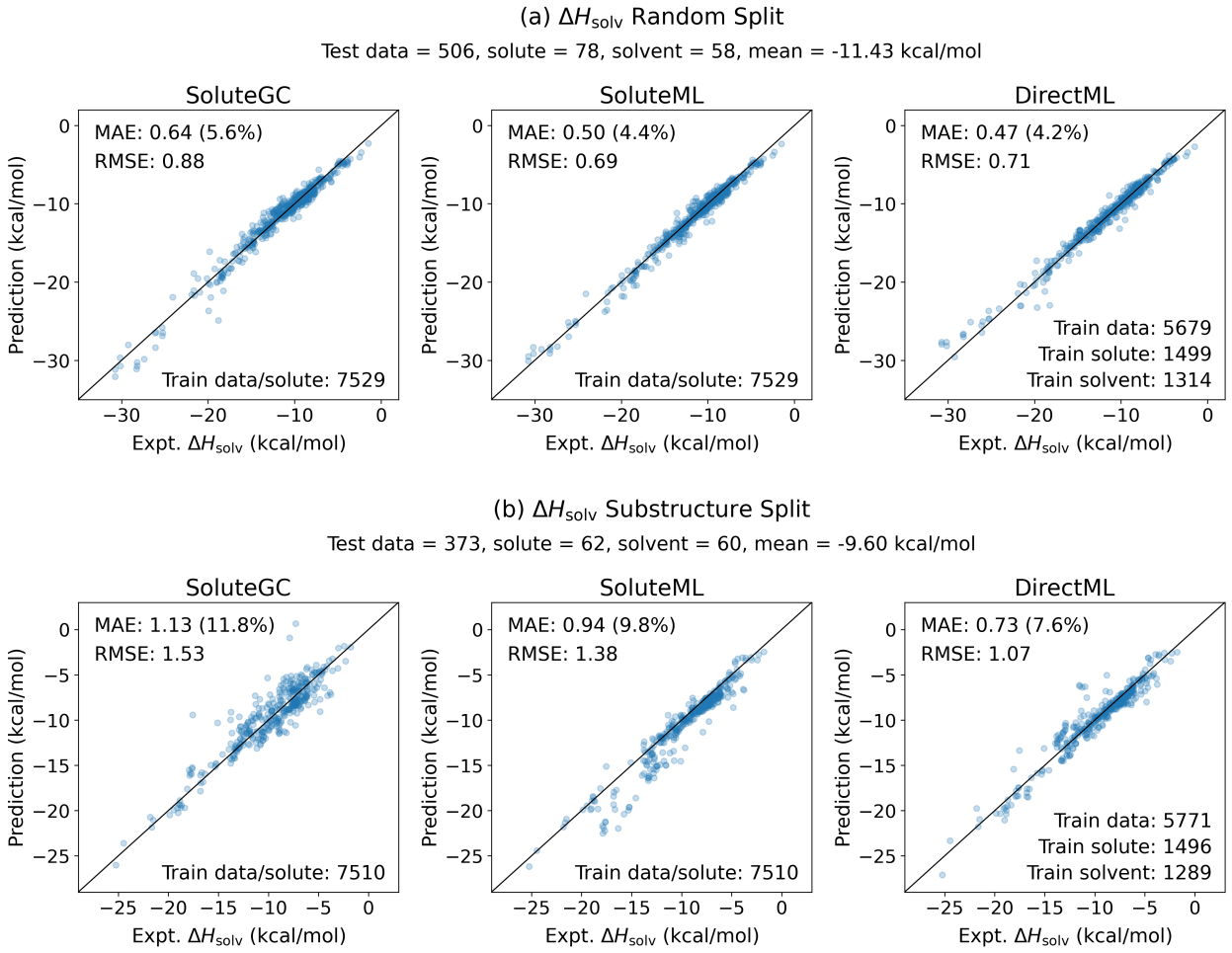

Figure 5: Parity plots for experimental and predicted $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ on the 10 % test sets. The plots only show the overlapping test data in each split type. The three different models are tested on pairs of unseen solutes and trained solvents for the random and substructure splits. The MAE and RMSE are in kcal/mol, and the PMAE is given in parenthesis. Information on the training and test data are given on the figures together with the overall errors.

for many substructures, but they have very different levels of accuracy for some substructures. For example, for the $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ predictions, the solutes containing methanesulfonamide functional group (SMARTS: NS(=O)(=O)C) are the main outliers of the SoluteGC and SoluteML models whereas the solutes with cyclopentene scaffold (C1=CCCC1) give the highest error for the DirectML model. The difference between the key outliers is more pronounced for the $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ predictions; here, the SoluteML model has much higher MAE than the other two models for the solutes containing adamantane scaffold (C1C2CC3CC1CC(C2)C3) while it has much lower error for imidazole scaffold (c1c[n&H1]cn1), which is the main outlier of the other two models.

In the random splits, the common outliers of the three models are cyclic or polycyclic so-

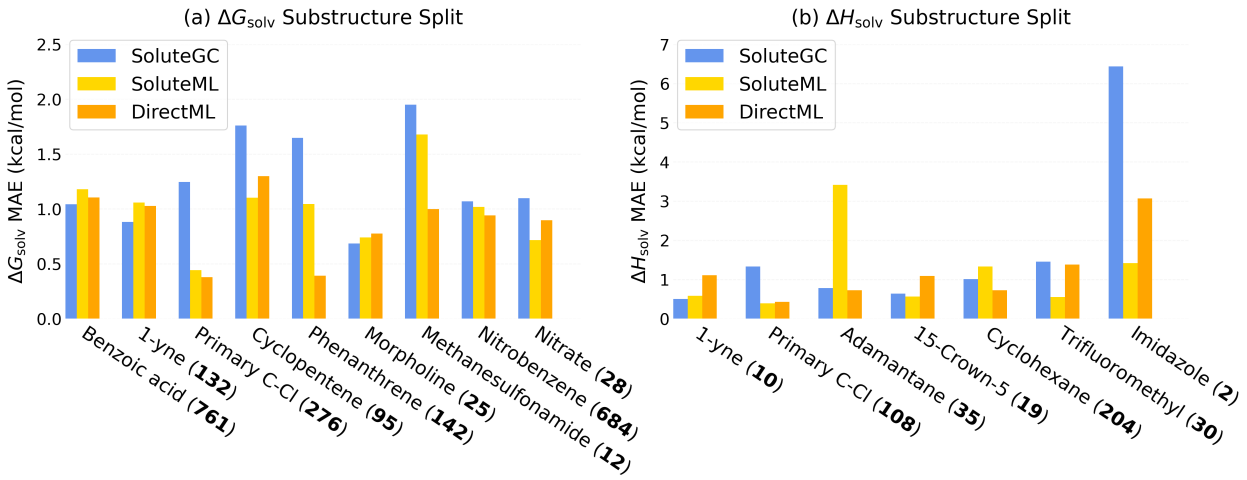

Figure 6: The MAE of each substructure for the prediction of $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ in the substructure split. Only the overlapping test data of the three models in each split type are compared in the plots. The number of solute data containing the corresponding substructure in each test set is given in parenthesis in bold.

lute compounds such as guanabenz, galantamine, methotrexate, and colchicine for ΔG_{solv} predictions and salicylamide, pyrene, 1,4-diphenylbenzene, and benzo-15-crown-5 for ΔH_{solv} predictions. However, the majority of the main outliers are found to be different for each model.

The solvent-wise errors of the three models are compared for the 30 most frequently appearing solvents in each test set as shown in Figure 7. Note that the figure shows the results on the unseen solutes in the trained solvents. For the majority of the solvents for $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ prediction, the DirectML model performs similar to or better than the SoluteML model, and the SoluteGC model gives the highest errors. In the random split set of $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, the SoluteML model performs better than the DirectML model for non-polar solvents such as alkanes, carbon tetrachloride (CCl4), isooctane, and p-xylene. However, the DirectML model gives the best predictions for the same non-polar solvents in the substructure split set. Similarly, the SoluteGC and SoluteML models have much higher errors for dimethylformamide (DMF), ethyl acetate, and ethyl ether than the DirectML model in the random split, but their performance is similar to the DirectML model in the substructure split for the same solvents (with exception of the SoluteML model for DMF). No clear correlation between the model performance and the types of solvents is found from the results as there are no specific types of solvents that a certain model always underperforms or outperforms. It is also noted that the three models have very similar error scales for most of the solvents.

Similar to the prediction of $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, the DirectML model performs better than the other two models for the majority of the solvents in the prediction of $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ in Figure 7. Some exceptions include hexadecane, ethanol, dichloromethane, and tert-butanol for which the SoluteML model gives much lower errors than the DirectML model in the

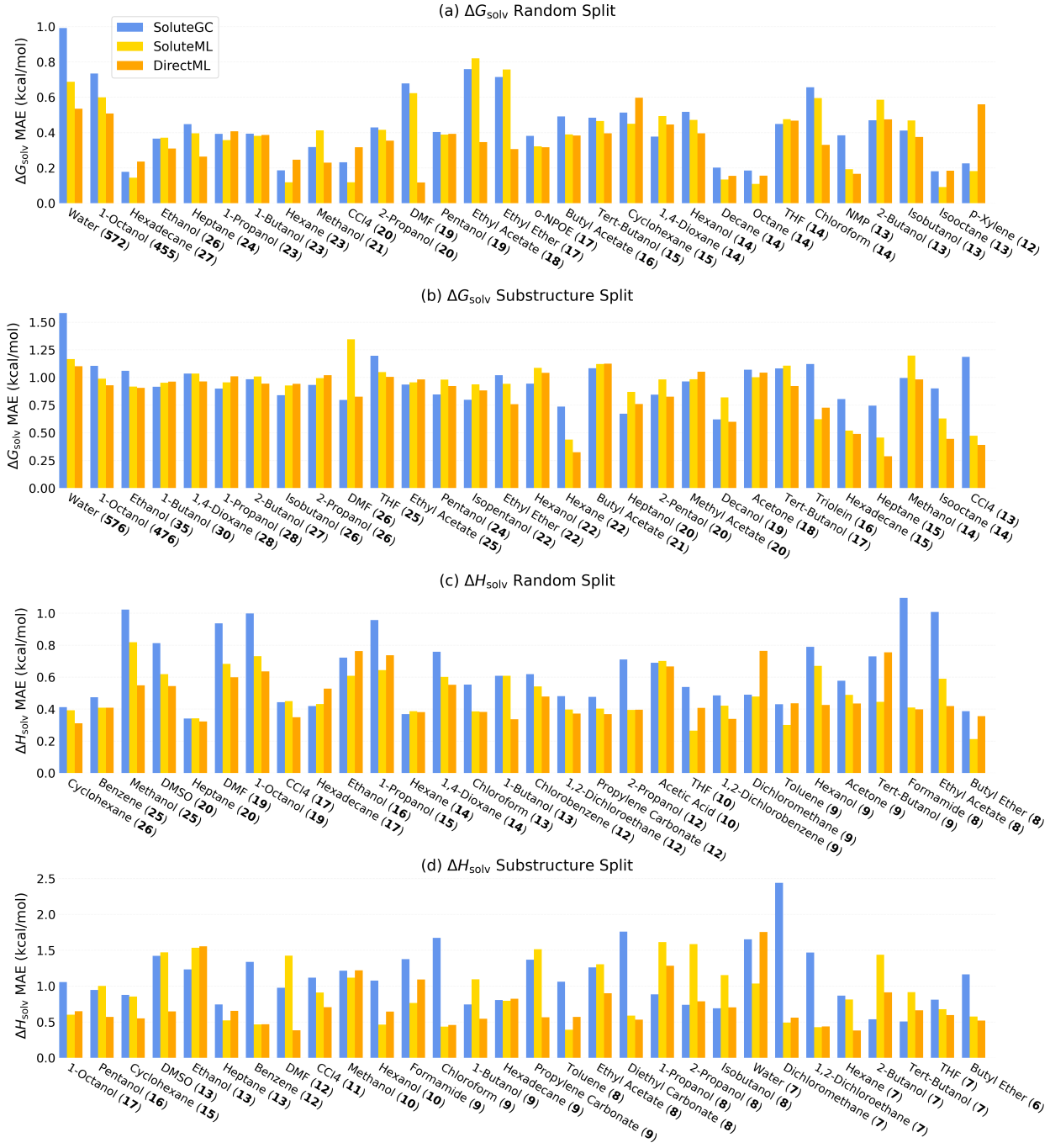

Figure 7: The MAE for the prediction of $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$ on the 30 most frequently appearing solvents in each test set. The performance of the different models is compared only using the overlapping test data. The number of solvent data in each test set is given in parenthesis in bold. All solvents are included in the training sets of the DirectML models. (CCl4: carbon tetrachloride, DMF: dimethylformamide, o-NPOE: 2-nitrophenyl octyl ether, THF: tetrahydrofuran, NMP: N-methyl-2-pyrrolidone, DMSO: dimethyl sulfoxide)

random split, but the SoluteML and DirectML models have similar prediction accuracy for the same solvents in the substructure split. The three models have different levels of accuracy for many solvents for the $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ prediction in the substructure split compared to the $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ prediction.

Overall, the DirectML model outperforms the SoluteML model despite having fewer solute data. The SoluteGC model gives the highest error but also contains the most useful information that relates physical contributions to the solvation properties to chemical substructures in the solutes. The underperformance of the SoluteML and the SoluteGC models are most likely related to the approximation of using the linear relationships for the calculation of energy-related solvation properties. For the SoluteGC model, an additional error comes from limiting most of the groups to the nearest neighboring atom interaction. The graph convolutional neural networks in the ML methods, on the other hand, allow the information of each atom to propagate into further-distanced atoms within the molecule and hence can include the non-nearest atom interaction.

Yet, as can be concluded from Figures 6 and 7, there is not one model that outperforms the others on all substructures and all solvents. Based on the observation that the three methods have different levels of accuracy for various solute substructures and solvents and have different outliers, we expect that using the average predictions of the different models may be able to give better predictions by suppressing the large errors from outliers. To test this, we compare the average predictions of the SoluteML and DirectML models (2 model average) and of the SoluteGC, SoluteML, and DirectML models (3 model average) with the experimental data. The resulting parity plots and error summary are shown in Figure 8 and Table 5. The average predictions indeed lead to slight improvements for most test sets compared to the single predictions from the DirectML model. More significant improvements are observed for some of the main outliers of the DirectML model where the other models had better performance.

In summary, even though the DirectML model performs better in general, the three models can offer different levels of information on the predicted values, and, when combined together, they can provide even more accurate predictions of $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$. The three models as a whole can be used to identify errors or outliers in any of the models. However, users of these models should still be cautious even when using the average values of the three models since it is possible that the particular solute-solvent pair of greatest interest to a user would be one that is significantly mis-predicted by all three models. We also caution that while the models work very well overall, this is not a guarantee that all the model parameters have been well-determined. In particular for the SoluteGC model, as is typical with large linear models, some linear combinations of group contribution values may have not been very well determined or tested by our data set. Therefore, it would be inadvisable to combine these group values with other group values determined some other way, without validating the newly formed combined model on independent data.

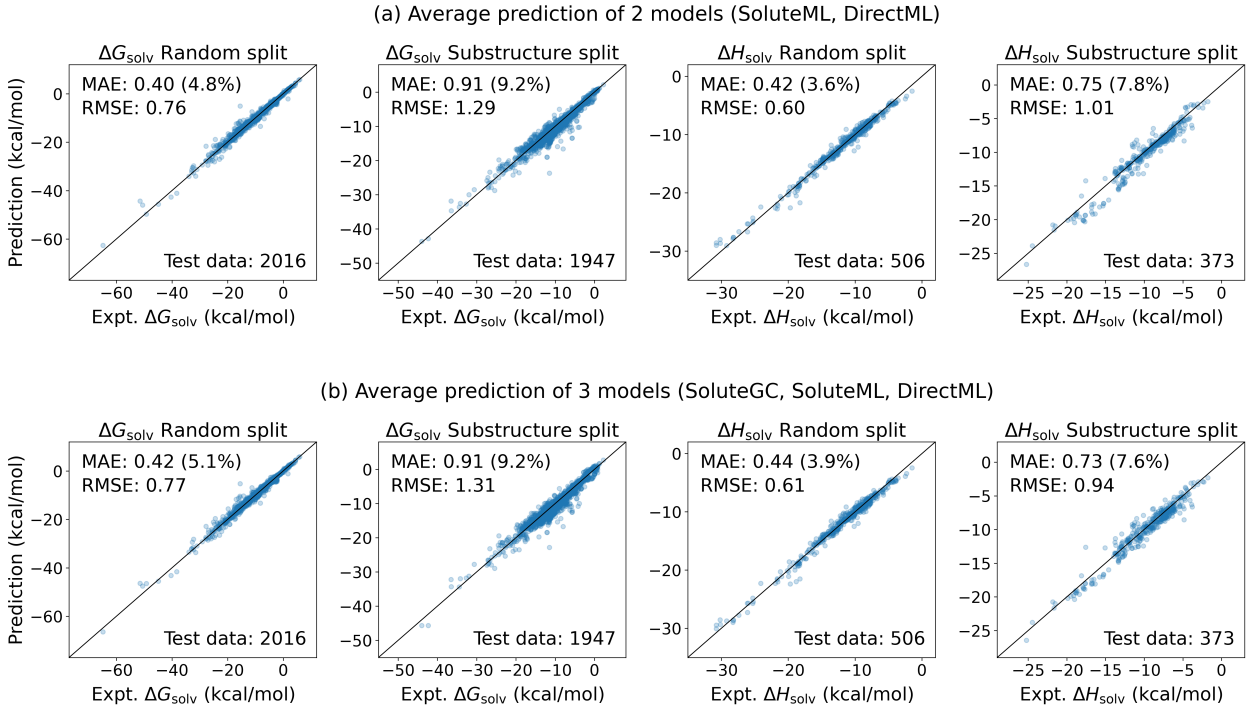

Figure 8: Parity plots between the experimental data and average predictions of the two different models (SoluteML, DirectML) and the three different models (SoluteGC, SoluteML, DirectML) on the overlapping test data. The parity plots are for $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$ predictions tested on pairs of unseen solutes and trained solvents for the random and substructure splits. The MAE and RMSE are in kcal/mol, and the PMAE is given in parenthesis. The number of test data are also given on the figures along with the test errors. The plots only show the overlapping test data in each split type, and therefore the test data are identical to those in Figures 4 and 5.

Table 5: Test set error summary of the SoluteGC, SoluteML, and DirectML models and the average predictions of multiple models on unseen solute and trained solvent pairs. "Avg. 2Models" represents the average predictions of the SoluteML and DirectML models, and "Avg. 3Models" represents the average predictions of the SoluteGC, SoluteML, and DirectML models. The MAE and RMSE are reported in kcal/mol. The lowest errors in each test set are marked in bold. The table shows the results on the overlapping test data, and therefore the test data are identical to those in Figures 4 and 5.

Model	ΔG_{solv} Random		ΔG_{solv} Substr.		ΔH_{solv} Random		ΔH_{solv} Substr.	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
SoluteGC	0.63	1.06	1.18	1.66	0.64	0.88	1.13	1.53
SoluteML	0.48	0.95	1.00	1.43	0.50	0.69	0.93	1.37
DirectML	0.41	**0.75**	**0.91**	1.32	0.47	0.71	**0.73**	1.07
Avg. 2Models	**0.40**	0.76	**0.91**	**1.29**	**0.42**	**0.60**	0.75	1.01
Avg. 3Models	0.42	0.77	**0.91**	1.31	0.44	0.61	**0.73**	**0.94**

4.2.2 Performance on Unseen Solutes and Unseen Solvents

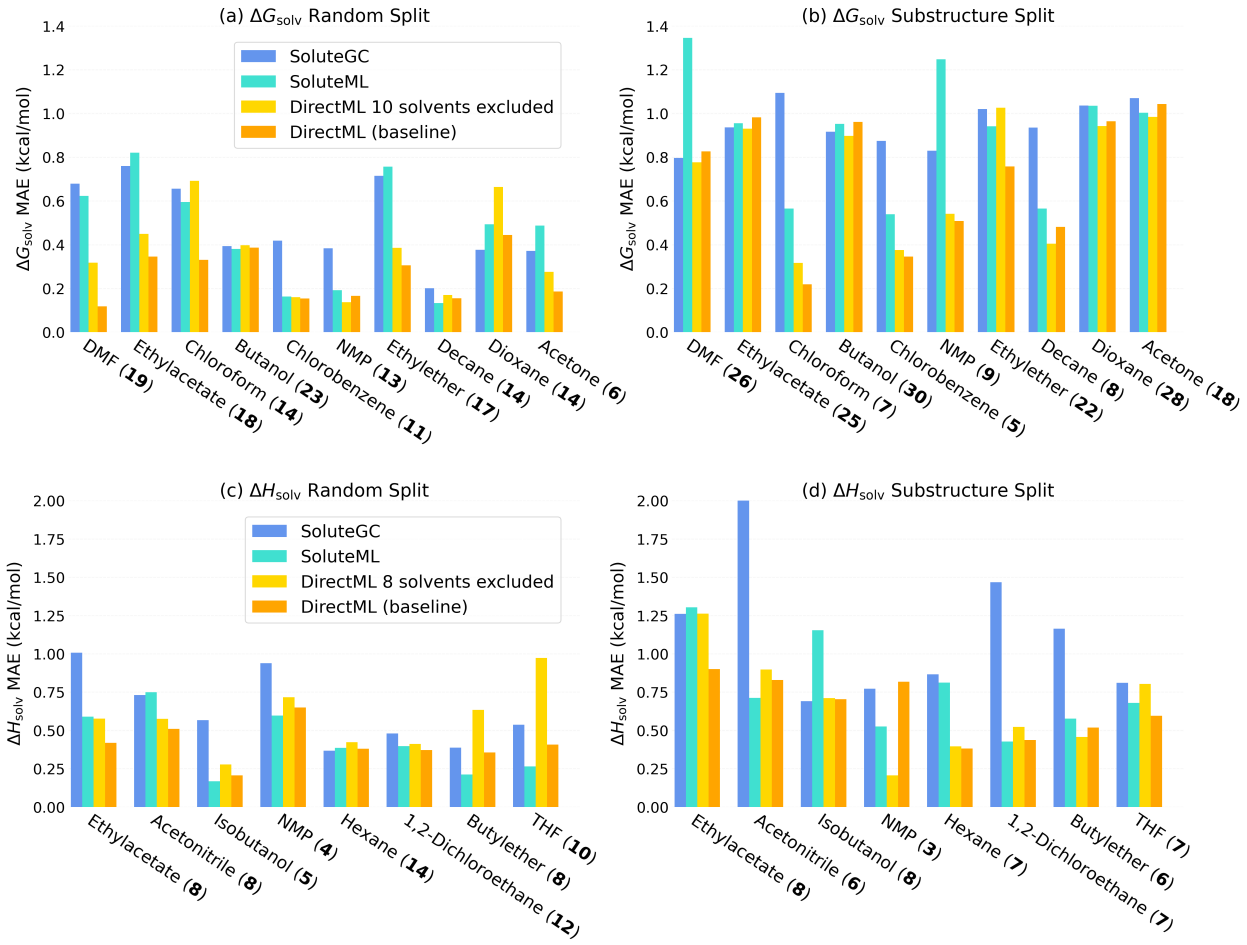

Figure 9: The MAE for the prediction of $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$ on the solvents excluded from the DirectML training and validation sets. The performance of the different models is compared using only the overlapping test data. The number of solvent data in the test set is given in parenthesis in bold. (DMF: dimethylformamide, NMP: N-methyl-2-pyrrolidone, THF: tetrahydrofuran)

To test the predictive performance of the DirectML model on pairs of unseen solutes and unseen solvents, 10 solvents and 8 solvents are excluded from the training and validation sets of dGsolvDB3 and dHsolvDB, respectively. New DirectML models are trained and validated on these reduced data sets, and their performance on the excluded solvents are compared with the baseline DirectML models trained using all solvents. The SoluteGC and SoluteML models from the earlier section are also compared to see whether the new DirectML models can still outperform them on the unseen solvents. The results on the random and substructure splits for $\Delta G_{\text{solv}}(298\,\text{K})$ and $\Delta H_{\text{solv}}(298\,\text{K})$ are presented in Figure 9.

For the majority of the out-of-sample solvents for $\Delta G_{\text{solv}}(298\,\text{K})$ prediction, the DirectML models trained on the reduced data set still outperform the SoluteGC and SoluteML models.

Compared to the baseline case, the new DirectML model has on average a higher error for the random split as expected, but achieves comparable or even better performance on some solvents for the substructure split. Most likely, this is because the prediction error of the substructure split is predominantly caused by the lack of solute substructures in the training set rather than missing solvent information. Moreover, the training set still includes solvents with similar structures to the excluded solvents. For example, even though decane is excluded, the training set still contains a series of other alkanes. Hence, for the substructure split, it is expected that the DirectML model still provides a similar performance even if solvents are excluded from the training set.

Similar results are observed for the prediction of $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$; the DirectML model has higher errors than the baseline model for all out-of-sample solvents in the random split and has similar or even lower errors for some solvents in the substructure split. Even though the DirectML model for the prediction of $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ is trained using much fewer data points (142 solvents when not including the solvents that are only present in self-solvation data) compared to the $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ model trained on 303 solvents, the error increase for the excluded solvents is on average similar to that of the $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ prediction.

Overall, the prediction error of the DirectML on out-of-sample solvents increases in the random splits on average by 0.1 - 0.2 kcal/mol and by 0.4 - 0.5 kcal/mol at most. On the contrary, a lack of solvent information does not affect the predictive performance of the DirectML model much in the substructure splits where the error is primarily caused by a lack of solute substructures in the training sets. The DirectML model can still provide better or similar predictions for most of the out-of-sample solvents compared to the SoluteML model, which uses empirical solvent parameters. However, caution should be made when applying the DirectML model to a solvent with a foreign structure or characteristic that is vastly different from the solvents in the training set as the error is usually much larger for a compound with a unique substructure.

4.3 Comparison to Existing Methods

Finally, the prediction errors of our models on unseen solutes that are presented in Section 4.2 are summarized in Table 6. In the table, the prediction errors of our models are reported using the entire 10 % test set data of each model while smaller common test sets were used in the previous sections for model comparison. All DirectML models in the table are trained, validated, and tested using the dGsolvDB3 data set for the solvation free energy prediction.

The prediction errors of the selected existing models from Section 3.3 are also included in Table 6. Lim and Jung[43] reported results on multiple split types for their MLSolvA model, but only the results on solute clustering are included in the table as this is most similar to the random and substructure-based solute splits used throughout our work. Similarly, for

the transfer learning model by Vermeire and Green,[44] the results on element-based solute splits are used for comparison. To the best of our knowledge, there is no ML model that can predict solvation enthalpies in a variety of solvents. One model reported by Jaquis et al.[102] is used for comparison, but it only considers ethanol as a solvent. Note that the size and the nature of the training and test sets used for each method varies, and therefore the relative ranking of these methods could change when evaluated on different data sets. The parity plots comparing the experimental data with the in-house calculations using the existing methods, namely the COSMO-RS and the UFZ-LSER GC methods, are provided in the Supporting Information Sections 4.3.1 and 4.3.2.

For both $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ predictions, our ML models achieve a similar or even better accuracy than the QM methods on random split but have higher errors on the substructure split. MLSolvA gives lower error than the DirectML and SoluteML models for the solute substructure splits. Nevertheless, it is difficult to draw a solid conclusion without knowing what substructures are found in MLSolvA's solute clustering. The transfer learning model by Vermeire and Green[44] yields relatively low error for the element-based solute splits, which are considered more challenging than the substructure-based splits. Their transfer learning model, however, is pre-trained on 1 million quantum calculations while our models are limited to the use of experimental data. For the solvation enthalpy prediction, our models are more accurate than the ML model by Jaquis et al.,[102] even though their model is limited to ethanol as a solvent.

The performance of the SoluteGC model on random solute split is similar to or better than that of the UFZ-LSER GC model evaluated on all solvation free energy and enthalpy data. Nevertheless, the reported errors of the UFZ-LSER in Table 6 are not true test errors since the training set of the UFZ-LSER is unknown. A considerable portion of its training set solutes are expected to overlap with the test set solutes used in the work, and hence we are unable to test the performance of the UFZ-LSER model on any solute splits.

Because of the paucity of experimental data, often all or nearly all data are used for training with no or a very small independent test set as it is done for our final models. The resulting final DirectML and SoluteGC models achieve very low training and testing errors similar to the expected experimental error bars as it can be seen from Table S7 in the Supporting Information. The final SoluteGC model has larger training errors, likely because the true solvation energies are not exactly linear in the functional groups. All of our final models can be accessed through various ways described in Section 6.

Table 6: Comparison of various QM, ML, and GC methods for solvation energy and enthalpy prediction at 298 K. The MAE and RMSE are in kcal/mol, and the number of each test set data (N_{test}) is presented. Our model results are written in bold. The results here for each of our models use its own full test set. In the comparisons in the text, a smaller common test set was used for all the models.

target	method	method type	test set split type	MAE	RMSE	N_{test}	Ref.
ΔG_{solv}	SMD/IEF-PCM/ M05-2X/6-31G*	QM	-	0.63	0.86	2346	[9] [a]
	COSMO-RS/ BP86/TZVPD-FINE	QM	-	0.46	0.77	14236	[b]
	MLSolvA	ML	K-mean solute cluster CV using scaffold-based split	0.62	1.15	6239	[43]
	Transfer learning	ML	element-based solute split (O excluded)	0.52	0.91	4684	[44]
	Transfer learning	ML	element-based solute split (Cl excluded)	0.45	0.63	1124	[44]
	DirectML	**ML**	**random solute**	**0.40**	**0.73**	**2138**	
	DirectML	**ML**	**substructure solute**	**0.89**	**1.32**	**2033**	
	DirectML	**ML**	**random solute (10 solvents excluded)**	**0.40**	**0.75**	**2138**	[c]
	DirectML	**ML**	**substructure solute (10 solvents excluded)**	**0.89**	**1.32**	**2033**	[c]
	SoluteML	**ML**	**random solute**	**0.48**	**0.95**	**2016**	
	SoluteML	**ML**	**substructure solute**	**1.01**	**1.45**	**1948**	
	UFZ-LSER	GC	evaluated on all ΔG_{solv} data [d]	0.78	1.42	16878	[e]
	SoluteGC	**GC**	**random solute**	**0.63**	**1.06**	**2016**	
	SoluteGC	**GC**	**substructure solute**	**1.18**	**1.66**	**1947**	
ΔH_{solv}	COSMO-RS/ BP86/TZVPD-FINE	QM	-	0.69	1.06	6058	[b]
	Jaquis et al. (only solvent ethanol)	ML	random solute	-	1.58	35	[102]
	DirectML	**ML**	**random solute**	**0.50**	**0.80**	**643**	
	DirectML	**ML**	**substructure solute**	**0.78**	**1.13**	**551**	
	DirectML	**ML**	**random solute (8 solvents excluded)**	**0.51**	**0.80**	**643**	[c]
	DirectML	**ML**	**substructure solute (8 solvents excluded)**	**0.78**	**1.12**	**551**	[c]
	SoluteML	**ML**	**random solute**	**0.50**	**0.69**	**506**	
	SoluteML	**ML**	**substructure solute**	**0.94**	**1.38**	**373**	
	UFZ-LSER	GC	evaluated on all ΔH_{solv} data [d]	0.59	0.98	4099	[e]
	SoluteGC	**GC**	**random solute**	**0.64**	**0.88**	**506**	
	SoluteGC	**GC**	**substructure solute**	**1.13**	**1.53**	**373**	

[a] The original authors used IEF-PCM protocol implemented in Gaussian03 for the SMD calculations. The errors were reported separately for neutral solutes in aqueous and non-aqueous solvents in the original paper, and we calculated the errors for all solvents based on the reported errors and number of data points. [b] The COSMO-RS calculations are performed in this work using COSMO*therm*[103] and COSMO*base*.[104] [c] These refer to the DirectML models from Section 4.2.2 for which 10 and 8 solvents are excluded from the training

and validation sets of dGsolvDB3 and dHsolvDB, respectively. [d] Note that these are not true test errors since the training set of the UFZ-LSER is unknown and its training set solutes are expected overlap with the test set solutes used in these error calculations. [e] The UFZ-LSER GC calculations are done in this work through the UFZ-LSER database[29] using in-house solvent parameters. The UFZ-LSER calculation is only available for molecular weight less than 1000 g/mol, and therefore, a few solute compounds could not be evaluated for ΔG_{solv} predictions.

5 Conclusions

A group contribution method (SoluteGC) and a machine learning model (SoluteML) are constructed for the Abraham solute parameter prediction that are used to estimate solvation free energy and enthalpy via the LSERs. Additionally, a machine learning model (DirectML) is developed for the direct prediction of the solvation free energy and enthalpy for a given solvent-solute pair. The predictive performance of the three models is evaluated on common test sets of the solvation free energy and enthalpy for out-of-sample solute compounds prepared using a random and substructure-based split. The results show that the DirectML model is superior to the SoluteGC and SoluteML models for both solvation energy and enthalpy predictions on all data splits. The SoluteML model performs similarly with the mean absolute errors (MAEs) around 0.1 kcal/mol higher than those of the DirectML. The SoluteGC model underperforms with the MAEs about 0.2 - 0.3 kcal/mol higher than those of the DirectML.

It is also found that adding $\log P$ data and a substantial amount of $\log K_{\mathrm{w}}$ data to the solvation free energy data set improves the performance of the DirectML model mainly for the substructure solute split. Although including these data introduces additional data uncertainty and causes the solvent data to be biased toward water and 1-octanol, the information gained from the additional data can compensate for these drawbacks and gives overall better results.

The present models and some other recently developed models[43,44] trained with large data sets are all accurate enough that they give $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ predictions close to the aleatoric uncertainties for random-split test data sets. Each model's predictions for solutes and solvents that are very different from those in its training data are less reliable, and hence an averaging or consensus-of-models approach is recommended. Here we provide a convenient set of 3 rather different models for $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and also provide 3 models for $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$. Together these models provide good inputs for predicting the solvation thermodynamics of a very large range of solutes and solvents at temperatures up to each solvent's critical point.[5]

Finally, we present our compiled solute parameter (SoluteDB), solvation free energy (dGsolvDBx), and solvation enthalpy (dHsolvDB) databases and provide public access to our

final prediction models through a simple web-based tool, software package, and GitHub. A web-based tool is designed for quick calculations and others are more suited for bulk, automated calculations.

6 Data and Software Availability

The in-house Abraham solute parameter database (SoluteDB) along with the in-house $\log K_{\mathrm{w}}$ and $\log P_{\mathrm{ow}}$ data are provided as a part of the Supplementary Information. The publicly available $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ data are also provided in the Supplementary Information. This excludes the $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$ data from the Minnesota solvation database, which is not open-source. For each entry of a solute-solvent pair, the list of all individual data points, mean value, and standard deviation are tabulated in our databases. Although we limit our work to the solute compounds containing H, C, N, O, S, P, F, Cl, Br, or I atoms, the open-source solute parameter, $\Delta G_{\mathrm{solv}}(298\,\mathrm{K})$, and $\Delta H_{\mathrm{solv}}(298\,\mathrm{K})$ data that contain the elements out of our scope are also provided as Supplementary Information.

The final SoluteGC, SoluteML, and DirectML models that are constructed using all data are made publicly available through GitHub, conda software package, and web-based search tool. The web-based tool is available on `https://rmg.mit.edu/database/solvation/search/` and is the simplest way to search for solute parameters, solvation free energies, and solvation enthalpies. Temperature-dependent ΔG_{solv} calculation based on our earlier work[5] is also available through the web-based tool for the solvents whose temperature-dependent densities can be computed by a free fluid modeling software CoolProp.[108] For bulk queries, one can download the source code using GitHub or install a conda package. The SoluteGC model can be accessed by installing the source code from RMG-Py and RMG-database git repositories (`https://github.com/ReactionMechanismGenerator`) with a sample code located at `https://github.com/ReactionMechanismGenerator/RMG-Py/blob/master/ipython/estimate_solvation_thermo_and_search_available_solvents.ipynb`. The SoluteML and DirectML models can be downloaded as a conda package from `https://anaconda.org/fhvermei/chemprop_solvation`. The source code for the ML models can be found from chemprop_solvation git repository (`https://github.com/fhvermei/chemprop_solvation`).

Acknowledgement

We gratefully acknowledge Eni S.p.A. for supporting this research. We also acknowledge additional financial support from the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS). Yunsie Chung acknowledges financial support from National Science Foundation (subawarded by the MolSSI under fellowship agreement number 479590-19825A). Florence Vermeire acknowledges financial support from the Belgian American Educational Foundation (BAEF). We thank Jonathan Zheng for his help with revising

the code for the web-based tool.

References

(1) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. Predicting aqueous solubilities from aqueous free energies of solvation and experimental or calculated vapor pressures of pure substances. *J. Chem. Phys.* **2003**, *119*, 1661–1670.

(2) Li, C.-J.; Chan, T.-H. *Comprehensive Organic Reactions in Aqueous Media*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007.

(3) Sixt, M.; Koudous, I.; Strube, J. Process design for integration of extraction, purification and formulation with alternative solvent concepts. *C. R. Chim.* **2016**, *19*, 733–748.

(4) Ruelle, P. The n-octanol and n-hexane/water partition coefficient of environmentally relevant chemicals predicted from the mobile order and disorder (MOD) thermodynamics. *Chemosphere* **2000**, *40*, 457–512.

(5) Chung, Y.; Gillis, R. J.; Green, W. H. Temperature-dependent vapor–liquid equilibria and solvation free energy estimation from minimal data. *AIChE J.* **2020**, *66*, e16976.

(6) Hildebrand, J. H. A History of Solution Theory. *Annu. Rev. Phys. Chem.* **1981**, *32*, 1–24.

(7) Cramer, C. J.; Truhlar, D. G. Molecular Orbital Theory Calculations of Aqueous Solvation Effects on Chemical Equilibria. *J. Am. Chem. Soc.* **1991**, *113*, 8552–8554.

(8) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges. *J. Chem. Theory Comput.* **2007**, *3*, 2011–2033.

(9) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(10) Klamt, A. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(11) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(12) Klamt, A.; Eckert, F. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilib.* **2000**, *172*, 43–72.

(13) Abraham, M. H.; Ibrahim, A.; Zissimos, A. M. Determination of sets of solute descriptors from chromatographic measurements. *J. Chromatogr. A* **2004**, *1037*, 29–47.

(14) Abraham, M. H.; Acree, W. E. Correlation and prediction of partition coefficients between the gas phase and water, and the solvents dodecane and undecane. *New J. Chem.* **2004**, *28*, 1538–1543.

(15) Kamlet, M. J.; Taft, R. W. The solvatochromic comparison method. I. The .beta.-scale of solvent hydrogen-bond acceptor (HBA) basicities. *J. Am. Chem. Soc.* **1976**, *98*, 377–383.

(16) Mintz, C.; Clark, M.; Acree, W. E.; Abraham, M. H. Enthalpy of Solvation Correlations for Gaseous Solutes Dissolved in Water and in 1-Octanol Based on the Abraham Model. *J. Chem. Inf. Model.* **2007**, *47*, 115–121.

(17) Jalan, A.; Ashcraft, R. W.; West, R. H.; Green, W. H. Predicting solvation energies for kinetic modeling. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2010**, *106*, 211–258.

(18) Bradley, J. C.; Abraham, M. H.; Acree, W. E.; Lang, A. S. Predicting Abraham model solvent coefficients. *Chem. Cent. J.* **2015**, *9*, 1–10.

(19) Abraham, M. H.; Acree, W. E. Estimation of heat capacities of gases, liquids and solids, and heat capacities of vaporization and of sublimation of organic chemicals at 298.15 K. *J. Mol. Liq.* **2020**, *317*, 113969.

(20) Abraham, M. H.; Acree, W. E. Estimation of enthalpies of sublimation of organic, organometallic and inorganic compounds. *Fluid Phase Equilib.* **2020**, *515*, 112575.

(21) Snelgrove, D. W.; Lusztyk, J.; Banks, J. T.; Mulder, P.; Ingold, K. U. Kinetic solvent effects on hydrogen-atom abstractions: Reliable, quantitative predictions via a single empirical equation. *J. Am. Chem. Soc.* **2001**, *123*, 469–477.

(22) Havelec, P.; Ševčík, J. G. Extended additivity model of parameter log(L16). *J. Phys. Chem. Ref. Data* **1996**, *25*, 1483–1493.

(23) Svozil, D.; Ševčík, J. G.; Kvasnička, V. Neural network prediction of the solvatochromic polarity/polarizability parameter πH2. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 338–342.

(24) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of molecular linear free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.

(25) Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. Estimation of Molecular Linear Free Energy Relationship Descriptors by a Group Contribution Approach. 2. Prediction of Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71–80.

(26) GHAFOURIAN, T.; DEARDEN, J. C. The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding-donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods. *J. Pharm. Pharmacol.* **2000**, *52*, 603–610.

(27) Zissimos, A. M.; Abraham, M. H.; Klamt, A.; Eckert, F.; Wood, J. A comparison between the two general sets of linear free energy descriptors of Abraham and Klamt. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1320–1331.

(28) Arey, J. S.; Green, W. H.; Gschwend, P. M. The electrostatic origin of Abraham's solute polarity parameter. *J. Phys. Chem. B* **2005**, *109*, 7564–7573.

(29) Ulrich, N.; Endo, S.; Brown, T.; Watanabe, N.; Bronner, G.; Abraham, M.; Goss, K.-U. UFZ-LSER database v 3.2 [Internet]. 2017; `http://www.ufz.de/lserd`.

(30) Liang, Y.; Torralba-Sanchez, T. L.; Di Toro, D. M. Estimating system parameters for solvent-water and plant cuticle-water using quantum chemically estimated Abraham solute parameters. *Environ. Sci.: Process. Impacts* **2018**, *20*, 813–821.

(31) ACD/Labs, ACD/ADME Suite 5.0, version 12.0. `https://www.acdlabs.com/`.

(32) Stenzel, A.; Goss, K. U.; Endo, S. Prediction of partition coefficients for complex environmental contaminants: Validation of COSMOtherm, ABSOLV, and SPARC. *Environ. Toxicol. Chem.* **2014**, *33*, 1537–1543.

(33) Brown, T. N.; Arnot, J. A.; Wania, F. Iterative fragment selection: A group contribution approach to predicting fish biotransformation half-lives. *Environ. Sci. Technol.* **2012**, *46*, 8253–8260.

(34) Brown, T. N. Predicting hexadecane-air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data. *SAR QSAR Environ. Res.* **2014**, *25*, 51–71.

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

(36) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.

(37) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(38) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database – version 2012, University of Minnesota, Minneapolis. 2012.

(39) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1338–1346.

(40) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(41) Hille, C.; Ringe, S.; Deimel, M.; Kunkel, C.; Acree, W. E.; Reuter, K.; Oberhofer, H. Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme. *J. Chem. Phys.* **2019**, *150*, 041710.

(42) Pathak, Y.; Mehta, S.; Priyakumar, U. D. Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks. *J. Chem. Inf. Model.* **2021**, *61*, 689–698.

(43) Lim, H.; Jung, Y. J. MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning. *J. Cheminformatics* **2021**, *13*, 1–10.

(44) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.

(45) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(46) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.

(47) Swain, M. PubChemPy. 2013; `https://pypi.org/project/PubChemPy/1.0/`.

(48) Swain, M. CIRpy. 2016; `https://pypi.org/project/CIRpy/`.

(49) Landrum, G. RDKit: Open-Source Cheminformatics. 2006; `http://www.rdkit.org/`.

(50) Grinberg Dana, A.; Ranasinghe, D.; Wu, H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. H. ARC - Automated Rate Calculator, version 1.1.0. https://github.com/ReactionMechanismGenerator/ARC, 2019.

(51) Moine, E.; Privat, R.; Sirjean, B.; Jaubert, J. N. Estimation of solvation quantities from experimental thermodynamic data: Development of the comprehensive CompSol databank for pure and mixed solutes. *J. Phys. Chem. Ref. Data* **2017**, *46*.

(52) Abraham, M. H.; Zissimos, A. M.; Acree, W. E., Jr. Partition of solutes from the gas phase and from water to wet and dry di-n-butyl ether: a linear free energy relationship analysis. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3732–3736.

(53) Abraham, M. H.; Zissimos, A. M.; Acree, W. E., Jr. Partition of solutes into wet and dry ethers; an LFER analysis. *New J. Chem.* **2003**, *27*, 1041–1044.

(54) Abraham, M. H.; Acree, W. E., Jr. Comparative analysis of solvation and selectivity in room temperature ionic liquids using the Abraham linear free energy relationship. *Green Chem.* **2006**, *8*, 906–915.

(55) Acree William E, J.; Abraham, M. H. The analysis of solvation in ionic liquids and organic solvents using the Abraham linear free energy relationship. *J. Chem. Technol. Biotechnol.* **2006**, *81*, 1441–1446.

(56) Abraham, M. H.; Enomoto, K.; Clarke, E. D.; Rosés, M.; Ràfols, C.; Fuguet, E. Henry's Law constants or air to water partition coefficients for 1,3,5-triazines by an LFER method. *J. Environ. Monit.* **2007**, *9*, 234–239.

(57) Sprunger, L. M.; Proctor, A.; Acree, W. E.; Abraham, M. H.; Benjelloun-Dakhama, N. Correlation and prediction of partition coefficient between the gas phase and water, and the solvents dry methyl acetate, dry and wet ethyl acetate, and dry and wet butyl acetate. *Fluid Phase Equilib.* **2008**, *270*, 30–44.

(58) Sprunger, L. M.; Gibbs, J.; Acree, W. E.; Abraham, M. H. Correlation and prediction of partition coefficients for solute transfer to 1,2-dichloroethane from both water and from the gas phase. *Fluid Phase Equilib.* **2008**, *273*, 78–86.

(59) Sprunger, L. M.; Achi, S. S.; Pointer, R.; Blake-Taylor, B. H.; Acree, W. E.; Abraham, M. H. Development of Abraham model correlations for solvation characteristics of linear alcohols. *Fluid Phase Equilib.* **2009**, *286*, 170–174.

(60) Sprunger, L. M.; Achi, S. S.; Acree, W. E.; Abraham, M. H.; Leo, A. J.; Hoekman, D. Correlation and prediction of solute transfer to chloroalkanes from both water and the gas phase. *Fluid Phase Equilib.* **2009**, *281*, 144–162.

(61) Abraham, M. H.; Acree, W. E.; Leo, A. J.; Hoekman, D. The partition of compounds from water and from air into wet and dry ketones. *New J. Chem.* **2009**, *33*, 568–573.

(62) Abraham, M. H.; Acree Jr, W. E.; Cometto-Muñiz, J. E. Partition of compounds from water and from air into amides. *New J. Chem.* **2009**, *33*, 2034–2043.

(63) Abraham, M. H.; Acree, W. E., Jr.; Leo, A. J.; Hoekman, D. Partition of compounds from water and from air into the wet and dry monohalobenzenes. *New J. Chem.* **2009**, *33*, 1685–1692.

(64) Sprunger, L. M.; Achi, S. S.; Pointer, R.; Acree, W. E.; Abraham, M. H. Development of Abraham model correlations for solvation characteristics of secondary and branched alcohols. *Fluid Phase Equilib.* **2010**, *288*, 121–127.

(65) Grubbs, L. M.; Saifullah, M.; De La Rosa, N. E.; Ye, S.; Achi, S. S.; Acree, W. E.; Abraham, M. H. Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents. *Fluid Phase Equilib.* **2010**, *298*, 48–53.

(66) Abraham, M. H.; Smith, R. E.; Luchtefeld, R.; Boorem, A. J.; Luo, R.; Acree, W. E. Prediction of Solubility of Drugs and Other Compounds in Organic Solvents. *J. Pharm. Sci.* **2010**, *99*, 1500–1515.

(67) Ye, S.; Saifullah, M.; Grubbs, L. M.; McMillan-Wiggins, M. C.; Acosta, P.; Mejorado, D.; Flores, I.; Acree, W. E.; Abraham, M. H. Determination of the Abraham model solute descriptors for 3,5-dinitro-2-methylbenzoic acid from measured solubility data in organic solvents. *Phys. Chem. Liq.* **2011**, *49*, 821–829.

(68) Stephens, T. W.; De La Rosa, N. E.; Saifullah, M.; Ye, S.; Chou, V.; Quay, A. N.; Acree, W. E.; Abraham, M. H. Abraham model correlations for solute partitioning into o-xylene, m-xylene and p-xylene from both water and the gas phase. *Fluid Phase Equilib.* **2011**, *308*, 64–71.

(69) Saifullah, M.; Ye, S.; Grubbs, L. M.; De La Rosa, N. E.; Acree, W. E.; Abraham, M. H. Abraham Model Correlations for Transfer of Neutral Molecules to Tetrahydrofuran and to 1,4-Dioxane, and for Transfer of Ions to Tetrahydrofuran. *J. Solution Chem.* **2011**, *40*, 2082–2094.

(70) Abraham, M. H.; Acree, W. E. The transfer of neutral molecules, ions and ionic species from water to benzonitrile; comparison with nitrobenzene. *Thermochim. Acta* **2011**, *526*, 22–28.

(71) Holley, K.; Acree, W. E.; Abraham, M. H. Determination of Abraham model solute descriptors for 2-ethylanthraquinone based on measured solubility ratios. *Phys. Chem. Liq.* **2011**, *49*, 355–365.

(72) Stephens, T. W.; Wilson, A.; Dabadge, N.; Tian, A.; Zimmerman, M.; Hensley, H. J.; Acree, W. E. J.; Abraham, M. H. Correlation of solute partitioning into isooctane from water and from the gas phase based on updated Abraham equations. *Global J. Phys. Chem.* **2012**, *3*, 1–16.

(73) Stephens, T. W.; Quay, A. N.; Chou, V.; Loera, M.; Shen, C.; Wilson, A.; Acree, W. E.; Abraham, M. H. Correlation of solute transfer into alkane solvents from water and from the gas phase with updated Abraham model equations. 2012; `https://digital.library.unt.edu/ark:/67531/metadc152452/`.

(74) Stephens, T. W.; Loera, M.; Calderas, M.; Diaz, R.; Montney, N.; Acree, W. E.; Abraham, M. H. Determination of Abraham model solute descriptors for benzoin based on measured solubility ratios. *Phys. Chem. Liq.* **2012**, *50*, 254–265.

(75) Abraham, M. H.; Acree, W. E. Linear free-energy relationships for water/hexadec-1-ene and water/deca-1,9-diene partitions, and for permeation through lipid bilayers; comparison of permeation systems. *New J. Chem.* **2012**, *36*, 1798–1806.

(76) Abraham, M. H.; Gola, J. R. M.; Gil-Lostes, J.; Acree, W. E.; Cometto-Muñiz, J. E. Determination of solvation descriptors for terpene hydrocarbons from chromatographic measurements. *J. Chromatogr. A* **2013**, *1293*, 133–141.

(77) Abraham, M. H.; Acree, W. E. Physicochemical and biochemical properties for the dialkyl phthalates. *Chemosphere* **2015**, *119*, 871–880.

(78) Bradley, J.-C.; Abraham, M. H.; Acree, W. E.; Lang, A. S. I. D.; Beck, S. N.; Bulger, D. A.; Clark, E. A.; Condron, L. N.; Costa, S. T.; Curtin, E. M.; Kurtu, S. B.; Mangir, M. I.; McBride, M. J. Determination of Abraham model solute descriptors for the monomeric and dimeric forms of trans-cinnamic acid using measured solubilities from the Open Notebook Science Challenge. *Chem. Cent. J.* **2015**, *9*, 11.

(79) Brumfield, M.; Wadawadigi, A.; Kuprasertkul, N.; Mehta, S.; Stephens, T. W.; Barrera, M.; De La Rosa, J.; Kennemer, L.; Meza, J.; Acree, W. E.; Abraham, M. H. Determination of Abraham model solute descriptors for three dichloronitrobenzenes from measured solubilities in organic solvents. *Phys. Chem. Liq.* **2015**, *53*, 163–173.

(80) Abraham, M. H.; Acree, W. E. Equations for water-triolein partition coefficients for neutral species; comparison with other water-solvent partitions, and environmental and toxicological processes. *Chemosphere* **2016**, *154*, 48–54.

(81) Abraham, M. H.; Acree, W. E. Descriptors for Pentane-2,4-dione and Its Derivatives. *J. Solution Chem.* **2017**, *46*, 1625–1638.

(82) Abraham, M. H.; Acree, W. E.; Liu, X. Partition of Neutral Molecules and Ions from Water to o-Nitrophenyl Octyl Ether and of Neutral Molecules from the Gas Phase to o-Nitrophenyl Octyl Ether. *J. Solution Chem.* **2018**, *47*, 293–307.

(83) Abraham, M. H.; Acree, W. E.; Cometto-Muñiz, J. E. Descriptors for terpene esters from chromatographic and partition measurements: Estimation of human odor detection thresholds. *J. Chromatogr. A* **2020**, *1609*.

(84) Sushko, I. et al. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554.

(85) Wishart, D. S. et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(86) Syracuse Research Corporation. Physical/Chemical Property Database (PHYSPROP); SRC Environmental Science Center: Syracuse, NY. 1994.

(87) Abraham, M. H.; Acree, W. E. The solubility of liquid and solid compounds in dry octan-1-ol. *Chemosphere* **2014**, *103*, 26–34.

(88) Acree, W.; Lang, A. Acree Enthalpy of Solvation Dataset. *Figshare* **2015**,

(89) Systems, D. C. I. SMARTS - A Language for Describing Molecular Patterns. `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`.

(90) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(91) 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.; Blondal, K.; West, R. H.; Goldsmith, C. F.; Green, W. H. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696.

(92) Benson, S. W.; Golden, D. M.; Haugen, G. R.; Shaw, R.; Cruickshank, F. R.; Rodgers, A. S.; O'neal, H. E.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279–324.

(93) Han, K.; Jamal, A.; Grambow, C. A.; Buras, Z. J.; Green, W. H. An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation. *Int. J. Chem. Kinet.* **2018**, *50*, 294–303.

(94) Green, W. H.; West, R. H. RMG - Reaction Mechanism Generator. 2021; `https://rmg.mit.edu/`.

(95) Lay, T. H.; Bozzelli, J. W.; Dean, A. M.; Ritter, E. R. Hydrogen atom bond increments for calculation of thermodynamic properties of hydrocarbon radical species. *J. Phys. Chem.* **1995**, *99*, 14514–14527.

(96) Walker, P. J.; Haslam, A. J. A New Predictive Group-Contribution Ideal-Heat-Capacity Model and Its Influence on Second-Derivative Properties Calculated Using a Free-Energy Equation of State. *J. Chem. Eng. Data* **2020**, *65*, 5809–5829.

(97) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(98) Hoerl, A. Application of ridge analysis to regression problems. *Chem. Eng. Prog.* **1962**, *58*, 54–59.

(99) Distributed Asynchronous Hyperparameter Optimization in Python. `https://github.com/hyperopt/hyperopt`.

(100) Bergstra, J.; Yamins, D.; Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, 2013; pp I–115 to I–23.

(101) Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.

(102) Jaquis, B. J.; Li, A.; Monnier, N. D.; Sisk, R. G.; Acree, W. E.; Lang, A. S. Using Machine Learning to Predict Enthalpy of Solvation. *J. Solution Chem.* **2019**, 564–573.

(103) Dassault Systèmes, BIOVIA COSMOtherm, Release 2020. `http://www.3ds.com`.

(104) Dassault Systèmes, BIOVIA COSMObase, Release 2020. `http://www.3ds.com`.

(105) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. New universal solvation model and comparison of the accuracy of the SM5.42R, SM5.43R, C-PCM, D-PCM, and IEF-PCM continuum solvation models for aqueous and organic solvation free energies and for vapor pressures. *J. Phys. Chem. A* **2004**, *108*, 6532–6542.

(106) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769–779.

(107) Guthrie, J. P. A blind challenge for computational solvation free energies: Introduction and overview. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(108) Bell, I. H.; Wronski, J.; Quoilin, S.; Lemort, V. Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library coolprop. *Ind. Eng. Chem. Res.* **2014**, *53*, 2498–2508.