Solvation free energies of anions: from curated reference data to predictive models

Thomas Nevolianis,[†,‖] Jonathan W. Zheng,[‡,‖] Simon Müller,[¶] Matthias Baumann,[†] Sofja Tshepelevitsh,[§] Ivari Kaljurand,[§] Ivo Leito,[§] Irina Smirnova,[¶] William H. Green,[‡] and Kai Leonhard[*,†]

†*Institute of Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany*

‡*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, 02139 Massachusetts, United States*

¶*Institute of Thermal Separation Processes, Hamburg University of Technology, 21073 Hamburg, Germany*

§*Institute of Chemistry, University of Tartu, 50411 Tartu, Estonia*

‖*Contributed equally to this work*

E-mail: Kai.Leonhard@ltt.rwth-aachen.de

Abstract

Predicting the physicochemical properties of ionizable solutes, including solubility and lipophilicity, is of broad significance. Such predictions rely on the accurate determination of solvation free energies for ions. However, the limited availability of high-quality reference data poses a challenge in developing accurate, inexpensive computational prediction methods. In this study, we address both issues of data quality and availability. We present three databases and models related to ionic phenomena:

(1) 8,241 pK_a datapoints across 8 solvents, (2) 5,536 gas-phase acidities from DLPNO-CCSD(T) QM calculations, and (3) 6,090 solvation free energies of anions across 8 solvents obtained from a thermodynamic cycle. We also report 6,088 solvation free energies of neutral conjugate solutes computed using the COSMO-RS method. The pK_a data were obtained from the iBonD database, cleaned, and combined with a separate compilation of trustworthy reference pK_a data. Gas-phase acidities were computed for most of the acids present in the pK_a corpus. Leveraging these data, we compiled values for solvation free energies of anions. We then trained several graph neural network models, which can be used as an alternative to QM approaches to quickly estimate these properties. The pK_a and gas-phase acidity models accept reaction SMILES strings of the acid dissociation as inputs, whereas the solvation energy model accepts the SMILES string of the anion. Our microscopic pK_a model achieves good accuracy, with an overall test mean average error of 0.58 units on unseen solutes and 0.59 on the SAMPL7 challenge (the lowest error so far among multi-solvent models). Our gas-phase acidity model had mean absolute errors slightly above 2 kcal mol^{-1} when evaluated against experimental data. The anionic solvation free energy model had mean absolute errors of less than 3 kcal mol^{-1} in several test evaluations, comparable to (though less reliable than) several widely-used QM-based solvation models. The models and data are free and publicly available.

Introduction

The accurate prediction of ionic solvation phenomena is essential for applications across numerous fields, including materials science,[1,2] renewable energy,[3,4] pharmaceuticals,[5–7] and environmental science.[8] Such predictions play key roles in the optimization of chemical processes[9,10] and prediction of behavior for biological systems.[11] In particular, the solvation free energy is fundamentally related to useful physicochemical properties. These include solubilities, acid dissociation constants (pK_a), partition coefficients, and others. A combination

of high accuracy and fast prediction speed are typically desired. As such, quantum chemical and machine learning (ML) approaches have become popular in recent years. Despite significant advancements in semi-empirical[12–25] approaches and data-driven methods,[26–34] accurate predictors for anionic solutes remain challenging to develop due in large part to the limited availability and accuracy of data.

The compilation of comprehensive databases for solvation free energies of ions is particularly difficult due to the complexities involved in measuring thermodynamic properties of charged solutes. Such properties cannot be accessed directly due to the principles of electroneutrality and numerous experimental and theoretical challenges. Instead, one must invoke *extrathermodynamic assumptions*,[35–38] which are unproven conjectures that allow bulk thermodynamic measurements of a salt to be partitioned into each individual ion. Even if such assumptions are made, solvation free energies are still difficult to obtain due to the vast differences in stability of ions between the gaseous and condensed phases.

Given these challenges, one strategy involves using a thermodynamic cycle, which relates the solvation free energy of charged solutes to experimentally accessible properties. One widely-used cycle includes gas-phase acidities, acid dissociation constants, and solvation free energies of neutral solutes, which are then "anchored" to the proton's solvation free energy in the desired solvent. Kelly *et al.* [39] used this approach with experimental gas-phase acidities and acid dissociation constants to report 121 aqueous solvation free energies of ions. Building on this work, Marenich *et al.* [13] extended the data to non-aqueous solvation free energies in solvents such as acetonitrile, dimethyl sulfoxide, and methanol, as part of the development of the Minnesota Solvation Database (MNSol) database,[13,40] which includes approximately 300 entries for ions. More recently, the Database of Ionic Solutes' Solvation Free Energies (DISSOLVE) database by Nevolianis *et al.* [41] expanded this dataset, including 330 entries for both aqueous and non-aqueous solvation free energies of ionic solutes, with reported standard uncertainties of 1.5 $\mathrm{kcal\,mol^{-1}}$ and 2.6 $\mathrm{kcal\,mol^{-1}}$, respectively. This work showed the feasibility of reliably computing gas-phase acidities and solvation free energies

for neutral solutes, which served as bottlenecks in compiling large sets of solvation free energies for ions. The IonSolv-Aq database introduced by Zheng *et al.* ,[42] with reported standard uncertainties of 2.5 kcal mol^{-1}, focused on hydration free energies of 118 anions and 155 cations using only experimentally-derived thermodynamic properties. Their study suggested that systematic deviations between experimental data and computed predictions (from the COSMO-RS and SMD solvation models) for ionic solutes can be corrected using empirical parameters. Despite the contributions of these databases (MNSol, DISSOLVE, and IonSolv-Aq), the number of entries remains insufficient for comprehensive parameterization of solvation models and development of accurate data-driven models.

Data-related challenges are also present for pK_a and gas-phase acidities. Although extensive pK_a data compilations exist, they are often of inconsistent data quality. For example, the Internet Bond-energy Databank (iBonD) database[43] includes over 30,000 experimental equilibrium acidity data points for roughly 20,000 unique compounds across various solvents. However, values from different authors sometimes differ by several pK_a units, due to inconsistencies in the literature regarding experimental determination of pK_a values. Such values for the same molecule can deviate by more than 10 pK_a units.[43] In particular, acid dissociation constants in non-aqueous solvents are often measured relative to each other and then anchored to the pK_a of a reference compound. Inconsistent choice of reference compounds can lead to high errors (for instance, two different pK_a values for a common anchor compound in acetone lead to pK_a values differing by nearly 3 pK_a units).[44,45] Additional errors are introduced from inconsistent treatment of ion pairing, particularly in poorly-screening solvents; homoconjugation, in which neutral acids form hydrogen-bonded adducts with their corresponding conjugate anions; and impurity of solvent (e.g., water content).[46] These errors alone can lead to differences of several pK_a units. One pK_a unit propagates to 1.4 kcal mol^{-1}, so such problems in data quality can lead to very large errors in reported free energies.

Experimental measurements of gas-phase acidities are quoted with uncertainties exceeding 2 kcal mol^{-1} (though most of that uncertainty is systematic, arising from the need to

anchor the gas-phase acidity scale).[47,48] Although gas-phase acidities can be predicted with relatively high accuracy using quantum chemical methods, such techniques are computationally expensive,[41] limiting the feasibility of high-throughput workflows. To this end, an optimal balance between computational cost and accuracy is thus crucial for predicting gas-phase acidities.

In this study, we introduce three new databases related to ionic phenomena. We present 6,090 solvation free energies of anionic solutes, as well as datasets of curated experimental pK_a data across 8 solvents and QM-computed gas-phase acidities that compose this set of anionic solvation free energies. We also present a set of solvation energies of neutral acids.

The pK_a data are collected from the iBonD database as well as a set of recently compiled and curated pK_a data, and jointly contain 8,241 entries. We curated these data by manually examining the pK_a values, checking the data for unreasonable values based on agreement between values from different sources, correlations between data in different solvents, acidity trends, and chemical feasibility. For most of these acids, we calculated their gas-phase acidities, benchmarking different levels of quantum chemical theory observed in our previous work[41] to achieve an effective balance of computational cost and accuracy. Using the COnductor like Screening MOdel for Real Solvents (COSMO-RS)[15,49] method, we computed the solvation free energies of the neutral acids, and leveraged previously-published data for the solvation free energy of the proton in both aqueous and non-aqueous solvents along with a thermodynamic cycle to obtain solvation free energies of anions.

We developed and evaluated Graph Neural Network (GNN) models for predicting each property (D-MPNN/FFNN as implemented in Chemprop v2[50]), providing insights into their relative strengths and limitations and discussing future research directions for modeling the complex chemistries of ionic solutes. To our knowledge, this work represents the first time that a machine learning model was developed for predicting gas-phase acidities and solvation free energies of anions.

Figure 1 depicts the workflow used for the development of the datasets. We utilize a

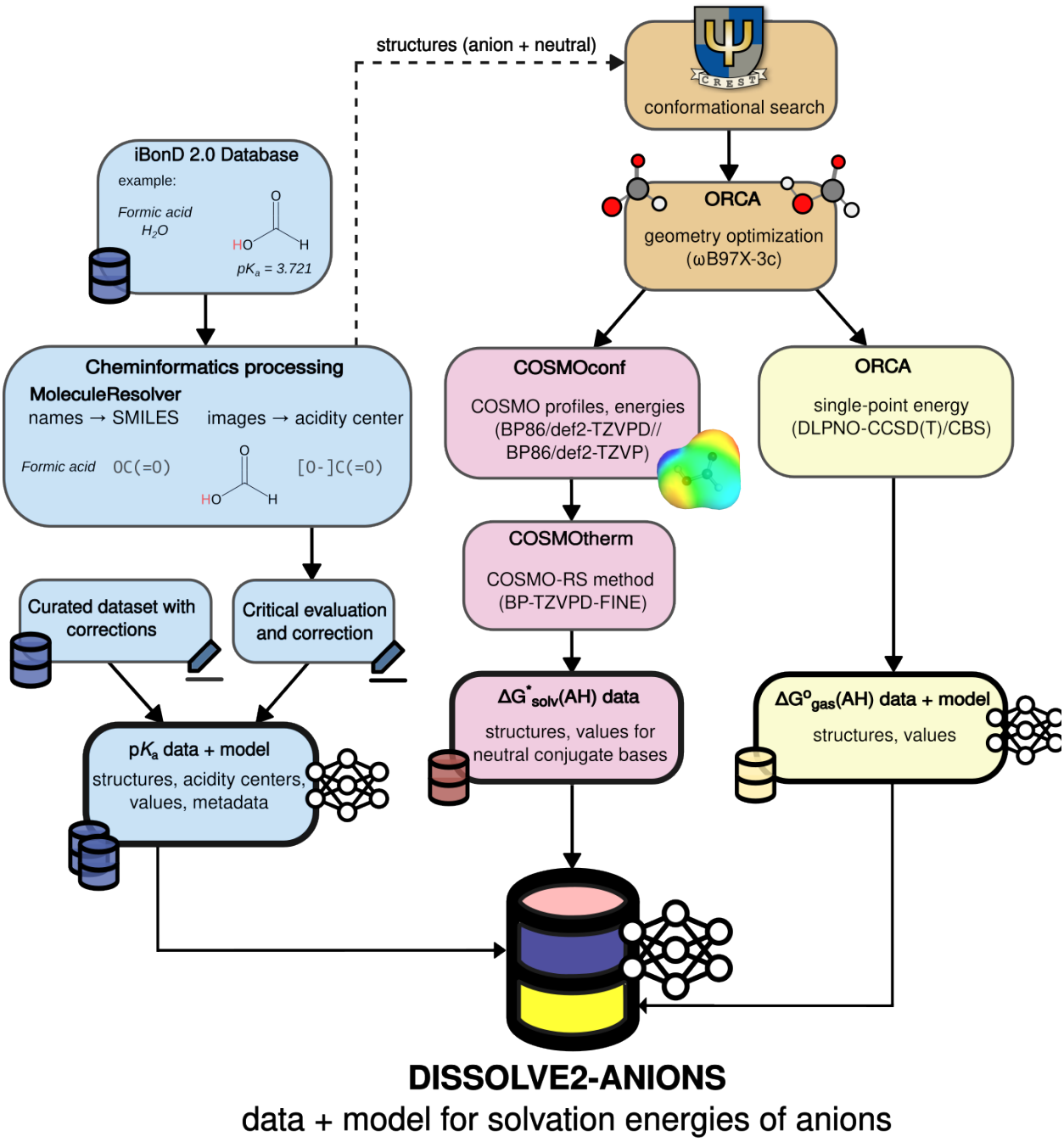

Figure 1: Workflow for constructing the DISSOLVE2-ANIONS databases.

thermodynamic cycle for calculating the solvation free energies of anions at 298 K:

$$\begin{aligned}\Delta G^*_{\text{solv}}(\text{A}^-) = &- \Delta G^o{}_{\text{acid, gas}}(\text{AH}) + \Delta G^*{}_{\text{acid, soln}}(\text{AH}) + \Delta G^*{}_{\text{solv}}(\text{AH}) \\ &- \Delta G^*{}_{\text{solv}}(\text{H}^+) - \Delta G^{o\rightarrow *}\end{aligned} \tag{1}$$

The acid dissociation free energy values in solution are expressed through the $\text{p}K_\text{a}$ values as $\Delta G^*{}_{\text{acid, soln}}(\text{AH}) = \ln(10) \cdot RT \cdot \text{p}K_\text{a}(\text{AH})$. $\Delta G^*_{\text{solv}}(\text{AH})$ is the solvation free energy of the neutral conjugate acid. $\Delta G^o_{\text{acid, gas}}(\text{AH})$ is the gas-phase acidity of the acid AH. The ^{o} symbol denotes an 1 atm gas-phase reference state, and the * denotes an 1 M reference state in either the gas or the solution phase. $\Delta G^{o\rightarrow *}$ is defined as the free energy difference between the 1 atm (or 1 bar) and the 1 M reference states of an ideal gas, and is equal to 1.9 kcal mol^{-1} at 298 K. $\Delta G^*_{\text{solv}}(\text{H}^+)$ is the proton's solvation free energy in a solvent. The gas-phase acidity is defined as $\Delta G^o_{\text{acid, gas}}(\text{AH}) = G^o{}_{\text{gas}}(\text{A}^-) + G^o{}_{\text{gas}}(\text{H}^+) - G^o{}_{\text{gas}}(\text{AH})$.

Results and Discussion

Overview of datasets

Table 1: Overview of datasets released in this work, and experimental data used to benchmark the calculations.

Property	**This work**			**External evaluation data**		
	Database name	**Description**	$\mathbf{n_{data}}$	**Dataset name**	**Description**	$\mathbf{n_{data}}$
$\text{p}K_\text{a}$	D2A-pKa	Literature	8,241	-	-	-
$\Delta G^o_{\text{acid, gas}}(\text{AH})$	D2A-dGgas	QM calcs	5,536	NIST/overlap	NIST Chemistry Webbook[51,52]	317
$\Delta G^*_{\text{solv}}(\text{AH})$	D2A-dGsolv-neutral	COSMO-RS	6,088	dGsolvDB3/overlap	Chung et al. (2023)[27]	624
$\Delta G^*_{\text{solv}}(\text{A}^-)$	D2A-dGsolv-anion	Thermo cycle	6,090	ExpSolv/overlap	Compilation from literature[24,27,52]	105
				MNSol/overlap	Minnesota Solvation Database[53]	171

The databases we compiled are summarized in Table 1. In this work, we use a systematic naming convention. DISSOLVE2-ANIONS refers to the set of *all* physicochemical data that we compiled or computed: $\text{p}K_\text{a}$, $\Delta G^o_{\text{acid, gas}}(\text{AH})$, $\Delta G^*_{\text{solv}}(\text{AH})$, $\Delta G^*_{\text{solv}}(\text{A}^-)$. The prefix D2A indicates that the database is a subset of the DISSOLVE2-ANIONS data that we curated in

Table 2: Overview of data splits and external data used for training and evaluating the GNN models. Stereoisomers of species in the training and validation sets were removed from the test sets.

Property	This work			External evaluation data		
	n_{train}	$n_{validation}$	n_{test}	Dataset name	Description	n_{data}
pK_a	6,774	846	603	SAMPL	SAMPL6 *acids* + SAMPL7 challenges[54,55]	32
				Zheng/nonoverlap	Challenge test set in Zheng et al., 2024[45]	3
$\Delta G^o_{acid,\ gas}$(AH)	4,496	562	464	NIST/nonoverlap	NIST Chemistry Webbook[51,52]	327
ΔG^*_{solv}(AH)	–	–	–	–	–	–
$\Delta G^*_{solv}(A^-)$	4,862	607	621	ExpSolv/nonoverlap	Compilation from literature[24,27,52]	8
				Pliego/nonoverlap	Data[41] with QM calculations[56]	7
				MNSol/nonoverlap	Minnesota Solvation Database[53] with QM calculations[24]	26

this work.

We also indicated data splits via suffixes. The suffix "/train" refers to the training/validation split, and "/test" refers to the test split. For external test data, "/overlap" indicates that the dataset was filtered such that its acids (or if appropriate, acid-solvent pairs) also show up in the respective property database of DISSOLVE2-ANIONS. These are used for assessing the uncertainty of the QM methods. On the other hand, "/nonoverlap" indicates that the acids/anions do not appear in the training/validation sets. These splits are useful for assessing the performance of the deep learning models, as they ensure that there is no data leakage. The data splits and external data for testing the models are summarized in Table 2. Digital versions of the data are available in the Zenodo (doi:10.5281/zenodo.13987781).

In the following sections, we will examine our compiled data for each property. We further assess the uncertainty of the data by comparing them with the external datasets labeled in Table 1. We then discuss the performance of the GNN models that we trained using Chemprop v2 by testing the model against datasets in Table 2. More information about the ML training procedure is provided in the Methods section.

pK_a Data and Modeling

D2A-pKa: Dataset Summary

Data in D2A-pKa were curated from two compilations of experimental data: iBonD,[43] as well as data from the "Acid dissociation constants in selected dipolar non-hydrogen-bond-donor solvents" project (IUPAC project 2015-020-2-500), a currently in press compilation of experimental pK_a data through 2024 which includes critical evaluation and in many cases corrections for all data entries.[57] The respective publication in *Pure and Applied Chemistry* is in press and is expected to be published in the near future. Both data sources were checked for errors and occasionally corrected. A more detailed discussion of the curation process is in the Methods section.

Figure 2 displays histograms of the pK_a values of the acids in several solvents and their corresponding molecular weight distributions (the distribution of ion types for pK_a values can be found in the SI).

The pK_a distributions in D2A-pKa vary significantly across different solvents, as shown in the Figure 2. For example, in water, pK_a values mostly fall within 0 to 14, but pK_a data in dimethyl sulfoxide (DMSO) is more widely distributed. These differences largely reflect the measurable pK_a ranges in each solvent (determined by the solvent's autoprotolysis constant)[46] as well as the different selections of acid types that have been measured and compiled in each solvent. For example, carboxylic acids make up nearly 40% of the pK_a data in water (and contribute to the peak near 4 pK_a units), but only 12% of the pK_a data reported in DMSO. The acidities of many C-H bonds have been measured in DMSO; these vary widely in their pK_a values. Given the different solvation characteristics of each solvent (which is observed in the variations in the pK_a value distributions), it is expected that models trained on D2A-pKa will also significantly depend on solvent.

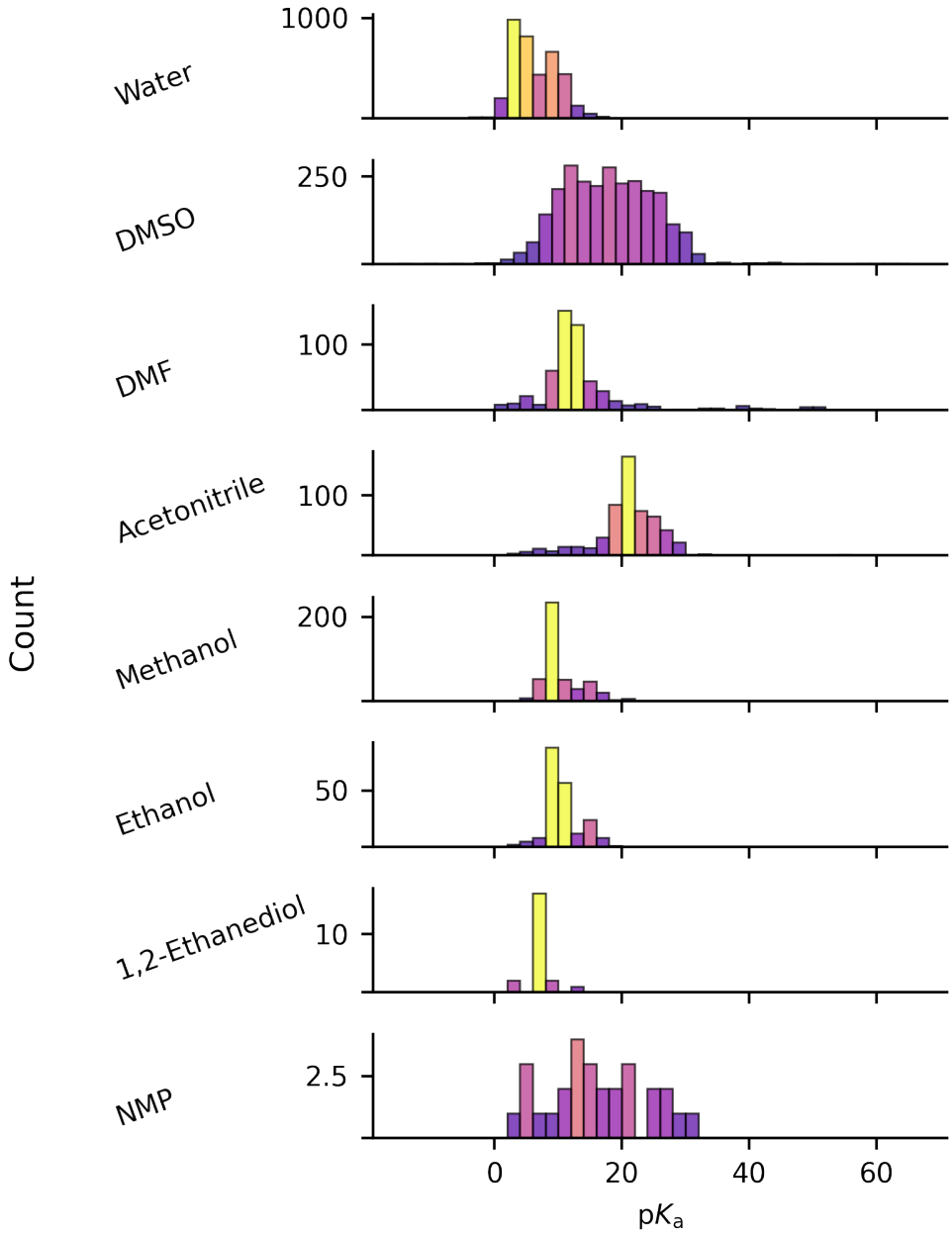

Figure 2: Distributions of pK_a of the acids in D2A-pKa. Many more acids have been measured in water and DMSO than in the other solvents. Brighter bars indicate higher density.

D2A-pKa: Data Uncertainty

The experimental uncertainties in pK_a (1σ standard error) are estimated as 0.23 kcal mol^{-1} for aqueous (0.17 pK_a units) and 0.41 kcal mol^{-1} for non-aqueous solvents (0.30 pK_a units) in pK_a values from previous literature estimates.[58–73]

pK_a model

Figure 3 shows the performance of the D-MPNN/FFNN model trained on D2A-pKa/train, tested on a subset consisting only of acids not seen during training (D2A-pKa/test).The model demonstrates acceptable predictive accuracy on the test set D2A-pKa/test, achieving an overall Mean Absolute Error (MAE) of 0.58 pK_a units and a Root-Mean-Square Error (RMSE) of 1.07 pK_a units (corresponding to about 1.4 kcal mol^{-1}). Some solvents have significant outliers; many of those correspond to "rare" acid sites such as carbon acids or nitrogen-centered acids, or small and unusual compounds such as ethenimine or hydrogen peroxide.

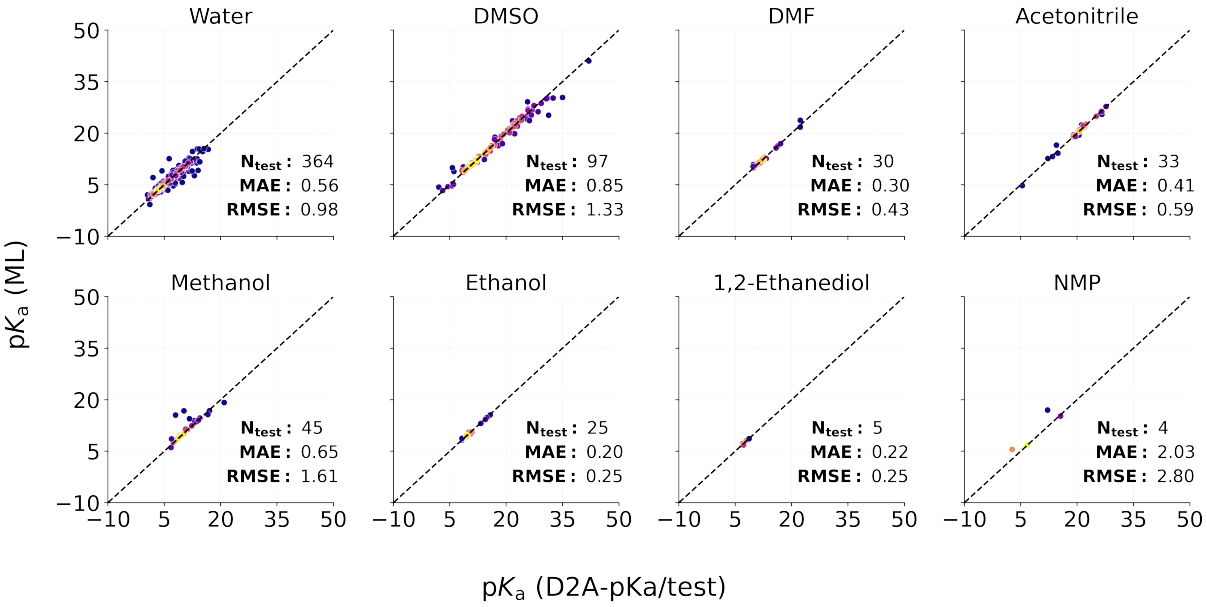

Figure 3: Performance of D-MPNN/FFNN model trained on pK_a data and tested against full test set. Brighter spots indicate higher density.

Other significant multi-solvent pK_a models have been trained, both using data from

iBonD: the XGBoost and neural network models by Yang *et al.* in 2020[74] and the AttenGpKa model by An *et al.* in 2024.[75] The Yang model is reported with an overall MAE of 0.87 pK_a units when trained across all 39 solvents. The AttenGpKa model similarly reports test statistics of 0.73 / 1.46 MAE / RMSE across a pool of 60 solvents. The error statistics we report herein are lower (overall 0.58 / 1.07 MAE / RMSE), albeit on a smaller set of only 8 solvents. However, it is insufficient to compare against different test splits across different solvents. We do expect that our data curation and cleaning potentially improved model performance, but cleaner data could also mean that test splits are less challenging, or sample more frequently from a solvent with abundant training data. Therefore, we compare our model against the SAMPL6 *acids*[54] and SAMPL7[55] aqueous pK_a challenges (SAMPL), which are widely used as model benchmarks. We ensured that none of the SAMPL molecules appeared in the D2A-pKa/train split.

Table 3: Comparison of MAE and RMSE values for different multi-solvent pK_a models across the aqueous SAMPL6 acids and SAMPL7 datasets.

Model	SAMPL6 acids MAE/RMSE	SAMPL7 MAE/RMSE
Yang XGB	0.79/1.10	1.48/1.62
Yang NN	0.93/1.34	0.93/1.16
AttenGpKa	**0.49**/**0.61**	1.00/1.24
Our model[a]	0.63/0.96	**0.59**/**0.78**

[a]Note that we removed entries SM28 and SM33 from SAMPL7 for consistency with the other models. If these datapoints are restored, our model's MAE/RMSE are 0.64/0.84.

Table 3 compares the performance of various multi-solvent pK_a models against SAMPL. Among them, our model performs well on the SAMPL6 challenge, although with worse performance than AttenGpKa, and performs the best for the SAMPL7 challenge.

It is worth noting that state-of-the-art models trained only on aqueous pK_a data perform generally better on the SAMPL6 acids benchmark. The QupKake model has reported MAE/RMSE values of 0.32/0.44 and 0.67/0.85 for the full SAMPL6 and 7 challenges.[76] Uni-pK_a reported 0.55/0.72 and 0.57/0.74 for those same benchmarks. Compared to these, our model performs comparably on the SAMPL7 challenge, with worse relative performance

on the SAMPL6 challenge (though, note, the test metrics for QupKake and Uni-pK_a were reported for the *full* test sets, so we could not compare only the acids). Such aqueous models include large corpuses of pretraining data in water from the ChEMBL database (more than 2 million values) as well as thermochemical information from quantum-chemical calculations. Our model was trained on a much smaller number of aqueous pK_a values - only 3,416 - so its low errors with the SAMPL7 set are somewhat surprising. There is also potential for data leakage in the other models, as some of the SAMPL molecules appear in the ChEMBL database used to pretrain those ML models, and it is unclear whether those datapoints were removed during training. We advise future model developers to exclude SAMPL compounds if pretraining on ChEMBL.

We also benchmarked the ML model against the data and semi-QM calculations in the Zheng/nonoverlap dataset, which includes 2 amino acids in methanol and 1 in ethanol. The semi-QM method estimates pK_a values in non-aqueous solvents using a thermochemical cycle, experimental aqueous pK_a data, and solvation free energies computed with COSMO-RS. The method includes fitting an empirical parameter δ which accounts for $\Delta G^*_{\text{solv}}(\text{H}^+)$ as well as any remaining systematic error.[77] The performance of the ML and semi-QM methods for these was comparable, with the semi-QM method leading to 0.90 and 1.34 pK_a unit RMSEs in methanol and ethanol, respectively, versus 1.03 and 0.92 pK_a units for the ML model.

Overall, the D-MPNN/FFNN model demonstrates fairly good predictive accuracy for pK_a values in solvents where sufficient amount of training data was available, even when applied to acids not included in the training set. We emphasize that such training was done on a relatively small subset of training data: only 6,774 datapoints, owing to examining only mono-anions and the rigorous cleaning process, much lower than other multi-solvent literature models which have been trained on 20,000+ datapoints. Variations of our model trained on slightly smaller splits of data demonstrate worse performance, indicating that continuing to add data would improve the model (see SI).

We believe that the good performance, in spite of a much lower total number of datapoints, reflects the effectiveness of cleaning and checking data. Future efforts should focus on refining the reference data in the solvents that were not explored in this work, which will enable development of accurate predictive models for those solvent media. We also suggest pretraining on pK_a values computed using quantum-chemical methods on a broad range of solutes/solvents,[45] to allow more reliable predictions for solvents and solutes very different from those studied experimentally.

$\Delta G^o_{\text{acid, gas}}$(AH) Data and Modeling

D2A-dGgas: Dataset Summary

To generate the $\Delta G^o_{\text{acid, gas}}$(AH) database, we took the SMILES strings of most acids from D2A-pKa and converted the 2D molecular graph representations into 3D conformers. We used molecular representations of anions based on deprotonating the neutral acids identified in the pK_a datasets at their labeled acidity centers. We used a CREST conformer generation workflow followed by DLPNO-CCSD(T)/CBS(aug-cc-pVDZ/pVTZ)// ωB97x-3c calculations. Further QM details are presented in the Methods section.

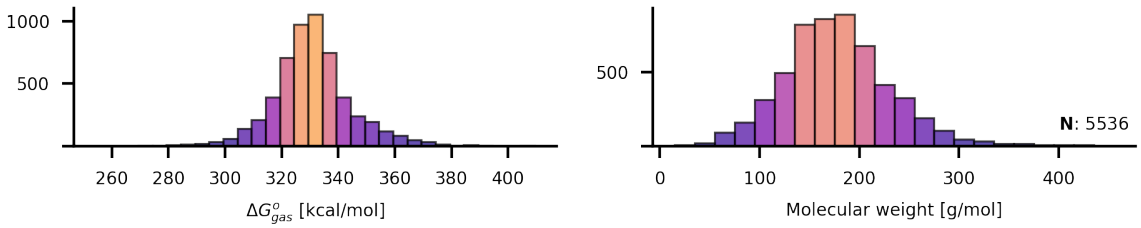

Figure 4: Distributions of gas-phase acidities as well as their molecular weight distributions in D2A-dGgas. Brighter bars indicate higher density.

Figure 4 presents the distributions of our computed (DLPNO-CCSD(T)/ωB97x-3c) gas-phase acidities $\Delta G^o{}_{\text{gas}}$(AH) alongside the molecular weight distributions of the corresponding acids (the distribution of ion types in our computed gas-phase acidity dataset can be found in the SI). Our computed gas-phase acidities range from 254 kcal mol^{-1} to 412 kcal mol^{-1}.

The distribution is centered around 320 kcal mol^{-1} to 340 kcal mol^{-1}.

The right panel displays the molecular weight distribution of the acids, which primarily fall between 50 and 350 g mol^{-1}, with a peak around 150 g mol^{-1} to 200 g mol^{-1}. Such molecules are small to medium-sized species, and nearly all are organic (roughly 30 species do not contain carbon).

D2A-dGgas: Data Uncertainty and Validation

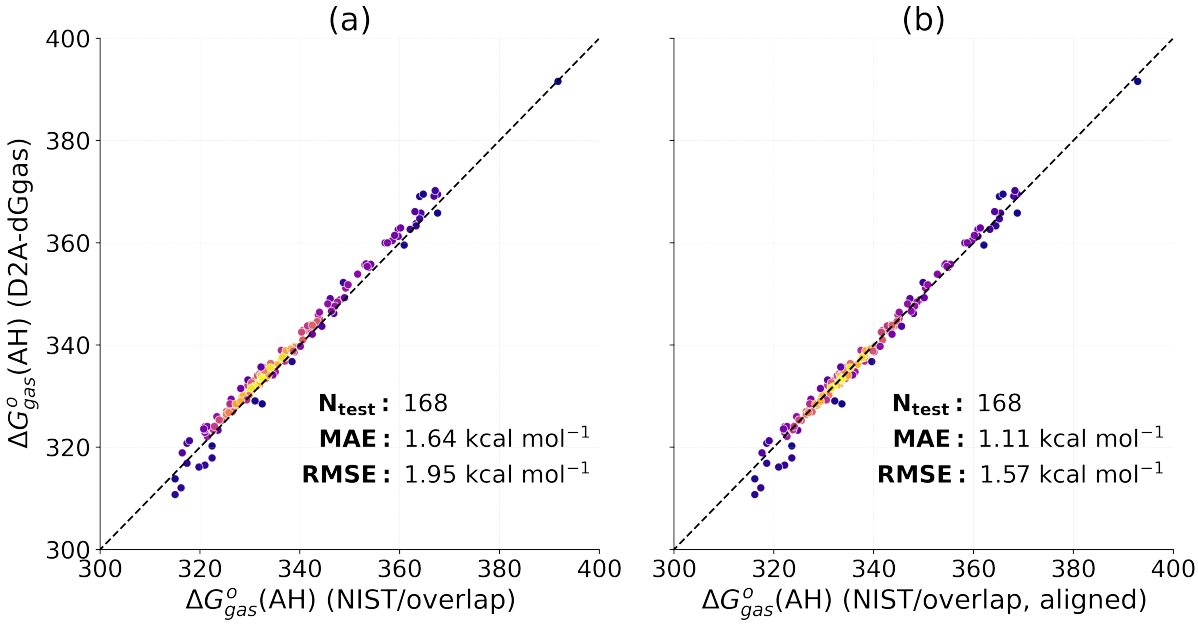

Figure 5: (a) Comparison between computed gas-phase acidities using the DLPNO-CCSD(T)/CBS//ωB97x-3c (normalPNO criteria) method and experimental values with low uncertainty from the NIST Webbook. Most of the computed values are higher than the experimental values. (b) Replotted after using a calibration factor to align the energy scales.

To better assess the accuracy of the $\Delta G^o_{\text{acid, gas}}$(AH) calculations in D2A-dGgas, we sampled a set of experimental $\Delta G^o_{\text{acid, gas}}$(AH) values from the NIST Chemistry Webbook (NIST/overlap). We used the provided names and chemical identifier information (whenever available) from the Webbook to determine acidity centers. Similar to the situation for pK_{a} data, potential error is introduced by using a single isomeric form to represent the anion, because experimental gas-phase acidities may correspond to numerous isomers, or the exact nature of the anion

is not known. We pruned the data, retaining only values with low reported uncertainties in order to establish a more reliable baseline. We also excluded carbanions due to difficulty in inferring the correct ionization sites.

Experimentally determined $\Delta G^o_{\text{acid, gas}}(\text{AH})$ values are often measured relative to each other by means of measuring equilibrium constants between two neutral compounds and their anionic forms. These relative affinities are then anchored to the value for a reference compound. The uncertainty of *relative* $\Delta G^o_{\text{acid, gas}}(\text{AH})$ measurements from these equilibrium measurements is cited as approximately 0.2 kcal mol^{-1}. However, anchoring the energy scale incurs a larger uncertainty, typically cited as around 2 kcal mol^{-1}.[47,48] For this reason, a systematic offset caused by a difference in anchor values may exist between computed values and curated experimental data. An uncertainty of around 2 kcal mol^{-1} is typically cited for $\Delta G^o_{\text{acid, gas}}(\text{AH})$, though most of the uncertainty is simply from the anchoring process.

For this reason, when comparing the DLPNO-CCSD(T)/CBS(normalPNO)//ωB97x-3c calculations to the NIST data, we aligned the values by adding a shift parameter to the NIST data that minimizes the RMSE of the values compared to the QM calculations. Prior to alignment, we observed a mean signed deviation of 1.2 kcal mol^{-1} between the QM values and the NIST data, indicative of systematic error. After alignment, we observed an RMSE value of 1.57 kcal mol^{-1} which we assign as the uncertainty of the gas-phase acidity QM calculations (see Figure 5). For consistency, all $\Delta G^o_{\text{acid, gas}}(\text{AH})$ data from NIST used in this work will be aligned to the QM data by shifting them by 1.2 kcal mol^{-1}.

$\Delta G^o_{\text{acid, gas}}(\text{AH})$ model

The test performance of the GNN model for gas-phase acidities on held-out QM data can be seen in Figure 6a. Although the model visually appears to predict most acidities fairly accurately, the test statistics of 2.35 / 3.73 kcal mol^{-1} MAE / RMSE are somewhat high, larger than the uncertainty in QM calculations of approximately 2 kcal mol^{-1} compared to the NIST data (and much larger than the *relative* uncertainty in experimental gas-phase acidity

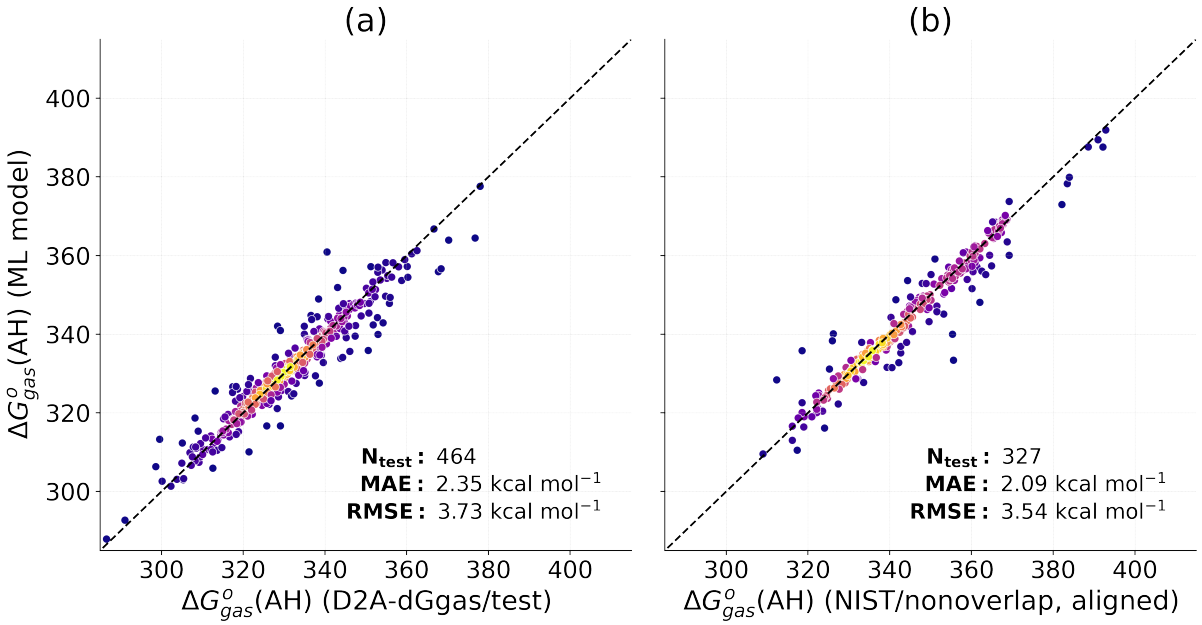

Figure 6: Performance of D-MPNN/FFNN model trained on gas-phase acidity QM calculations against (a) held-out gas-phase QM data and (b) experimental data shifted up by 1.2 kcal mol^{-1}. If using unshifted NIST data, MAE / RMSE are 2.26 / 3.56 kcal mol^{-1} instead. Brighter spots indicate higher density.

data, cited to be approximately 0.2 kcal mol^{-1} [47,48]). Because we are training and predicting on QM calculations, we do not expect systematic error associated with experimental data to manifest, though systematic errors due to QM calculations may still appear. Therefore, such errors are due to inherent deficiencies in model training (such as low amounts of training data) and QM calculation error.

The ratio of acceptable error to the range of property values is quite low. Gas-phase acidities span roughly 100 kcal mol^{-1}, but values accurate enough for room-temperature kinetic predictions should have errors less than 2 kcal mol^{-1}, which is lower than the aleatoric uncertainty of experimental gas-phase acidity data, i.e. we desire to predict gas-phase acidities with <2% error. Hence, although the error metrics are somewhat high for the gas-phase acidities (the RMSEs are around 4 kcal mol^{-1}, twice as large as tolerable error), the percent errors of about 4% are fairly low, and the model generally captures the overall trends and relative acidities among different compounds.

To further assess the accuracy of our model, we compared predictions of the model trained on QM data to the experimental data in NIST/nonoverlap, from which we also removed carbanions and anions whose acidity centers were not included in the D2A-dGgas/train set. The model performance on NIST/nonoverlap (Figure 6b) is comparable to that on the QM data (Figure 6a), with an RMSE of 3.54 compared to 3.73 $\mathrm{kcal\,mol^{-1}}$. The NIST/nonoverlap set includes many compounds whose acidities are systematically underpredicted by the ML model by 20 $\mathrm{kcal\,mol^{-1}}$ or more; these points drive the high RMSE values. Many of these are boron acids, weak carbon acids, and pyrazoles. The very poor performance for those outliers is likely driven by poor extrapolation to unseen chemistries. We had initially observed high errors for iminols (which tautomerize to amides) as well as SMILES containing neutral species in keto form and anions in enolate form. Correcting the tautomers encoded in the SMILES strings led to considerably lower errors (see SI).

It is worth emphasizing that model performance strongly depends on functional group. Errors associated with carboxylates and phenolates are lower than those of other groups, with RMSEs near 2 $\mathrm{kcal\,mol^{-1}}$, close to the aleatoric limit of the QM calculations. At the other extreme, RMSEs for oxime anions exceed 8 $\mathrm{kcal\,mol^{-1}}$ (see SI). We advise potential users of this model to be wary of such potential stratification of performance based on functional groups. Our expectation is that errors from predicting gas-phase acidities of carboxylic acids and phenols will be significantly lower than those of other functional groups.

There is both a scarcity of data as well as ambiguity in the quality of experimental data. For instance, although the relative uncertainties of $\Delta G^{o}_{\mathrm{acid,\ gas}}(\mathrm{AH})$ are quoted as just 1 $\mathrm{kJ\,mol^{-1}}$,[47] values from different authors in the literature sometimes differ by several $\mathrm{kcal\,mol^{-1}}$.[78,79] In some cases, the gas-phase acidities were measured at temperatures significantly higher than 298 K, but our model was trained with calculations at 298 K. Such potential issues may contribute to the relatively high errors seen herein.

We encourage future work to focus on more accurately computing gas-phase acidities for diverse sets of molecules. Increasing the number of training datapoints would improve the

model performance (see SI) as well as methods for compensating for tautomerization, such as data augmentation with tautomer enumeration, or canonicalization of input tautomer structures.

ΔG^*_{solv}(AH) Data and Modeling

D2A-dGsolv-neutral: Dataset Summary

We calculated ΔG^*_{solv}(AH) values for the neutral acids in D2A-dGgas, with the intention of using them in a thermodynamic cycle to calculate $\Delta G^*_{\text{solv}}(A^-)$. We employed the COSMO-RS[15,49] method, which is described in the Methods section.

Figure 7 shows the distribution of ΔG^*_{solv}(AH) for different solvents. The number of ΔG^*_{solv}(AH) data is just two less than that of $\Delta G^*_{\text{solv}}(A^-)$ (lower because some acids can lose a proton at different locations). The ΔG^*_{solv}(AH) values range from -36 to 0 kcal mol^{-1}, and are mostly concentrated between -20 to -5 kcal mol^{-1}. The solvation energy distributions are mostly unimodal, and water is the most widely-represented solvent.

D2A-dGsolv-neutral: Data Uncertainty and Validation

For predicting the solvation free energy of neutral solutes using COSMO-RS, an RMSE of 0.67 kcal mol^{-1} was previously estimated for both aqueous and non-aqueous solvents.[26,80] To assess our values, we found the overlapping solutes in the experimental neutral solvation energy database `dGsolvDB3` (dGsolvDB3/overlap) and compared the D2A-dGsolv-neutral values to the ones therein.

Figure 8 shows that the error between neutral solvation energies is significantly higher than the RMSE of 0.67 kcal mol^{-1} reported previously.[26] The overall MAE / RMSE was 1.09 / 1.57 kcal mol^{-1}, driven by the high representation of data in water. Although RMSEs in DMSO and NMP are lower than the previous 0.67 kcal mol^{-1} benchmark, they represent only a very small portion of the data. The RMSE values in water, DMF, methanol, and ethanol are much larger.

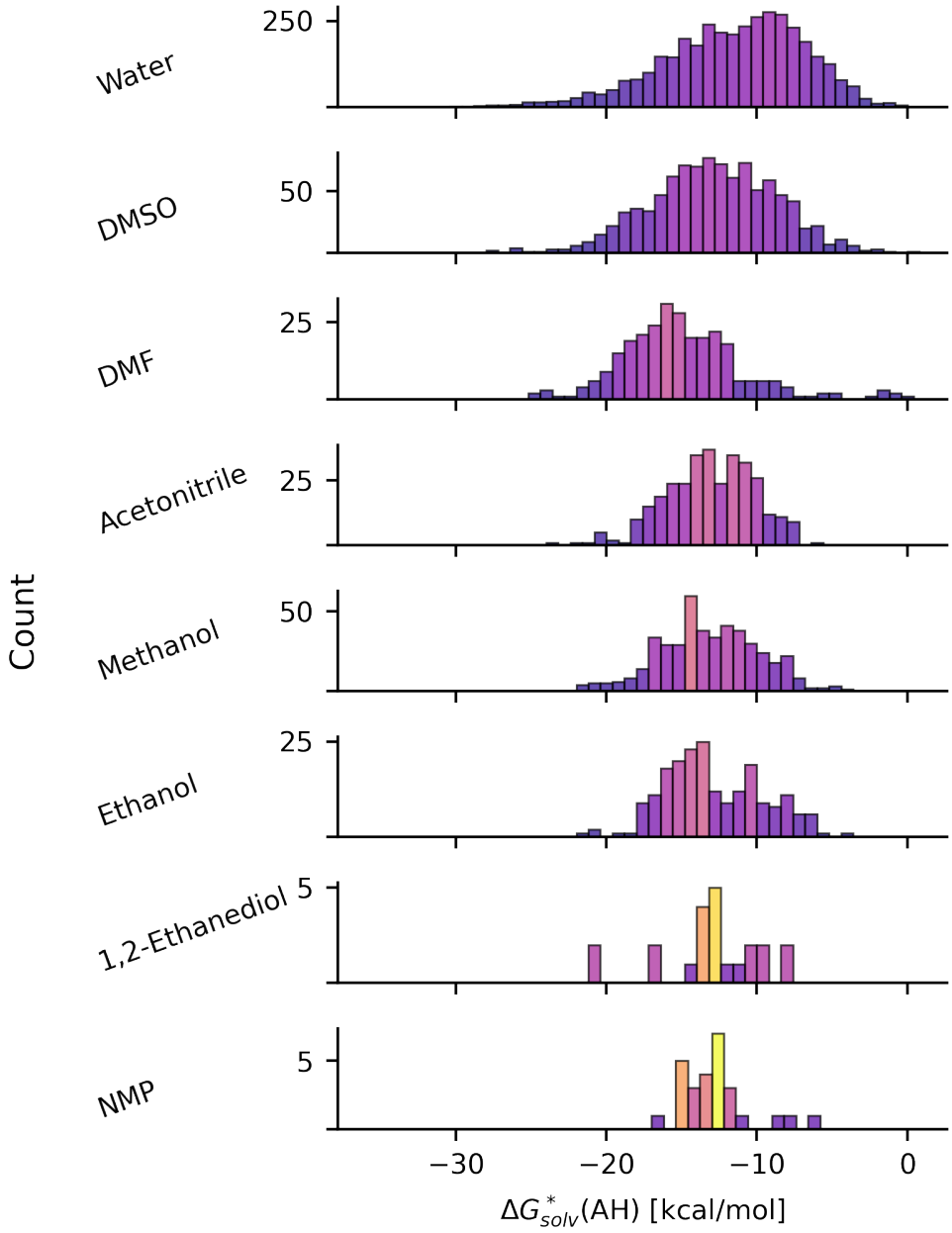

Figure 7: Distributions of ΔG^*_{solv}(AH) and molecular weights of the neutral acids in D2A-dGsolv-neutral. Brighter bars indicate higher density.

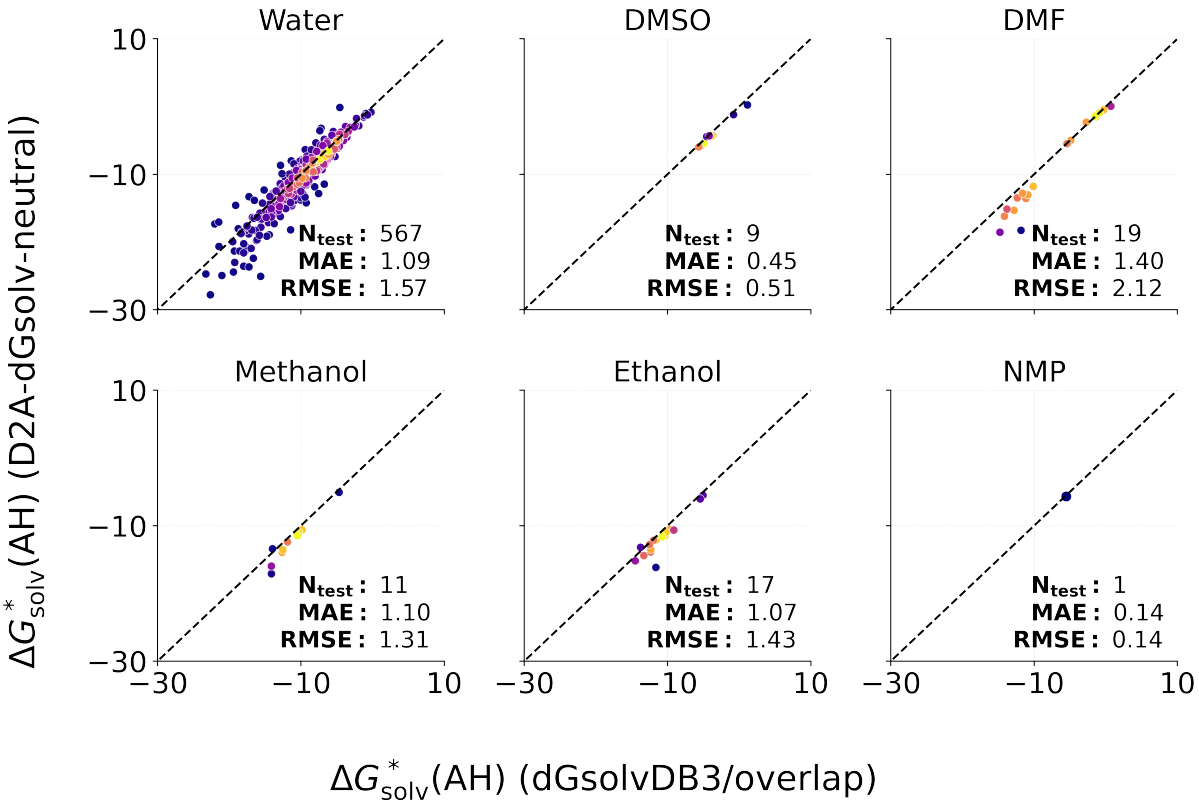

Figure 8: Comparison of data from D2A-dGsolv-neutral to the values from the experimental dGsolvDB3/overlap dataset. The overall MAE / RMSE is 1.09 / 1.57 kcal mol^{-1}.

The difference between the COSMO-RS calculated values and those from dGsolvDB3/overlap are significant. We observed that ΔG^*_{solv}(AH) calculations more negative than -10 kcal mol^{-1} tended to have much higher errors than those more positive. The solutes with highest error tended to have many carbon atoms and include at least one carboxylic acid site. Further tests showed this to also occur when comparing CombiSolv-QM and CombiSolv-Exp, which are separate datasets of COSMO-RS calculations and solvation energies for neutral solutes (see SI).[26] The higher errors in this negative solvation energy regime might be due to greater difficulty in accurately measuring experimental solvation energies for those compounds, or may be from greater errors in COSMO-RS. The relative contribution of each potential source of error is currently unclear to the authors.

Regardless of its origin, we attribute the higher ΔG^*_{solv}(AH) uncertainty in this work to a higher proportion of "challenging" solutes, especially carboxylic acids. Because the solutes in ΔG^*_{solv}(AH) are all acids, and many have solvation energies lower than -10 kcal mol^{-1} (see Figure 7), we assign our high uncertainty value of 1.57 kcal mol^{-1} as the standard uncertainty of ΔG^*_{solv}(AH).

ΔG^*_{solv}(**AH**) model

The DirectML GNN model by Chung *et al.* [27] is used later in this work to compute solvation free energies of neutral species. The Chung model was trained on the `dGsolvDB3` dataset, and the authors therein reported a 0.29 / 0.56 kcal mol^{-1} MAE / RMSE, randomly holding out 10% of the data as test data. Model errors were higher for substructure splitting (0.81 / 1.24 kcal mol^{-1}), indicating potential challenges with generalizing to new chemistries. We refer readers to the corresponding literature reference for further details.

Because numerous ΔG^*_{solv}(AH) models have been trained and evaluated on combinations of COSMO-RS solvation energies[26] and experimental data,[27] we did not train a ΔG^*_{solv}(AH) model.

$\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$ Data and Modeling

D2A-dGsolv-anion: Dataset Summary

We constructed D2A-dGsolv-anion by combining experimental $\mathrm{p}K_\mathrm{a}$ values, QM-computed $\Delta G^o_{\mathrm{acid,\ gas}}(\mathrm{AH})$ values, COSMO-RS values for $\Delta G^*_{\mathrm{solv}}(\mathrm{AH})$, and the solvation free energy of the proton in each solvent using Equation 1. This final dataset consists of 4,468 anionic solutes corresponding to 6,090 $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$ values across 8 solvents.

Figure 9 displays histograms of $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$. Each row corresponds to a specific solvent. Given the observed variations in data distributions across different solvents, and also the solvents' varying dielectric constants or solvating abilities and different anions' very different electric charge distributions, it is expected that the performance of the models trained on D2A-dGsolv-anion will depend significantly on the solvent.

The solvation free energies, which range from approximately -100 to -40 $\mathrm{kcal\,mol^{-1}}$, are concentrated between -90 and -50 $\mathrm{kcal\,mol^{-1}}$, with substantial variation depending on the solvent. Most of the distributions are unimodal and slightly right-skewed. Water is the most widely-represented solvent in this dataset. The molecular weight distributions follow the same patterns as observed for D2A-dGsolv-neutral.

Some of the variation in $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$ is due to different types of ions sampled; the solvation free energy of ions depends strongly on the ionization center. The ions represented in water are more diverse, leading to a longer left and right tail.

The relative centering of each distribution will change if different values are used for the solvation free energies of protons. In this work, we used the values reported in a critical evaluation by Marcus *et al.* ,[81] which include transfer free energies for the proton from water using the TATB assumption. We added those transfer free energies to the *absolute* intrinsic solvation free energy of the proton in water ($-254.3\,\mathrm{kcal\,mol^{-1}}$) as also reported by Marcus.[82] We include only solvents with "consensus" values as reported by Marcus: water, methanol, acetonitrile, dimethylsulfoxide (DMSO), N-methylpyrrolidone (NMP), ethanol,

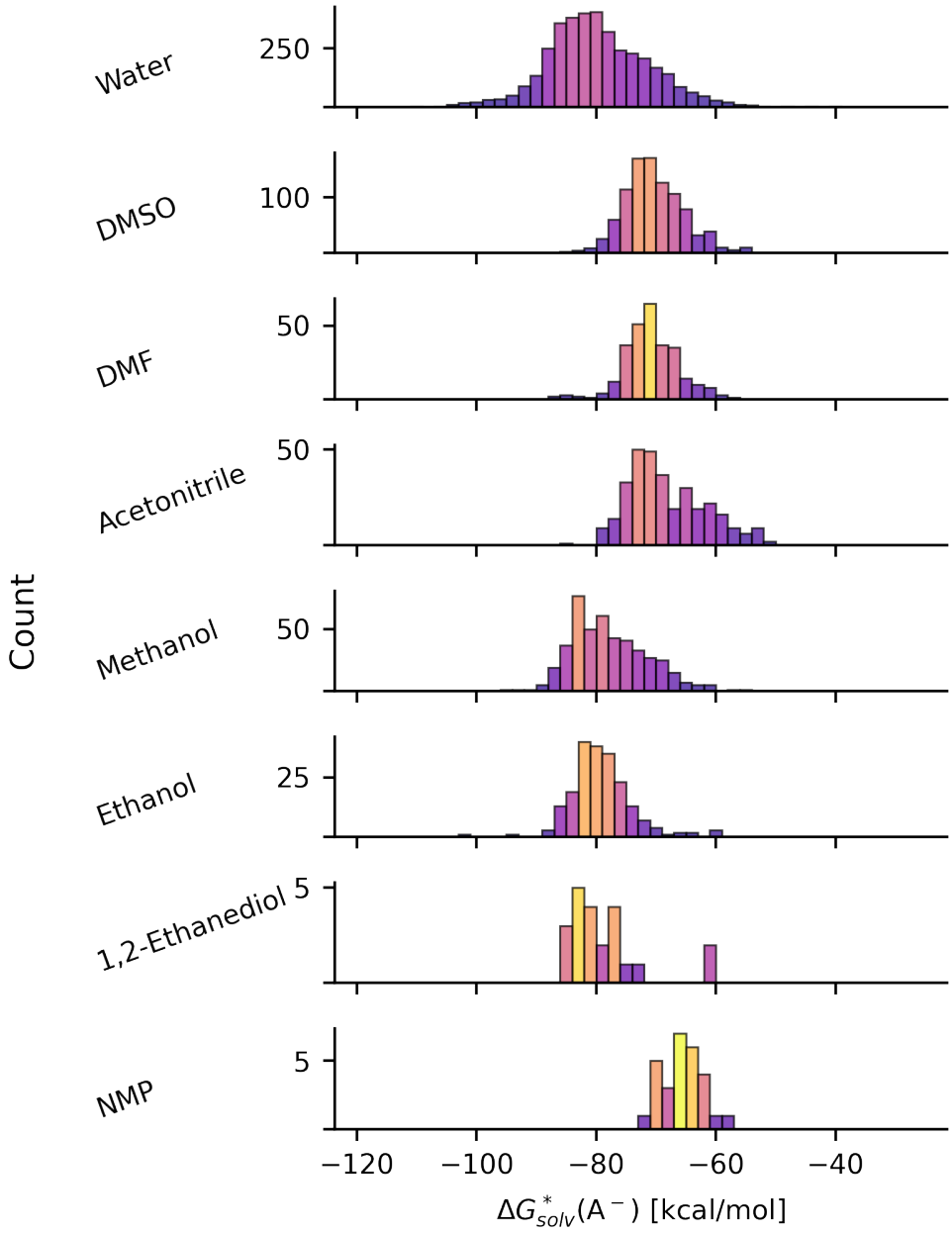

Figure 9: Distributions of solvation free energies of anions in D2A-dGsolv-anion. Brighter bars indicate higher density.

1,2-ethanediol, and dimethylformamide (DMF). The complete set of $\Delta G^*_{\mathrm{solv}}(\mathrm{H}^+)$ values used in this study is provided in Table 4.

Table 4: Transfer energies and *absolute* solvation free energies of the proton across different solvents. To calculate the *absolute* $\Delta G^*_{\mathrm{solv}}(\mathrm{H}^+)$ values, the Marcus reference value in water of $-254.3\,\mathrm{kcal\,mol^{-1}}$ [82] is used. All values are in kcal mol^{-1}.

IUPAC name	$\Delta G^*_{\mathrm{transfer}}(\mathrm{H}^+)$	$\Delta G^*_{\mathrm{solv}}(\mathrm{H}^+)$
Methanol	2.5	-251.8
Ethanol	2.7	-251.6
1,2-Ethanediol	1.2	-253.1
N-Methylpyrrolidone	-6.0	-260.3
Acetonitrile	11.0	-243.2
Dimethyl sulfoxide	-4.6	-258.9
Dimethyl formamide	-4.3	-258.6

Figure 10 shows the distribution of anion occurrences across different chemical classes within D2A-dGsolv-anion based on substructure matching (the proportions of the atom types can be found in SI). Carboxylates are by far the most common, with nearly 2,500 entries; the azanide and phenolate classes follow with around 1,000 occurrences each, while carbanions and alkoxides are moderately represented. Other classes, such as aromatic alkoxides and thiolate anions are less frequent. Only a few datapoints are included in classes such as non-carbon oxoacid anions, or peroxyacid anions. Similar to solvent distribution, the observed variations in number of training examples across different chemical classes suggest that models trained on D2A-dGsolv-anion will exhibit significant dependence on the chemical class.

D2A-dGsolv-anion: Data Uncertainty and Validation

To assess the uncertainty of values in D2A-dGsolv-anion, we analyze the propagation of uncertainties from individual sources. Here, we provide a summary of the key points for clarity.

For the pK_{a}, $\Delta G^o_{\mathrm{acid,\,gas}}(\mathrm{AH})$, and $\Delta G^*_{\mathrm{solv}}(\mathrm{AH})$ values, we adopt the uncertainties from literature studies, which are described above.

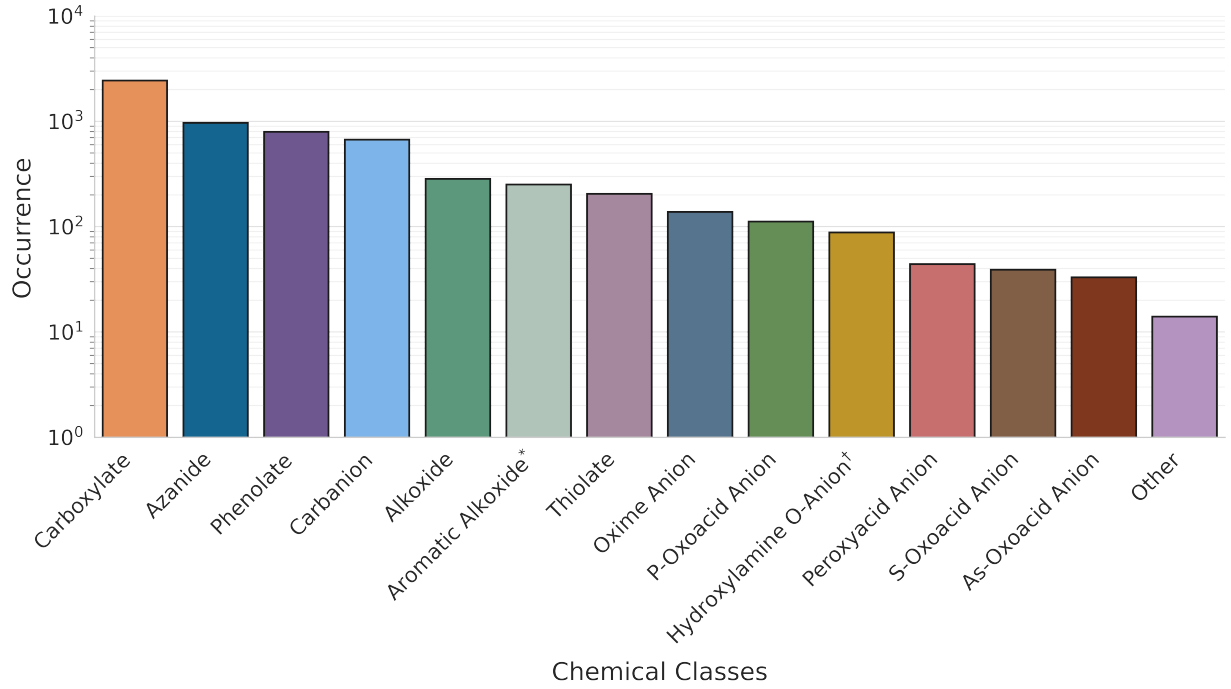

Figure 10: Distribution of chemical classes for D2A-dGsolv-anion. * refers to aromatic alkoxides excluding phenolates. † refers to hydroxylamine O-anions excluding oxime anions. Anion classes were determined by substructure matching SMARTS patterns using the RDKit package.

An unavoidable, large source of error is the proton's solvation free energy, $\Delta G^*_{\text{solv}}(\text{H}^+)$. This term anchors relative acidity scales to a single absolute energy scale, and is reliant on the choice of *extrathermodynamic assumption*. Uncertainties associated with this property are high, exceeding $2\,\text{kcal mol}^{-1}$ in non-aqueous solvents.[36,83,84] The uncertainty of the value in water is slightly lower, sometimes reported as under $1.5\,\text{kcal mol}^{-1}$.[36,83,84] However, any error associated with this term manifests as a systematic offset, which can be corrected in downstream modeling applications and which usually cancels out when computing the quantities of interest in an experiment.

The published uncertainty (1σ) in estimating $\Delta G^*_{\text{solv}}(\text{H}^+)$ is $1.7\,\text{kcal mol}^{-1}$ for aqueous[81] and $2.4\,\text{kcal mol}^{-1}$ for non-aqueous solvents.[36,83,84] As discussed previously, the proton solvation free energy error is systematic for each solvent and not random.[24,42]

Combining these analyses, the overall uncertainties for $\Delta G^*_{\text{solv}}(\text{A}^-)$ in D2A-dGsolv-anion are estimated to be $2.8\,\text{kcal mol}^{-1}$ in water and $3.3\,\text{kcal mol}^{-1}$ in non-aqueous solvents.

These uncertainties were calculated through uncertainty propagation analysis, which involves combining the individual uncertainty contributions from the sources using the root-sum-of-squares method, assuming that the uncertainty contributions are independent. The overall uncertainties can be interpreted as 1σ confidence intervals, which have similar coverage probability to RMSE.

Table 5: Propagation of 1σ uncertainties represented as RMSE, for *absolute* solvation free energies of anions in D2A-dGsolv-anion, incorporating individual contributions from Equation (1). All values are in kcal mol^{-1}.

	methodology	aqueous	non-aqueous
$\Delta G^o_{\text{acid, gas}}$	DLPNO-CCSD(T)/CBS	1.57	1.57
$\Delta G^*_{\text{acid, soln}}$	Expt. pK_a ref.[72]	0.23	0.41
$\Delta G^*_{\text{solv}}(\text{AH})$	COSMO-RS[26,80]	1.57	1.57
$\Delta G^{*,\text{Lit.}}_{\text{solv}}(\text{H}^+)$	Ref.[82]	1.7	2.4
$\Delta G^*_{\text{solv}}(\text{A}^-)$	(*absolute error*)	2.8	3.3

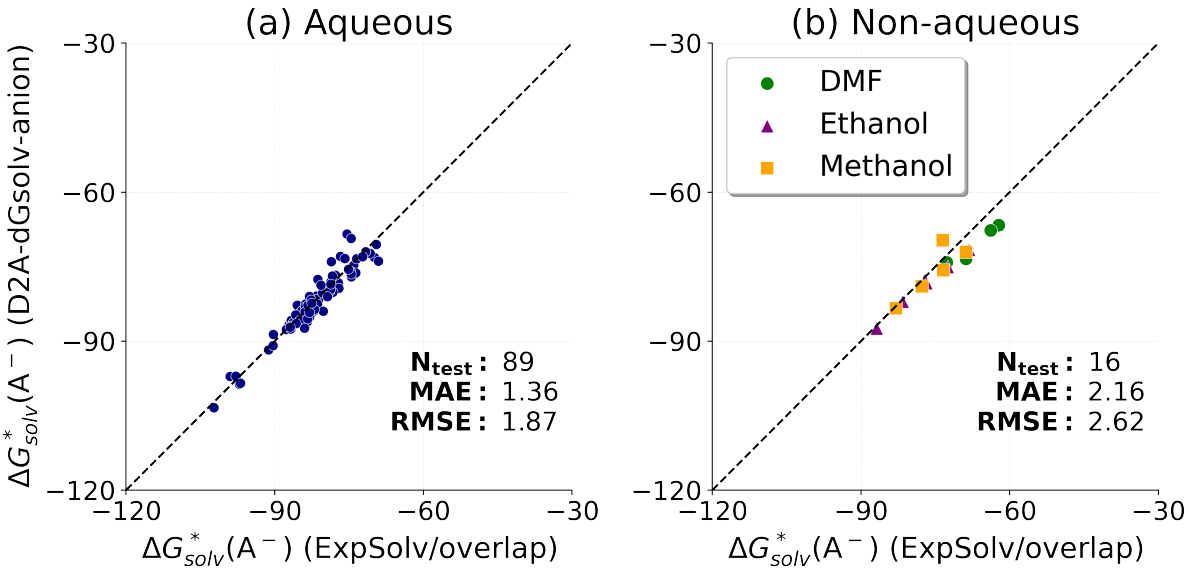

Figure 11: Comparison of some values in D2A-dGsolv-anion with those in experimental dataset ExpSolv/overlap for (a) aqueous and (b) non-aqueous values. The summary statistics shown in (b) represent the entire set of non-aqueous values. Because the $\Delta G^o_{\text{acid, gas}}$(AH) term comes from NIST data, the values have been aligned to the QM scale by -1.2 kcal mol^{-1}. After shifting, the individual values for each solvent are (MAE / RMSE): 1.33 / 1.72 for ethanol, 2.21 / 2.53 for methanol, and 3.56 / 3.78 for DMF. All values are in kcal mol^{-1}.

To further assess the quality of data in D2A-dGsolv-anion, we compiled a small set of anionic solvation free energies composed only from experimental data. We obtained ΔG^*_{solv}(AH) values from dGsolvDB3/overlap[27] and $\Delta G^o_{\text{acid, gas}}$(AH) values from NIST/overlap,[51] shifting the $\Delta G^o_{\text{acid, gas}}$(AH) values by 1.2 kcal mol^{-1} to align with the QM calculations. We then used these data to compose a set of solvation free energies of anions (ExpSolv/overlap). We compared D2A-dGsolv-anion data to ExpSolv/overlap, comprised of 89 anions in water, 4 in DMF, 5 in methanol, and 7 in ethanol (Figure 11). We use the same values of $\Delta G^*_{\text{solv}}(H^+)$ and pK_a in both sets, so the remaining disagreement is from differences in $\Delta G^o_{\text{acid, gas}}$(AH) and ΔG^*_{solv}(AH).

Much of the deviation between D2A-dGsolv-anion and ExpSolv/overlap are from different values for gas-phase acidities. The MAE/RMSE between the QM-computed and experimental gas-phase acidities alone in this subset is 1.02/1.64 kcal mol^{-1}, which is consistent with the

errors of 1.33 / 1.57 kcal mol^{-1} from comparing NIST/overlap with D2A-dGgas (see Figure 5) as well as the higher uncertainty of the gas-phase calculations (Table 5).

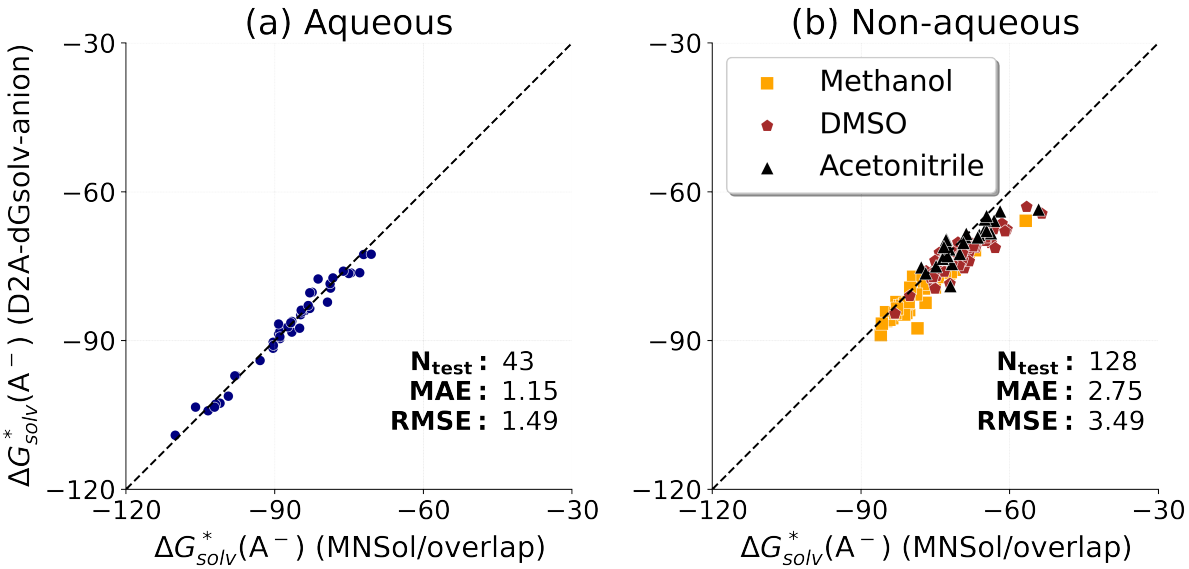

Figure 12: Comparison of some values in D2A-dGsolv-anion with those in experimental dataset MNSol/overlap for (a) aqueous and (b) non-aqueous values. The summary statistics shown in (b) represent the entire set of non-aqueous values. The values have been systematically aligned to match the $\Delta G^*_{\text{solv}}(H^+)$ values we used. Because the $\Delta G^o_{\text{acid, gas}}(AH)$ term comes from NIST data, the values have been aligned to the QM scale by -1.2 kcal mol^{-1}. After shifting, the individual values for each solvent are (MAE / RMSE): 2.49 / 3.10 for methanol, 3.32 / 4.02 for DMSO, and 2.11 / 3.01 for acetonitrile. All values are in kcal mol^{-1}.

We also repeated this analysis with the values in MNSol/overlap (see Figure 12). We observe similar test statistics: an overall lower RMSE for water and higher RMSE in non-aqueous solvents methanol, DMSO, and acetonitrile. Unfortunately, the constituent values for $\Delta G^*_{\text{solv}}(A^-)$ are not available in MNSol, and so we could not assess their disagreement with the D2A-dGsolv-anion values.

The D2A-dGsolv-anion uncertainties at first glance seem to match the expected uncertainties of the $\Delta G^*_{\text{solv}}(A^-)$ data as detailed previously. However, for comparison to "experimental" data, we must exclude consideration of systematic errors because the same value of $\Delta G^*_{\text{solv}}(H^+)$ is used in both datasets, and we assume that systematic offset in $\Delta G^o_{\text{acid, gas}}(AH)$ is captured in the gas-phase correction term. The remaining non-systematic uncertainties

(that is, from the $\Delta G^*_{\mathrm{solv}}(\mathrm{AH})$ calculations, $\mathrm{p}K_{\mathrm{a}}$, and *relative* error in $\Delta G^o_{\mathrm{acid,\,gas}}(\mathrm{AH})$) in D2A-dGsolv-anion is approximately 1.6 kcal mol^{-1} across all solvents. The RMSE values in Figures 11 and 12 are higher in non-aqueous solvents than 1.6 kcal mol^{-1}, so the actual uncertainties of the non-aqueous data in D2A-dGsolv-anion are higher than expected. Considering systematic errors, this would suggest that the actual uncertainties of the non-aqueous $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$ values are closer to 4 kcal mol^{-1}. Future improvements could arise from using more accurate QM methods for both $\Delta G^o_{\mathrm{acid,\,gas}}(\mathrm{AH})$ and $\Delta G^*_{\mathrm{solv}}(\mathrm{AH})$.

$\Delta G^*_{\mathrm{solv}}(\mathbf{A}^-)$ direct model

We investigated the prediction for the solvation free energies of anions using two different approaches. This approach directly trains a GNN model on D2A-dGsolv-anion/train for predicting $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$.

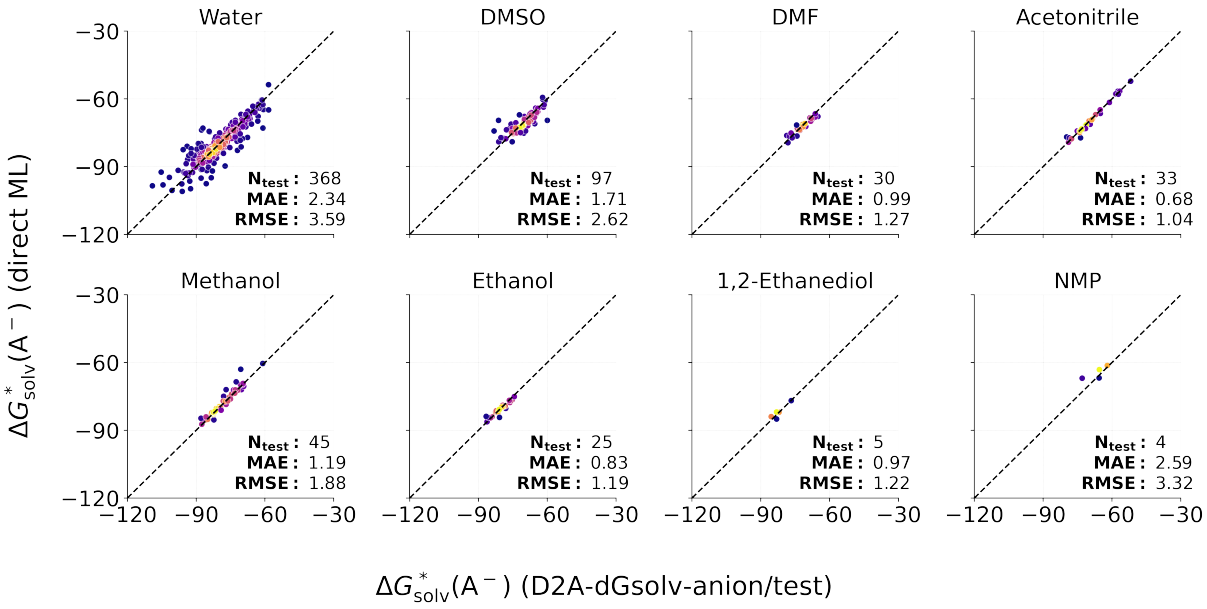

Figure 13: Performance of D-MPNN/FFNN model on the held-out test data (units in kcal mol^{-1}). Model was trained directly on anion solvation free energies.

Figure 13 shows the performance of the D-MPNN/FFNN model, which was trained directly on D2A-dGsolv-anion/train to directly predict $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$. The model's predictions are validated on D2A-dGsolv-anion/test, composed of data that are not seen during the

training process. The overall MAE / RMSE was 1.93 / 3.08 kcal mol^{-1}.

In water, which is the most significant solvent in the dataset, the model predicts with an RMSE of 3.59 kcal mol^{-1}, demonstrating reasonable accuracy for most organic anions tested, but having the tendency to vastly over- or underpredict for a small set of outliers. Such errors tended to be higher in the more negative solvation free energy range. We observed that errors are higher for peroxyacid anions and hydroxylamine O-anions, whose solvation free energy distributions are centered lower than other types of anions. Those types of ions are far less represented in non-aqueous solvents (see SI) and only make up a small portion of the training dataset. Hence, some of the higher error may represent the ion-type dependence of the model, in which the water test set is more challenging, because of a more diverse set of molecules. The ability to further understand ion-type dependence is limited in large part to the lower availability of data in other solvents.

Although carboxylates are widely represented in the dataset, the model errors in water for such species are around 4 kcal mol^{-1}. Upon further investigation of the data, we found that the MAE for carboxylates with only one carboxylic acid group was just 1.6 kcal mol^{-1}, whereas the error for compounds with two or more carboxyl groups was 6.1 kcal mol^{-1}. Errors are similar for COSMO-RS calculations, which are discussed in the SI; the MAE for those classes above are 2.3 kcal mol^{-1} and 5.0 kcal mol^{-1}, suggesting that these high errors are potentially due to issues with the data or QM calculations rather than the machine learning model. It is unclear to the authors at the moment why exactly model performance is significantly worse for dicarboxylic acids compared to monocarboxylic acids, given that only dissociation of the first proton was considered, so that the acids essentially behaved as monoprotic acids.

For non-aqueous solvents like methanol and DMSO, the model performs slightly better, with RMSEs of 1.88 kcal mol^{-1} and 2.62 kcal mol^{-1}, respectively. In solvents such as ethanol, acetonitrile, and DMF, the RMSE is even lower - less than 1.5 kcal mol^{-1}, representing good performance on the unseen solutes. Part of this apparent performance improvement is due to

the lower diversity of compound types in such solvents, leading to less challenging compounds in the test sets. We do not know how accurately these models predict other types of anions in these solvents, because we lack relevant data. In general, we recommend caution when using this model for values in solvents trained on only a few datapoints, including NMP and 1,2-ethanediol.

Similar trends (poor performance in water, better performance in solvents where the represented acids are less diverse) were also observed using the COSMO-RS method; in fact, the relative ordering of which solvent predictions had higher error is nearly identical (see SI). COSMO-RS is driven by quantum mechanics including some parameterization on data; hence, errors from the COSMO-RS method tend to be based in physical limitations (e.g. more challenging quantum-chemical systems to model), whereas our ML model is purely data-driven. Therefore, it is somewhat unexpected that the errors would be so similar. With the exception of 1,2-ethanediol, the RMSE of the direct ML predictions are lower than those of the COSMO-RS predictions, showing overall better performance of the ML method.

To further test the GNN models, we test against the values in ExpSolv/nonoverlap dataset, a compilation of $\Delta G^*_{\text{solv}}(\text{A}^-)$ values that only uses experimental data for the thermodynamics (i.e., no properties from quantum chemical calculations). In ExpSolv/nonoverlap, the $\Delta G^o_{\text{acid, gas}}(\text{AH})$ values were obtained from NIST/nonoverlap[51] and the neutral solvation free energies are from dGsolvDB3/nonoverlap.[27] The pK_a data in both ExpSolv/nonoverlap and D2A-dGsolv-anion/test are the same (from D2A-pKa), and hence there is no error contribution associated with those pK_a values. This test set includes 8 values, all in water.

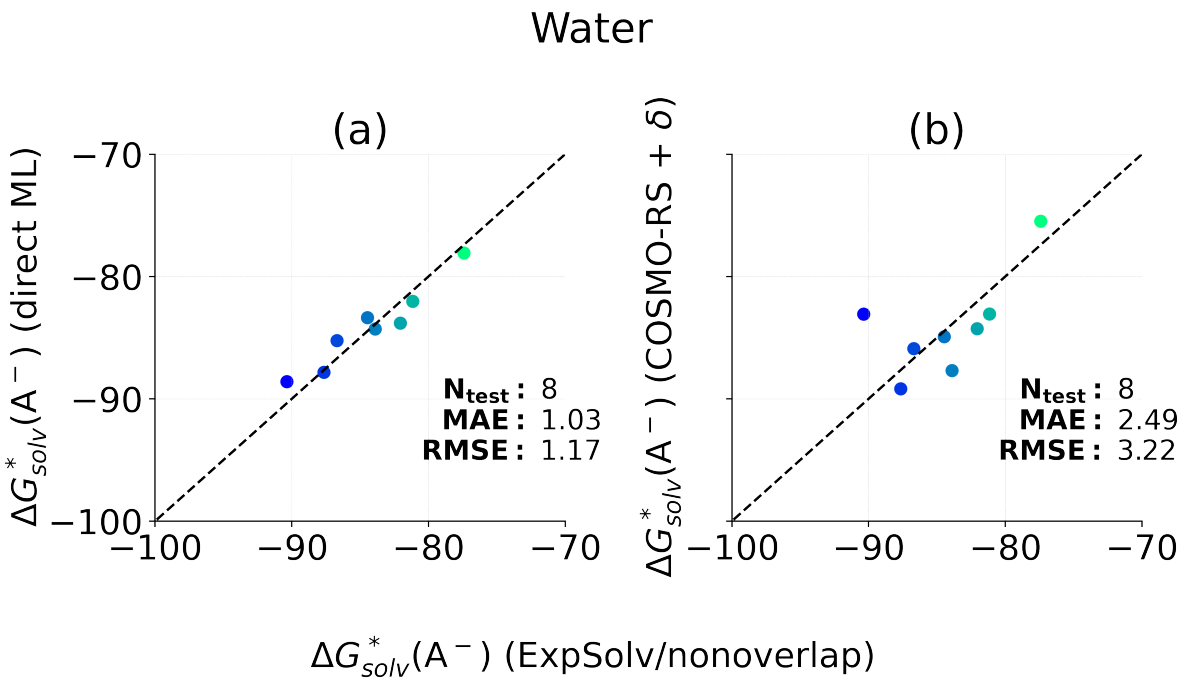

Figure 14: Comparison of (a) direct prediction method and (b) COSMO-RS method to ExpSolv/nonoverlap (all units in kcal mol^{-1}). Points with the same shading correspond to the same experimental data. Because the $\Delta G^o_{\text{acid, gas}}$(AH) term comes from NIST data, the values have been aligned to the QM scale by -1.2 kcal mol^{-1}. δ refers to the parameter that systematically shifts COSMO-RS calculations to match the experimental data.

Figure 14a depicts good performance of the model on ExpSolv/nonoverlap, with an RMSE of 1.17 kcal mol^{-1} in water. The errors are slightly lower than the uncertainties of $\Delta G^*_{\text{solv}}(\text{A}^-)$ shown in Figure 11 (RMSE values of 1.49 kcal mol^{-1} in water, comparing D2A-dGsolv-anion to ExpSolv/overlap). As shown in Figure 14b, the COSMO-RS method again performs worse than the direct ML model. Note that the COSMO-RS values were shifted by a parameter δ that minimizes the error of the COSMO-RS errors against ExpSolv/overlap data, to account for systematic misalignments in energy scales.

We further tested the ML model on datasets with existing quantum-chemical calculations, so we could benchmark our method against QM methods. In particular, we were interested in comparing our results to cluster-continuum quantum chemistry approaches, which have been used in recent years to compute solvation free energies of ionic solutes.[20,23] To further benchmark our model, we obtained a set of molecules with experimental values and hybrid cluster-

Table 6: Comparison of hybrid cluster-continuum method and direct ML method for solvation energies of anions (kcal mol^{-1})

Compound	**Cluster-continuum method error**	**Direct ML error**
Fluoride anion	**-1.1**	21.3
Chloride anion	-1.5	**-1.0**
Bromide anion	**-1.0**	-5.1
Iodide anion	**-1.3**	-9.7
Nitrate anion	3.3	**0.4**
Indolate	4.2	**-2.2**
Tricyanomethanide	**2.3**	-3.4

continuum calculations of solvation free energies in acetonitrile (Pliego/nonoverlap).[56] The seven solutes are not in our training corpus (even including in non-acetonitrile solvents). Four of the compounds are hydrogen halides (not well-represented in our training data), one is nitric acid, and the rest are organic acids. We used the direct ML model to predict $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$. The results are summarized in Table 6.

The results indicate that generally the cluster-continuum method is more reliable, especially for out-of-domain species. Halide anions do not appear in the training corpus, and the element iodine does not appear at all. On the other hand, the model does perform quite well for the organic compounds and the nitrate anion, indicating that the model performs comparably to solvation models when the solutes are anions of relatively simple organic acids.

Table 7: Errors of different solvation models (CRS: COSMO-RS and CC: Cluster-continuum) in prediction solvation free energies of anions (kcal mol^{-1}) for individual predictions as well as MAE/RMSE metrics for all anions in a solvent. Lowest errors are bolded.

Anion's conjugate acid	**Solvent**	**CRS Err.**	**CC Err.**	**Direct ML Err.**
Acetonitrile	Water	-2.5	-3.4	**1.7**
Chloroacetic acid	Water	**-0.2**	-0.8	3.8
Chloroform	Water	**2.1**	-2.1	-12.6
Dimethyl sulfoxide	Water	-7.6	-7.4	**-5.3**
MAE/RMSE (Water)		**3.1**/**4.1**	3.4/4.2	5.8/7.1
2-nitrobenzoic acid	Acetonitrile	-1.7	**0.5**	5.8
4-hydroxybenzoic acid	Acetonitrile	-5.4	-6.1	**1.9**
Benzenesulfonamide	Acetonitrile	**0.6**	2.0	9.1
MAE/RMSE (Acetonitrile)		**2.6**/**3.3**	2.9/3.7	5.6/6.3
1,2,3-triazole	DMSO	**-0.2**	1.2	5.6
1,2,4-triazole	DMSO	**1.4**	1.7	6.8
2-nitrobenzoic acid	DMSO	-3.0	-3.5	**-0.1**
Adenine	DMSO	0.9	**0.8**	3.5
Benzenesulfonamide	DMSO	-1.9	**-1.2**	1.9
Carbazole	DMSO	-1.7	**-0.8**	1.4
Dimethyl sulfoxide	DMSO	**0.1**	1.7	3.8
Imidazole	DMSO	**1.3**	2.0	6.5
Pyrazole	DMSO	-2.5	**-2.4**	3.2
Tetrazole	DMSO	-1.2	**-0.6**	5.5
Water	DMSO	7.0	**6.6**	22.5
MAE/RMSE (DMSO)		**1.9**/2.7	2.0/**2.6**	5.5/8.0
2-fluorobenzoic acid	Methanol	**0.1**	0.4	0.3
2-nitrobenzoic acid	Methanol	-3.1	-3.0	**-2.9**
3-chlorophenol	Methanol	3.8	**1.8**	1.9
3-methoxybenzoic acid	Methanol	-1.8	-1.5	**-0.3**
3-methylbenzoic acid	Methanol	-1.3	**-0.1**	-0.5
4-hydroxybenzoic acid	Methanol	-8.1	-7.3	**-6.7**
Chloroacetic acid	Methanol	**-0.3**	0.9	1.1
Cyclobutanecarboxylic acid	Methanol	-0.3	**0.2**	1.1
MAE/RMSE (Methanol)		2.4/3.5	**1.9**/2.9	**1.9**/**2.7**

To further test the models, we compared the direct ML model with 26 anions in the Minnesota Solvation Database[53] (MNSol/nonoverlap) along with previous COSMO-RS cal-

culations at the BP-TZVPD-FINE level and and cluster-continuum calculations, shown in Table 7.[24] We aligned the proton solvation free energies to allow for fair comparison, and adjusted the experimental $\Delta G^*_{\text{solv}}(\text{A}^-)$ values by -1.2 for the MNSol/overlap data because their gas-phase acidities came from NIST. In methanol, the direct ML model performs about as well if not slightly better than the other methods. The solutes in this set are all simple organic acids' conjugate anions, similar to the solutes that appear in the data used to train the model. In the other solvents, the ML model overall does not perform well. Individual predictions tend to be either very accurate (successful interpolation) or very inaccurate (high errors from extrapolation). For instance, in water, two of the lowest errors are from the direct ML model, but several outliers are also present including one that exceeds 12 kcal mol^{-1}, showing the variability of the model depending on provided inputs. The solutes with high errors tend to be exceptional compounds - such as the anion conjugates of chloroform, water, and DMSO. The model also generally perform poorly with azole anions in DMSO.

In summary, the direct ML model's performance strongly depends on the class of solute - in some cases, model performance is comparable to the cluster-continuum and COSMO-RS solvation models (see SI for errors by anion type). However, the unsatisfactory ability of the model to extrapolate precludes it from modeling more general solvation phenomena. We therefore suggest using the model only for small organic acids, or in scenarios where several kcal mol^{-1} error are tolerable; for instance, ranking relative energies or in screening candidate compounds. Despite the model's apparent shortcomings, it still provides a very fast way of approximating $\Delta G^*_{\text{solv}}(\text{A}^-)$. Finetuning on additional data may improve the performance of future models.

$\Delta G^*_{\text{solv}}(\mathbf{A}^-)$ composite model

The composite model approach trains separate GNN models on D2A-dGgas/train and D2A-pKa/train, and utilizes Chung's $\Delta G^*_{\text{solv}}(\text{AH})$ model. These pK_{a}, $\Delta G^o_{\text{acid, gas}}(\text{AH})$, and $\Delta G^*_{\text{solv}}(\text{AH})$ model predictions are then combined using Equation 1 to predict $\Delta G^*_{\text{solv}}(\text{A}^-)$. Because the

underlying data do not need to contain the same solutes, this method could take advantage of larger data corpuses during training. For instance, in this work, the pK_a dataset is larger than the $\Delta G^o_{\text{acid, gas}}$(AH) dataset. Alternatively, individual parts of each model could be swapped out with higher-accuracy QM methods.

Figure 15 displays the performance of the composite prediction method, evaluated on D2A-dGsolv-anion/test.

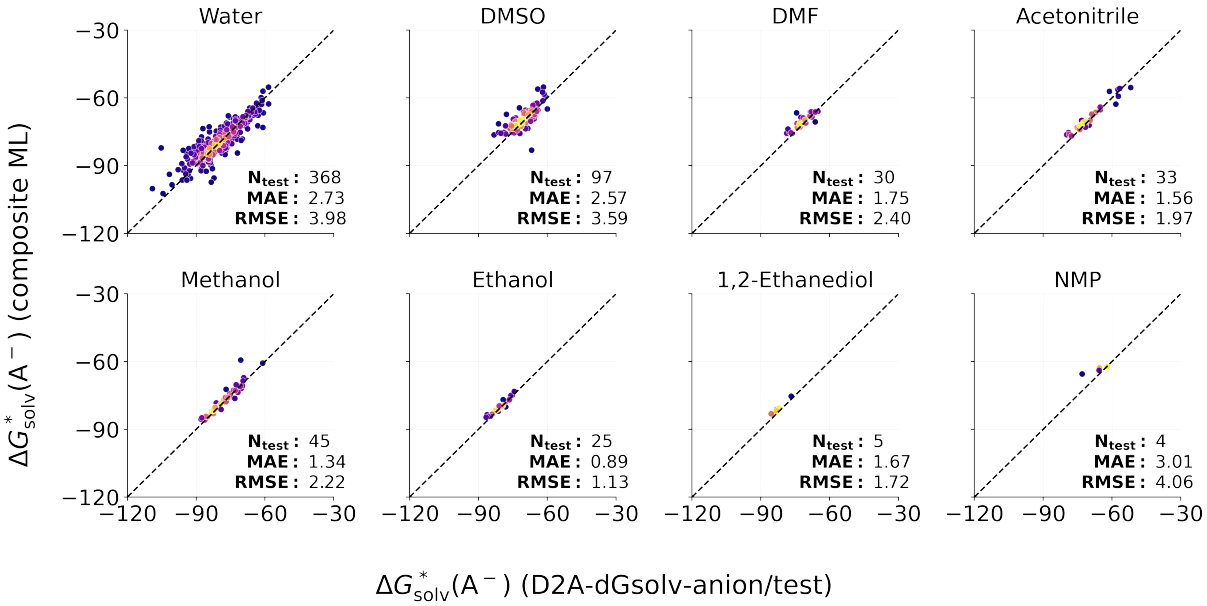

Figure 15: Performance of D-MPNN/FFNN model trained on separate components of thermochemical cycle (all units in kcal per mol). Brighter spots indicate higher density.

In all cases except for in ethanol, the composite prediction model performs worse than the direct prediction model. For water, a RMSE of 3.98 kcal mol^{-1} is observed, which is higher than the 3.59 kcal mol^{-1} achieved by the direct prediction model. The errors are high across all solvents, but for the most part follow the ranking of accuracies as seen for direct prediction method (Figure 13).

One benefit of this method is that the sources of error can be examined. We examined the difference between the individual model predictions and the data. Table 8 summarizes the main sources of error by property. The gas-phase model introduces error consistent with the benchmarks observed in this study. The solvation free energy predictions by Chung et

Table 8: Comparison of property errors for D-MPNN predictions and values used in the composite model for D2A-dGsolv-anion/test (kcal mol^{-1} for energies). The pK_a errors associated with the solution-phase acid dissociation are 0.58 and 1.07 for MAE and RMSE, respectively.

Property	Data type	Model MAE	Model RMSE
$\Delta G^{o}_{\text{acid, gas}}$(AH)	QM	1.98	3.28
$\Delta G^{*}_{\text{solv}}$(AH)	COSMO-RS	1.53	2.20
$\Delta G^{*}_{\text{acid, soln}}$(AH)	Experimental	0.80	1.46
$\Delta G^{*}_{\text{solv}}(A^{-})$	–	2.41	3.56

al.'s D-MPNN model contribute to significant error, despite reported MAEs of less than 1 kcal mol^{-1} in the original manuscript.[27] These high errors are consistent with our benchmarks that showed high deviation (1.09 / 1.57 kcal mol^{-1}) between D2A-dGsolv-neutral and dGsolvDB3/overlap.

Although the prediction quality is overall worse, the composite prediction method has the benefit of modularity. The models could be switched out for more accurate machine learning models developed in the future, or even with QM calculations if a higher level of accuracy is desired. There are numerous combinations of models for different properties that can be chosen, which future work should investigate further. In particular, using an accurate solvation method to compute $\Delta G^{*}_{\text{solv}}$(AH) with the other ML models might result in model performance comparable or better than the direct method. We therefore see potential use cases for both the direct and composite prediction methods.

Summary of modeling results

In this work, we have examined the *uncertainty* in the data and the *error* in several different models, all evaluated against many test sets. To reduce the burden for the reader, we summarize the results in tables.

Table 1 summarizes the data used to evaluate data uncertainty, and Table 5 reports the evaluated uncertainties. In contrast, Table 2 summarizes the data splits and external

Table 9: Overview of modeling results, with MAEs and RMSEs reported for each test set. Units for free energies are in kcal mol^{-1}, and for pK_a are dimensionless.

Property	Dataset name	Description	MAE	RMSE
pK_a	D2A-pKa/test	Test split	0.58	1.07
	SAMPL6 acids	SAMPL6 challenge, aqueous, acids[54]	0.63	0.96
	SAMPL7	SAMPL7 challenge, aqueous[55]	0.59	0.78
	Zheng/nonoverlap	Challenge test set in Zheng et al., 2024[45]	1.00	1.07
$\Delta G^o_{\text{acid, gas}}$(AH)	D2A-dGgas/test	Test split	2.35	3.73
	NIST/nonoverlap	NIST Chemistry Webbook[51,52]	2.09	3.54
ΔG^*_{solv}(AH)	–	–	–	–
ΔG^*_{solv}(A$^-$)	D2A-dGsolv-anion/test	Test split	1.93	3.08
	D2A-dGsolv-anion/test-composite	Test split with composite model	2.41	3.56
	ExpSolv/nonoverlap	Compilation from literature[24,27,52]	1.03	1.17
	Pliego/nonoverlap	Data[41] with QM calculations[56]	6.16	9.18
	MNSol/nonoverlap	Minnesota Solvation Database[53] with QM calculations[24]	4.46	6.45

test sets used to evaluate the models. Table 9, shown in this section, reports the model performances.

Across the $\Delta G^o_{\text{acid, gas}}$(AH) and ΔG^*_{solv}(A$^-$) models, we observed a poor ability of the models to extrapolate to unfamiliar chemistries. We also found that in general, the models had higher errors for smaller molecules, possibly because differences between molecules with few heavy atoms are more likely to arise from differences in acidity centers and functional groups (which correspond strongly with ΔG^*_{solv}(A$^-$)) rather than side groups, branching, and similar features; see SI for more details.

Considering the high potential for error, it may be useful for users to leverage uncertainty estimation techniques to assess the reliability of each prediction. Information about using ensemble variance to estimate error can be found in the SI.

Conclusions

In this study, we developed the DISSOLVE2-ANIONS set of databases. The database of ΔG^*_{solv}(A$^-$) values, D2A-dGsolv-anion, comprises of 6,090 entries focused primarily on the room-temperature solvation free energies of anions across the solvents water, methanol, acetonitrile, DMSO, ethanol, NMP, 1,2-ethanediol, and DMF. To construct this dataset, we curated two other large subsets of thermodynamic data: 5,536 QM calculations of gas-phase

acidities (D2A-dGgas) and 8,241 pK_a values (D2A-pKa) across those same 8 solvents. These D2A-dGgas is the first corpus of $\Delta G^o_{\text{acid, gas}}(\text{AH})$ values provided digitally at a large scale with no limitations on usage. The pK_a dataset introduced herein is among the first to introduce a rigorous cleaning process, which improves the reliability of the data for modeling purposes. We also provided a dataset of solvation free energies of neutral acids from the COSMO-RS method.

Using these datasets, we trained D-MPNN models to predict each of those properties. The pK_a and gas-phase acidity models take as input the reaction SMILES corresponding to the neutral acid losing a proton to form the anion. The $\Delta G^*_{\text{solv}}(\text{A}^-)$ model takes the anion's SMILES as input. The direct prediction method, which involves training the model directly on $\Delta G^*_{\text{solv}}(\text{A}^-)$, generally outperformed the composite prediction method by a significant margin. This method achieved a lower RMSE across most solvents, particularly excelling in common solvents like ethanol and acetonitrile when evaluated on D2A-dGsolv-anion/test. Predictions were worse when compared to other data sources, often larger than in the assessed uncertainties D2A-dGsolv-anion, suggesting that further model improvements should come from improving both the quantity of training data as well as the accuracy of the underlying computations. The model predictions are sensitive to the similarity of the target compound and its similarity to the training sets, and model accuracy for $\Delta G^*_{\text{solv}}(\text{A}^-)$ is mixed. However, we believe the $\Delta G^*_{\text{solv}}(\text{A}^-)$ model could still provide a useful alternative to solvation models in cases where error up to $3\,\text{kcal}\,\text{mol}^{-1}$ on average is tolerable, especially for small, simple organic acids.

We emphasize again that such model predictions are dependent on both solvent and ionization center. Across all models, we observed better performance for phenolates, carboxylates (whose conjugate acids are monocarboxylic acids), and thiolates, with worse performance for other types of ions. We hence urge caution when using the models for wide varieties of chemistries.

In contrast, the composite prediction method decomposes the solvation free energy of

an anion into its component parts - $\Delta G^{o}_{\text{acid, gas}}(\text{AH})$, $\text{p}K_\text{a}$, $\Delta G^{*}_{\text{solv}}(\text{H}^{+})$, and $\Delta G^{*}_{\text{solv}}(\text{AH})$ - before summing these components using a thermodynamic cycle. The method introduced additional sources of error, but its individual thermodynamic components can be examined to assess possible sources of error, or swapped out with more accurate methods. Therefore, the composite method may be a suitable method when insufficient data is available to compose data for the direct method.

In addition to solvation free energies of anions, we also developed models to predict $\text{p}K_\text{a}$ values and gas-phase acidities. The $\text{p}K_\text{a}$ model demonstrated good performance, achieving an overall MAE of 0.58 $\text{p}K_\text{a}$ units on the full unseen test set. Moreover, when compared to other state-of-the-art multi-solvent $\text{p}K_\text{a}$ models from the literature, our $\text{p}K_\text{a}$ model showed competitive results in the SAMPL6 challenge and outperformed others in the SAMPL7 challenge. This good performance, despite a smaller training dataset, highlights the potential benefits of cleaning and curating data. Conversely, while the gas-phase acidity model provided reasonable predictions, it exhibited higher error margins, particularly when validated against experimental data.

All of the models developed in this study are freely available on Zenodo and can be used with the Chemprop package version $\geq$2.0.

There are potential limitations to these models. For one, we assumed that the $\text{p}K_\text{a}$ data are microscopic, despite the fact that most (if not all) data come from *macroscopic* measurements. This does not become an issue for monoprotic acids, or acids with acidity centers of distinctly different acid strengths, but can be problematic for polyacids and species with multiple thermodynamically relevant protomers. Furthermore, we made no corrections regarding the acidity centers of the iBonD data, and so any errors in transcription may appear herein. Such issues may also limit the accuracy of our evaluations against experimental data, though any such errors would make the model appear worse than it is. The models also are only trained on single tautomeric forms, and may provide poor results when provided with alternate tautomeric forms of a species.

The thermodynamics in this work are defined at the infinite dilution limit. However, in many practical applications, bulk thermodynamics are also influenced by counterions including through effects on solute activity and ion pairing. The values of $\Delta G^*_{\mathrm{solv}}(\mathrm{A}^-)$, though usable in theoretical studies and dilute systems, may not accurately capture the energetics of concentrated electrolyte systems. Using models such as COSMO-RS-ES[85] which explicitly account for electrolytic interactions may help, though ultimately collecting additional experimental data at high salt concentrations over different temperature ranges is still required.

Our model errors also are higher than the assessed aleatoric uncertainties in the experimental data, suggesting that further improvements in data quantity and QM accuracy will lead to further improvements.

We also used values of $\Delta G^*_{\mathrm{solv}}(\mathrm{H}^+)$ to anchor the solvation free energies, but the preferred value in each solvent is not yet clear in the literature. Different values of the proton solvation free energies are often found in the literature.[41,42,53] Model users should take care to subtract the difference between the desired proton solvation energy scales and the values used in this work.

Looking forward, future efforts in this area should focus on improving the $\Delta G^o_{\mathrm{acid,\ gas}}(\mathrm{AH})$ and $\Delta G^*_{\mathrm{solv}}(\mathrm{AH})$ models by incorporating additional molecular features and expanding the training datasets with more high-quality experimental data or calculations. Additionally, extending models to cover new families of acids and a broader range of solvents will be essential for broadening their applicability. We hope to release a curated dataset for singly-charged cations in the future, which will further enhance the predictive power and utility of these models across diverse chemical environments.

Methods

pK_a data curation details

We start by collecting pK_a values from the iBonD database. Although the database is accessible online, its format is not ideal for data science purposes, requiring additional steps to make the data usable for our analysis.

To address this, we employed a Python script to extract the species' chemical names, along with their corresponding pK_a values, solvent details, and publication references from the iBonD database. This information was then connected with images from iBonD indicating the molecules' dissociation sites (see Figure 1). We assume here that the acidity sites identified by the iBonD curators are correct, and that the values curated (most of which are macroscopic constants) can be expressed as *microscopic* values (i.e. that the ionization centers of the acid are sufficiently distinct so that the experimentally-obtained pK_a value can be unambiguously assigned to a specific tautomer[54]).

In this work, we focus only on anionic solutes with -1 charge. To ensure the dataset's integrity, we focus solely on stable solutes, excluding radicals, deuterated molecules, and species with charge states other than 0 and -1.

In our curation process, we focused on thoroughly reviewing the pK_a values for the acids (and thereby solvation free energies of anions) across several key solvents: water, methanol, N-methylpyrrolidin-2-one (NMP), ethanol, 1,2-ethanediol, DMSO, acetonitrile, and dimethyl formamide (DMF). These solvents are widely used and have "consensus" proton transfer free energies.[81] We cross-referenced the pK_a values for the same compounds across different sources and solvents to identify any discrepancies. Compounds with significant pK_a variance (greater than 0.2 and 0.5 pK_a units in water and organic solvents, respectively) between different data sources were flagged, and in cases where the variance was smaller, the values were averaged. Any data points that were deemed questionable, outside the reliable measurement range, or likely incorrect were excluded from the final dataset. Throughout

this process, we considered special cases where solvent or substituent effects might cause outliers that do not fit typical correlations.

At this stage, the data remain unsuitable for direct application due to the necessity of converting the IUPAC names of the solutes to molecular graphs (e.g., SMILES strings). To address the conversion from IUPAC names to SMILES of the acids, we use the MoleculeResolver,[86] a Python package that converts IUPAC names to SMILES by crosschecking different chemical repositories as well as by using the rules-based OPSIN algorithm. We manually reviewed any molecules where MoleculeResolver was unsuccessful, and either attempted to convert them to SMILES strings or excluded them from the dataset if not possible. To identify the acidity center for neutral acids, which was then used to derive the SMILES for anions, we used a tailored Python tool designed to process images from the iBonD database alongside the neutral SMILES of the acids. This Python tool renders the 2D structures of solutes in both their neutral and charged states, enabling users to precisely identify and select the dissociation sites for the generation of the anion's SMILES. This process was repeated across the entire iBonD database to ensure that the generated SMILES matched the deprotonation sites indicated in iBonD, thereby maintaining the integrity of the data. As a result, the final dataset contains SMILES representation for both neutral and anionic species, making it convenient for use.

We note that a recent study by An and collaborators[75] has also provided SMILES and values from the iBonD database, including atom indices of dissociation sites, while applying specific curation steps such as removing compounds with significant $\mathrm{p}K_\mathrm{a}$ variance between sources and averaging datapoints where the variance was smaller. Our approach is similar, but DISSOLVE2-ANIONS includes both neutral and anionic SMILES structures, and implements an extensive data cleaning process for entries with high variance. Both datasets are suitable for $\mathrm{p}K_\mathrm{a}$ prediction, though we recommend using our curated set for the nine solvents described herein due to its more rigorous cleaning.

Among the organic solvents, the amounts of $\mathrm{p}K_\mathrm{a}$ data in the solvents acetonitrile, DMSO,

and DMF from iBonD are among the largest. Due to the large numbers of studies done, many systematic errors related to inconsistent alignment of acidity scales or treatment of solvent effects have also been introduced. For this set of three solvents, we discarded the iBonD data. In their place, we used trusted values from the "Acid dissociation constants in selected dipolar non-hydrogen-bond-donor solvents" project, whose values were critically evaluated and sometimes corrected by us.[57]

QM details

Gas-phase acidities

We identified the most stable conformers in the gas phase for both neutral and charged solutes by using CREST at the GFN2-xTB level of theory.[87,88] From these, the single most stable conformer for each solute was utilized as the starting point for geometry optimization and frequency analysis in the gas phase, employing the ωB97x-3c level of theory[89] via ORCA 5.0.3.[90–93] The rigid rotor harmonic oscillator[94] approximation is used with the frequency calculations and all reported thermodynamic properties include the zero-point energy. Following geometry optimization, we conducted single point energy calculations at the DLPNO-CCSD(T)/CBS(aug-cc-pVDZ/pVTZ) level of theory using ORCA 5.0.3. For the Complete Basis Set (CBS) calculations, we use the two-parameter expression with an integer exponent of power four:[95,96]

$$E(n) = E_{\mathrm{CBS}} + \frac{A}{\left(n + \frac{1}{2}\right)^4} \tag{2}$$

for n=2 and n=3 at aug-cc-pVDZ/pVTZ levels of theory. We find that this method yields similar results to the standard exponential[96,97] or the mixed Gaussian/exponential expression[96,98] CBS calculations for n=2, 3, and 4 at aug-cc-pVDZ/aug-cc-pVTZ/aug-cc-pVQZ levels of theory. These latter methods, which require three parameters, necessitated about 1.1 core hours for n=2, 7.7 core hours for n=3, and 27.6 core hours for n=4 for a molecule

with 14 heavy atoms on the RWTH HPC cluster. Consequently, the three-parameter methods are significantly more computationally expensive than two-parameter methods due to the additional cost associated with the aug-cc-pVQZ calculations. Further benchmarking confirmed that the combination of methods DLPNO-CCSD(T)/CBS//ωB97x-3c led to low uncertainty in the computations (see Figure S1 and S2 in SI).

COSMO-RS method for solvation free energies

Gas-phase energies and screening charge profiles were computed at the BP86/TZVPD//BP86/TZVP level of theory using COSMOconf 2023[99] with TURBOMOLE 7.7[100] to generate an ensemble of conformers, followed by COSMOtherm 2023[101] with BP-TZVPD-FINE-23 parameterization to derive the solvation free energies.

ML model details

Model architecture Graph neural network (GNN) models have demonstrated their capability to predict various molecular properties directly from the molecular structure with high accuracy.[26,102–108] In these models, molecules are depicted as graphs, where atoms serve as nodes and bonds as edges, each associated with feature vectors that encapsulate atom and bond characteristics. For a comprehensive review of GNN models, we refer readers to the relevant literature.[109–112]

In our work, we employ the Directed-Message Passing Neural Network (D-MPNN) / Feed-Forward Neural Network (FFNN) model implemented in Chemprop v2, a molecular deep learning architecture that has proven effective in predicting a variety of molecular properties with high accuracy.[50,113] Both the molecular representation *and* the FFNN weights are simultaneously learned during training. The final trained models are available on Zenodo (doi:10.5281/zenodo.13987781).

All models were trained with a condensed graph of reaction in a solvent, in which each reaction corresponds to the neutral acid dissociating to the anionic conjugate base.[114] Dur-

ing training, the condensed graph of reaction is embedded by one D-MPNN, and the solvent molecule is embedded by a separate D-MPNN. Both learned embeddings are then concatenated before being introduced to the FFNN.

Splitting For each property, we split the test data once such that each solvent included only solutes *unseen* in the training set. If a solute appears in the test set, it does not appear anywhere in the training set. This resulted in an overall 90% training/validation and 10% test set split for the solvation free energies of ions. The ratio was slightly different among solvents, with every test set representing at least 10% of the total datapoints in that solvent. For instance, 24 training/validation points and 4 test data corresponded to the solvent NMP for solvation energies of anions, representing an 86% - 14% split rather than the overall 90% - 10% across all solvents. We did not split by functional groups, due to the inconsistent availability of data with each functional group in each solvent. From the test sets, we also identified that approximately 15 species for each property were stereoisomers of compounds in the training and validation sets; we removed those from the test sets.

The test set was held constant and used for all GNN models that were trained in this study. Because ionic solvation free energies were only constructed for *overlapping* gas-phase acidities and pK_a data, and because pK_a and solvation free energy data can be stratified by solvents whereas gas-phase acidities cannot, there were sometimes different amounts of data available for the pK_a and gas-phase acidity data than that for the ionic solvation free energies. This resulted in different split proportions. In all cases, the train - validation ratio was maintained as 8:1. For gas-phase data, 8% of the data was held out as the test set, whereas for pK_a, 7% was held out.

Enforcing these data splitting constraints, especially to ensure that species in the training set were not seen in the testing set, may cause test distributions to be slightly different than those of training data. Additionally, the use of only one random split with already relatively low dataset sizes introduces the possibility of biasing the train and test distributions. As

a result, the test performances reported herein should be considered more pessimistic than testing on completely random splits.

Featurization and ensembling During training, we use the standard features provided by Chemprop. For each property, the training and validation data were striped into nine folds with a ratio of 8:1 training to validation. We optimized the model's hyperparameters on one split of data, using 50 iterations across the default search space using the Ray Tune package. The optimal hyperparameters for each property were selected based on the lowest validation error. A model was trained for each fold over 50 epochs, resulting in an ensemble of 9 models. Each model was trained with random initializations of weights and the hyperparameters obtained previously.

Consideration of pretraining Previous work has shown that transfer learning can sometimes improve model performance.[26,115] We did not implement transfer learning for our gas-phase QM calculations because of potential concerns about replicating the NIST data, which is governed by the NIST Standard Reference Data act.[52] We also attempted to pretrain on ChEMBL pK_a data, but observed that this did not improve model performance. Hence, we report here only models without any multifidelity learning.

COSMO-RS benchmark of $\Delta G^*_{\mathrm{solv}}(\mathbf{A}^-)$

The quantum chemical details are the same as discussed for computing the solvation energies of neutral solutes.

The mean average errors of solvation free energies of ions computed using models such as COSMO-RS[15,49] and Solvation Model based on Density (SMD)[116] have been shown to exceed $4\,\mathrm{kcal\,mol^{-1}}$, which can be reduced to approximately $2\,\mathrm{kcal\,mol^{-1}}$ if a systematic correction term is included.[24,42]

To correct for systematic offsets (from misalignment of energy scales and also from solvation model error), we first determined the systematic offset parameters δ that would minimize

the mean squared error of the predictions for each solvent using D2A-dGsolv-anion/train.[24,42] These values are then added to the COSMO-RS predictions and compared to D2A-dGsolv-anion/test. Benchmarking results are shown in the SI.

Data availability

All quantum chemical data generated, curated experimental pK_a values, and the trained machine learning models developed in this study are available on Zenodo (doi:10.5281/zenodo.13987781). Due to constraints associated with the experimental gas-phase acidity data, we are unable to release that subset of data used in this work.

Code availability

The source codes of this work are available at

Chemprop - `https://github.com/chemprop`

MoleculeResolver - `https://github.com/MoleculeResolver/molecule-resolver`

Supporting Information

Benchmarking of computational methods, dataset statistics, chemical class and molecular weight distributions, model performance comparisons, error analysis by functional group and ion type, and supplementary figures and references.

Acknowledgement

T.N. and K.L gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with project number 191948804. J.W.Z.

and W.H.G. acknowledge the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS) for funding. J.W.Z. acknowledges the Takeda Fellowship for funding. Work at Tartu was supported by the Estonian Research Council grant PRG2557 and by the Estonian Ministry of Education and Research (TK210). We acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing high performance computing (HPC) resources that were used to perform the COSMOtherm simulations and Chemprop model training in this study. We acknowledge the RWTH HPC under project p0020138 for providing computing resources that were used to perform the other quantum chemical simulations in this study.

References

(1) Liu, J. et al. Materials Science and Materials Chemistry for Large Scale Electrochemical Energy Storage: From Transportation to Electrical Grid. *Advanced Functional Materials* **2012**, *23*, 929–946.

(2) Liu, Y.; He, G.; Jiang, H.; Parkin, I. P.; Shearing, P. R.; Brett, D. J. L. Cathode Design for Aqueous Rechargeable Multivalent Ion Batteries: Challenges and Opportunities. *Advanced Functional Materials* **2021**, *31*, 2010445.

(3) Malavasi, L.; Fisher, C. A. J.; Islam, M. S. Oxide-ion and proton conducting electrolyte materials for clean energy applications: structural and mechanistic features. *Chemical Society Reviews* **2010**, *39*, 4370.

(4) Basdogan, Y.; Maldonado, A. M.; Keith, J. A. Advances and challenges in modeling solvated reaction mechanisms for renewable fuels and chemicals. *WIREs Computational Molecular Science* **2019**, *10*.

(5) Marrucho, I.; Branco, L.; Rebelo, L. Ionic Liquids in Pharmaceutical Applications. *Annual Review of Chemical and Biomolecular Engineering* **2014**, *5*, 527–546.

(6) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

(7) Egorova, K. S.; Gordeev, E. G.; Ananikov, V. P. Biological Activity of Ionic Liquids and Their Application in Pharmaceutics and Medicine. *Chemical Reviews* **2017**, *117*, 7132–7189.

(8) Mohan, M.; Simmons, B. A.; Sale, K. L.; Singh, S. Multiscale molecular simulations for the solvation of lignin in ionic liquids. *Scientific Reports* **2023**, *13*.

(9) Nevolianis, T.; Wolter, N.; Kaven, L. F.; Krep, L.; Huang, C.; Mhamdi, A.; Mitsos, A.; Pich, A.; Leonhard, K. Kinetic Modeling of a Poly(N-vinylcaprolactam-co-glycidyl methacrylate) Microgel Synthesis: A Hybrid In Silico and Experimental Approach. *Industrial & Engineering Chemistry Research* **2023**, *62*, 893–902.

(10) Kaven, L. F.; Keil, J.; Wolter, N.; Nevolianis, T.; Pich, A.; Leonhard, K.; Mhamdi, A.; Mitsos, A. Dynamic Modeling for Synthesis of Tailored Microgels with Charged Domains. *Industrial & Engineering Chemistry Research* **2024**, *63*, 7727–7742.

(11) Lopatkin, A. J.; Collins, J. J. Predictive biology: modelling, understanding and harnessing microbial complexity. *Nature Reviews Microbiology* **2020**, *18*, 507–520.

(12) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews* **2005**, *105*, 2999–3094, PMID: 16092826.

(13) 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. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396, PMID: 19366259.

(14) Klamt, A. Conductor-Like Screening Model for Real Solvents: A New Approach to the

Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry* **1995**, *99*, 2224–2235.

(15) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *The Journal of Physical Chemistry A* **1998**, *102*, 5074–5085.

(16) Smith, E. J.; Bryk, T.; Haymet, A. D. J. Free energy of solvation of simple ions: Molecular-dynamics study of solvation of Cl- and Na+ in the ice/water interface. *The Journal of Chemical Physics* **2005**, *123*, 034706.

(17) Klamt, A.; Eckert, F.; Arlt, W. COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures. *Annual Review of Chemical and Biomolecular Engineering* **2010**, *1*, 101–122.

(18) Xi, C.; Zheng, F.; Gao, G.; Song, Z.; Zhang, B.; Dong, C.; Du, X.-W.; Wang, L.-W. Ion Solvation Free Energy Calculation Based on Ab Initio Molecular Dynamics Using a Hybrid Solvent Model. *Journal of Chemical Theory and Computation* **2022**, *18*, 6878–6891, PMID: 36253911.

(19) Xu, L.; Coote, M. L. Methods To Improve the Calculations of Solvation Model Density Solvation Free Energies and Associated Aqueous pKa Values: Comparison between Choosing an Optimal Theoretical Level, Solute Cavity Scaling, and Using Explicit Solvent Molecules. *The Journal of Physical Chemistry A* **2019**, *123*, 7430–7438, PMID: 31382743.

(20) Pliego Jr, J. R.; Riveros, J. M. Hybrid discrete-continuum solvation methods. *WIREs Computational Molecular Science* **2020**, *10*, e1440.

(21) Simm, G. N.; Türtscher, P. L.; Reiher, M. Systematic microsolvation approach with a cluster-continuum scheme and conformational sampling. *Journal of Computational Chemistry* **2020**, *41*, 1144–1155.

(22) Rufino, V. C.; Pliego Jr, J. R. Single-ion solvation free energy: A new cluster–continuum approach based on the cluster expansion method. *Physical Chemistry Chemical Physics* **2021**, *23*, 26902–26910.

(23) Pliego, J. R. Cluster expansion of the solvation free energy difference: Systematic improvements in the solvation of single ions. *The Journal of Chemical Physics* **2017**, *147*, 034104.

(24) Kröger, L. C.; Müller, S.; Smirnova, I.; Leonhard, K. Prediction of Solvation Free Energies of Ionic Solutes in Neutral Solvents. *The Journal of Physical Chemistry A* **2020**, *124*, 4171–4181, PMID: 32336096.

(25) Itkis, D.; Cavallo, L.; Yashina, L. V.; Minenkov, Y. Ambiguities in solvation free energies from cluster-continuum quasichemical theory: lithium cation in protic and aprotic solvents. *Physical Chemistry Chemical Physics* **2021**, *23*, 16077–16088.

(26) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021**, *418*, 129307.

(27) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy. *Journal of Chemical Information and Modeling* **2022**, *62*, 433–446.

(28) Alibakhshi, A.; Hartke, B. Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. *Nature Communications* **2021**, *12*.

(29) Lim, H.; Jung, Y. MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning. *Journal of Cheminformatics* **2021**, *13*.

(30) Weinreich, J.; Browning, N. J.; von Lilienfeld, O. A. Machine learning of free energies

in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation. *The Journal of Chemical Physics* **2021**, *154*, 134113.

(31) Fowles, D. J.; McHardy, R. G.; Ahmad, A.; Palmer, D. S. Accurately predicting solvation free energy in aqueous and organic solvents beyond 298 K by combining deep learning and the 1D reference interaction site model. *Digital Discovery* **2023**, *2*, 177–188.

(32) Zhang, Z.-Y.; Peng, D.; Liu, L.; Shen, L.; Fang, W.-H. Machine Learning Prediction of Hydration Free Energy with Physically Inspired Descriptors. *The Journal of Physical Chemistry Letters* **2023**, *14*, 1877–1884.

(33) Yu, J.; Zhang, C.; Cheng, Y.; Yang, Y.-F.; She, Y.-B.; Liu, F.; Su, W.; Su, A. SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digital Discovery* **2023**, *2*, 409–421.

(34) Röcken, S.; Burnet, A. F.; Zavadlav, J. Predicting solvation free energies with an implicit solvent machine learning potential. 2024; `https://arxiv.org/abs/2406.00183`.

(35) Coetzee, J. F.; Dollard, W. J.; Istone, W. K. Measurement of real free energies of transfer of individual ions from water to other solvents with the jet cell. *Journal of Solution Chemistry* **1991**, *20*, 957–975.

(36) Huenenberger, P.; Reif, M. *Single-Ion Solvation*; Theoretical and Computational Chemistry Series; The Royal Society of Chemistry, 2011; pp P001–664.

(37) Pollard, T.; Beck, T. L. Quasichemical analysis of the cluster-pair approximation for the thermodynamics of proton hydration. *The Journal of Chemical Physics* **2014**, *140*, 224507.

(38) Vlcek, L.; Chialvo, A. A. Single-ion hydration thermodynamics from clusters to bulk solutions: Recent insights from molecular modeling. *Fluid Phase Equilibria* **2016**, *407*, 58–75.

(39) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Aqueous Solvation Free Energies of Ions and Ion-Water Clusters Based on an Accurate Value for the Absolute Aqueous Solvation Free Energy of the Proton. *The Journal of Physical Chemistry B* **2006**, *110*, 16066–16081.

(40) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Generalized Born Solvation Model SM12. *Journal of Chemical Theory and Computation* **2013**, *9*, 609–620.

(41) Nevolianis, T.; Baumann, M.; Viswanathan, N.; Kopp, W. A.; Leonhard, K. DISSOLVE: Database of ionic solutes' solvation free energies. *Fluid Phase Equilibria* **2023**, *571*, 113801.

(42) Zheng, J. W.; Green, W. H. Experimental Compilation and Computation of Hydration Free Energies for Ionic Solutes. *The Journal of Physical Chemistry A* **2023**, *127*, 10268–10281.

(43) Cheng, J.-P.; Yang, J.-D.; Xue, X.-S.; Ji, P.; Li, X.; Wang, Z. iBonD Website. `http://ibond.nankai.edu.cn/`.

(44) Sülzner, N.; Haberhauer, J.; Hättig, C.; Hellweg, A. Prediction of Acid pKa Values in the Solvent Acetone Based on COSMO-RS. *Journal of Computational Chemistry* **2022**, *43*, 1011–1022.

(45) Zheng, J. W.; Al Ibrahim, E.; Kaljurand, I.; Leito, I.; Green, W. H. pKa prediction in non-aqueous solvents. *Journal of Computational Chemistry* **2025**, *46*, e27517.

(46) Kütt, A.; Selberg, S.; Kaljurand, I.; Tshepelevitsh, S.; Heering, A.; Darnell, A.; Kaup-

mees, K.; Piirsalu, M.; Leito, I. pKa values in organic chemistry–Making maximum use of the available data. *Tetrahedron letters* **2018**, *59*, 3738–3748.

(47) Bartmess, J. E. Gas-phase equilibrium affinity scales and chemical ionization mass spectrometry. *Mass Spectrometry Reviews* **1989**, *8*, 297–343.

(48) Ervin, K. M. Experimental techniques in gas-phase ion thermochemistry. *Chemical Reviews* **2001**, *101*, 391–444.

(49) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *Journal of Physical Chemistry* **1995**, *99*, 2224–2235.

(50) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling* **2023**, *64*, 9–17.

(51) Linstrom, P. NIST Chemistry WebBook, NIST Standard Reference Database 69. 1997; `http://webbook.nist.gov/chemistry/`.

(52) Lias, S. G.; Bartmess, J. E. NIST Gas-Phase ion Thermochemistry. `https://webbook.nist.gov/chemistry/ion/`.

(53) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota solvation database (MNSOL) version 2012. 2020.

(54) Işık, M.; Levorse, D.; Rustenburg, A. S.; Ndukwe, I. E.; Wang, H.; Wang, X.; Reibarkh, M.; Martin, G. E.; Makarov, A. A.; Mobley, D. L., et al. pKa measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. *Journal of computer-aided molecular design* **2018**, *32*, 1117–1138.

(55) Bergazin, T. D.; Mobley, D. L.; Amezcua, M.; Grosjean, H.; Isik, M.; Slochower, D.; Chodera, J.; Tielker, N.; Ray, D.; Sasmal, S.; Murakumo, K. samplchallenges/SAMPL7: Version 1.1: Update logP analysis; release PHIP2 analysis. 2021; `https://doi.org/10.5281/zenodo.5637494`.

(56) Pliego Jr, J. R. Hybrid Cluster-Continuum Method for Single-Ion Solvation Free Energy in Acetonitrile Solvent. *The Journal of Physical Chemistry A* **2024**, *128*, 6440 – 6449.

(57) Leito, I.; Kaljurand, I.; Piirsalu, M.; Tshepelevitsh, S.; Zheng, J.; Roses, M.; Gal, J.-F. Acid dissociation constants in selected dipolar non-hydrogen-bond-donor solvents. *Pure and Applied Chemistry* **2024**, in press. The full dataset is available at `https://doi.org/10.5281/zenodo.12608876`.

(58) Coetzee, J. F.; Padmanabhan, G. R. Properties of Bases in Acetonitrile as Solvent. IV. Proton Acceptor Power and Homoconjugation of Mono- and Diamines. *Journal of the American Chemical Society* **1965**, *87*, 5005–5010.

(59) Cohen, S. G., Streitwieser, A., Taft, R. W., Eds. *Progress in Physical Organic Chemistry*; Progress in Physical Organic Chemistry; John Wiley & Sons, Inc: Hoboken, NJ, USA, 1963.

(60) Izmailov, N. A.; Chernyi, V. S.; Spivak, L. L. Thermodynamic properties of nonaqueous electrolyte solutions. 14. Calculation of the transport energy of acids from one solvent to another. *Zhurnal Fizicheskoi Khimii* **1963**, 822.

(61) Jasinski, T.; Stefaniuk, K. Miareczkowanie slabych kwasow w srodowisku sulfotlenku dwumetylowego. *Chemia Analityczna (Warsaw)* **1965**, *10*.

(62) Juillard, J. *Bulletin de la Société Chimique de France* **1966**, 1727.

(63) Juillard, J. Ph. D. Thesis, University of Clermont-Ferrand, Clermont-Ferrand, France, 1967.

(64) Kolthoff, I. M.; Chantooni, M. K. Intramolecular hydrogen bonding in monoanions of o-phthalic acid and the homologous oxalic acid series in acetonitrile. *Journal of the American Chemical Society* **1975**, *97*, 1376–1381.

(65) Juillard, J.; Dondon, M.-L. *Bulletin de la Société Chimique de France* **1963**, 2535.

(66) Konovalov, O. M. Effect of solvent on change in free energy of carboxylic acid molecules in solution. *Zhurnal Fizicheskoi Khimii* **1965**, 693.

(67) Kolthoff, I. M.; Bruckenstein, S.; Chantooni, M. K. Acid-Base Equilibria in Acetonitrile. Spectrophotometric and Conductometric Determination of the Dissociation of Various Acids 1. *Journal of the American Chemical Society* **1961**, *83*, 3927–3935.

(68) Kolthoff, I. M.; Chantooni, M. K. Calibration of the Glass Electrode in Acetonitrile. Shape of Potentiometric Titration Curves. Dissociation Constant of Picric Acid 1. *Journal of the American Chemical Society* **1965**, *87*, 4428–4436.

(69) T. Jasinski,; A. A. El-Harakany,; F. G. Halaka,; H. Sadek, Potentiometric Study of Acid-Base Interactions in Acetonitrile. *Croatica Chemica Acta* **1978**, *51*.

(70) Tshepelevitsh, S.; Kütt, A.; Lõkov, M.; Kaljurand, I.; Saame, J.; Heering, A.; Plieger, P. G.; Vianello, R.; Leito, I. On the Basicity of Organic Bases in Different Media. *European Journal of Organic Chemistry* **2019**, *2019*, 6735–6748.

(71) Olmstead, W. N.; Margolin, Z.; Bordwell, F. G. Acidities of water and simple alcohols in dimethyl sulfoxide solution. *The Journal of Organic Chemistry* **1980**, *45*, 3295–3299.

(72) Settimo, L.; Bellman, K.; Knegtel, R. M. A. Comparison of the accuracy of experimental and predicted pKa values of basic and acidic compounds. *Pharmaceutical research* **2014**, *31*, 1082–1095.

(73) Sooväli, L.; Kaljurand, I.; Kütt, A.; Leito, I. Uncertainty estimation in measurement of pKa values in nonaqueous media: A case study on basicity scale in acetonitrile medium. *Analytica chimica acta* **2006**, *566*, 290–303.

(74) Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. Holistic prediction of the pKa in diverse solvents based on a machine-learning approach. *Angewandte Chemie* **2020**, *132*, 19444–19453.

(75) An, H.; Liu, X.; Cai, W.; Shao, X. AttenGpKa: A Universal Predictor of Solvation Acidity Using Graph Neural Network and Molecular Topology. *Journal of Chemical Information and Modeling* **2024**, *64*, 5480–5491.

(76) Abarbanel, O. D.; Hutchison, G. R. QupKake: Integrating Machine Learning and Quantum Chemistry for Micro-pKa Predictions. *Journal of Chemical Theory and Computation* **2024**, *20*, 6946–6956.

(77) Rossini, E.; Netz, R. R.; Knapp, E.-W. Computing pKa Values in Different Solvents by Electrostatic Transformation. *Journal of chemical theory and computation* **2016**, *12*, 3360–3369.

(78) Alconcel, L. S.; Continetti, R. E. Dissociation dynamics and stability of cyclopentoxy and cyclopentoxide. *Chemical physics letters* **2002**, *366*, 642–649.

(79) Garver, J. M.; Yang, Z.; Kato, S.; Wren, S. W.; Vogelhuber, K. M.; Lineberger, W. C.; Bierbaum, V. M. Gas phase reactions of 1, 3, 5-triazine: proton transfer, hydride transfer, and anionic σ-adduct formation. *Journal of The American Society for Mass Spectrometry* **2011**, *22*.

(80) Letcher, T. M. *Development and Applications in Solubility*; Royal Society of Chemistry, 2007.

(81) Marcus, Y. Thermodynamic functions of transfer of single ions from water to nonaqueous and mixed solvents: Part I - Gibbs free energies of transfer to nonaqueous solvents. *Pure and Applied Chemistry* **1983**, *55*, 977–1021.

(82) Marcus, Y.; Marcus, Y. Ions. *Ions in Water and Biophysical Implications: From Chaos to Cosmos* **2012**, 49–98.

(83) Himmel, D.; Goll, S. K.; Leito, I.; Krossing, I. Anchor points for the unified Brønsted acidity scale: the rCCC model for the calculation of standard Gibbs energies of proton solvation in eleven representative liquid media. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2011**, *17*, 5808–5826.

(84) Malloum, A.; Fifen, J. J.; Conradie, J. Solvation energies of the proton in methanol revisited and temperature effects. *Physical chemistry chemical physics : PCCP* **2018**, *20*, 29184–29206.

(85) Müller, S.; de Castilla, A. G.; Taeschler, C.; Klein, A.; Smirnova, I. Evaluation and refinement of the novel predictive electrolyte model COSMO-RS-ES based on solid-liquid equilibria of salts and Gibbs free energies of transfer of ions. *Fluid phase equilibria* **2019**, *483*, 165–174.

(86) Contributors, M. R. molecule-resolver (version 0.1.2). 2024; `https://pypi.org/project/molecule-resolver/0.1.2/`, Accessed: 2024-08-05.

(87) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

(88) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(89) Müller, M.; Hansen, A.; Grimme, S. ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set. *The Journal of Chemical Physics* **2023**, *158*.

(90) Neese, F. The ORCA program system. *WIREs Comput Mol Sci* **2012**, *2*, 73–78.

(91) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput Mol Sci* **2018**, *8*.

(92) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of chemical physics* **2020**, *152*, 224108.

(93) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Computational Molecular Science* **2022**, *12*, e1606, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1606.

(94) Atkins, P.; Friedman, R. *Molecular Quantum Mechanics*, 5th ed.; Oxford University Press: Oxford, 2011.

(95) Martin, J. M. Ab initio total atomization energies of small molecules — towards the basis set limit. *Chemical Physics Letters* **1996**, *259*, 669–678.

(96) Vasilyev, V. Online complete basis set limit extrapolation calculator. *Computational and Theoretical Chemistry* **2017**, *1115*, 1–3.

(97) Feller, D. Application of systematic sequences of wave functions to the water dimer. *The Journal of Chemical Physics* **1992**, *96*, 6104–6114.

(98) Woon, D. E.; Dunning, T. H. Benchmark calculations with correlated molecular wave functions. VI. Second row A2 and first row/second row AB diatomic molecules. *The Journal of Chemical Physics* **1994**, *101*, 8877–8893.

(99) BIOVIA COSMOconf 2023. Dassault Systèmes. `https://www.3ds.com`.

(100) TURBOMOLE V7.7 2022, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from
`https://www.turbomole.org`.

(101) BIOVIA COSMOtherm 2023. Dassault Systèmes. `https://www.3ds.com`.

(102) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling* **2017**, *57*, 1757–1772.

(103) Brozos, C.; Rittig, J. G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Graph neural networks for surfactant multi-property prediction. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **2024**, *694*, 134133.

(104) Sanchez Medina, E. I.; Linke, S.; Stoll, M.; Sundmacher, K. Graph neural networks for the prediction of infinite dilution activity coefficients. *Digital Discovery* **2022**, *1*, 216–225.

(105) Greenman, K. P.; Green, W. H.; Gómez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chemical Science* **2022**, *13*, 1152–1162.

(106) Rittig, J. G.; Ben Hicham, K.; Schweidtmann, A. M.; Dahmen, M.; Mitsos, A. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. *Computers and Chemical Engineering* **2023**, *171*, 108153.

(107) Qin, S.; Jiang, S.; Li, J.; Balaprakash, P.; Lehn, R. C. V.; Zavala, V. M. Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium. *Digital Discovery* **2023**, *2*, 138–151.

(108) Nevolianis, T.; Rittig, J. G.; Mitsos, A.; Leonhard, K. Multi-fidelity graph neural networks for predicting toluene/water partition coefficients. **2024**, Preprint.

(109) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. Proceedings of the 34th International Conference on Machine Learning. 2017; pp 1263–1272.

(110) Rittig, J. G.; Gao, Q.; Dahmen, M.; Mitsos, A.; Schweidtmann, A. M. In *Machine Learning and Hybrid Modelling for Reaction Engineering*; Zhang, D., Del Río Chanona, E. A., Eds.; Royal Society of Chemistry, 2023; pp 159–181.

(111) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Communications Materials* **2022**, *3*, 93.

(112) Schweidtmann, A. M.; Rittig, J. G.; Weber, J. M.; Grohe, M.; Dahmen, M.; Leonhard, K.; Mitsos, A. Physical pooling functions in graph neural networks for molecular property prediction. *Computers and Chemical Engineering* **2023**, *172*, 108202.

(113) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

(114) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *Journal of Chemical Information and Modeling* **2021**, *62*, 2101–2110.

(115) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. *The Journal of Physical Chemistry A* **2019**, *123*, 5826–5835.

(116) 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. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396.

TOC Graphic