Graph-Based Approaches for Predicting Solvation Energy in Multiple Solvents: Open Datasets and Machine Learning Models

Logan Ward,[*,†] Naveen Dandu,[‡] Ben Blaiszik,[†,¶] Badri Narayanan,[‡,§] Rajeev S. Assary,[‡] Paul C. Redfern,[‡] Ian Foster,[†,‖,¶] and Larry A. Curtiss[‡]

†*Data Science and Learning Division, Argonne National Laboratory, Lemont, IL*

‡*Materials Science Division, Argonne National Laboratory, Lemont, IL*

¶*Globus, University of Chicago, Chicago, IL*

§*Department of Mechanical Engineering, University of Louisville, Louisville, KY*

‖*Department of Computer Science, University of Chicago, Chicago, IL*

E-mail: lward@anl.gov

Abstract

The solvation properties of molecules, often estimated using quantum chemical simulations, are important in the synthesis of energy storage materials, drugs, and industrial chemicals. Here, we develop machine learning models of solvation energies to replace expensive quantum chemistry calculations with inexpensive-to-compute message-passing neural network models that require only the molecule graph as inputs. Our models are trained on a new database of solvation energies for 130 258 molecules taken from the QM9 dataset computed in five solvents (acetone, ethanol, acetonitrile, dimethyl sulfoxide, and water) computed via an implicit solvent model. Our best model achieves a Mean Absolute Error (MAE) of 0.5 kcal/mol for molecules with nine

or fewer nonhydrogen atoms and 1 kcal/mol for molecules with between 10 and 14 nonhydrogen atoms. We make the entire dataset of 651 290 computed entries openly available, and provide simple web and programmatic interfaces to enable others to run our solvation energy model on new molecules. The model calculates the solvation energies for molecules using only the SMILES string and also provides an estimate of whether each molecule is within the domain of applicability of our model. We envision that the dataset and models will provide functionality needed for the rapid screening of large chemical spaces to discover improved molecules for many applications.

Introduction

The ability of a molecule to dissolve in a particular solvent is a key factor in determining whether it will be suitable for many technological applications, such as drug molecules that must dissolve in the bloodstream or redox-active molecules that must remain in solution inside a flow battery. Chemists assess the solubility of a molecule by computing the solvation energy (ΔG_{solv}) – a measure of the strength of interaction between a molecule and a solvent – from quantum mechanical simulations.[1] While methods for computing solvation energy are well established,[2] their computational expense is large enough (minutes to hours per compound on a single processor) to impose serious limitations on how many molecules can be assessed. Evaluating thousands of chemicals requires significant computational resources, yet would explore a small fraction of modern chemical search spaces that approach billions of molecules.[3–6] Faster approaches, such as machine-learned QSAR models,[7] are required to reach the necessary evaluation throughput to scan for promising leads in these large chemical spaces.

The quality of a machine-learned model of a chemical property is highly dependent on the size, breadth, and quality of the available data. Existing machine learning models for solvation energy have been trained using small databases of experimental data, which limits their applicability. One such database, FreeSolv,[8] has been used extensively to create ma-

chine learning (ML) models for hydrogenation energy based on diverse techniques including convolutional neural networks[9,10] and fingerprint-based machine learning.[11,12] Similar experimental databases, like those from the University of Minnesota,[13] provide solvation data for many solvents and have also been used as the basis for machine learning models, as illustrated by pioneering work from Borhani et al[14] and Subramanian et al.[15] The limited size and chemical diversity of experimental data (only a few hundred molecules) makes them impractical to use with high-accuracy, deep learning models (e.g., Refs.[16–18]). This issue can be addressed by developing large and diverse datasets of solvation properties (¿100,000 molecules) via computations that provide sufficient representation of different regions of the chemical space.

Here, we report on a new large dataset, QM9-Solvation, of computed solvation energies (ΔG_{solv}) for 130 258 molecules in multiple solvents, and provide an initial set of machine learning models trained using this data. The dataset was generated from DFT simulations using the SMD Solvation Model[19] and is openly available for use by the chemistry community.[20] We also demonstrate a state-of-the-art ML prediction model of molecular solvation energies trained on this database. Our model uses a message-passing architecture that includes properties of both the solvent and solute, which we demonstrate is capable of learning solvent/solute relationships from over a half-million training entries. We established a domain of applicability region for this model and find that it is possible to predict solvation energies of molecules larger than those in our training set, and have made the models freely available through GitHub and DLHub.[21,22] We envision that our database and molecules will help spur rapid advancements in the ability for scientists to efficiently assess the solvation properties of molecules.

Methods

Our work involved first creating a database of solvation energies using quantum chemistry computations and creating models to predict solvation energy using machine learning. We describe the quantum chemistry and machine learning aspects of our work separately.

Quantum chemistry

We performed Density functional theory (DFT) calculations using Gaussian16 software to compute solvation energies of molecules.[23] All calculations were performed using the B3LYP functional and 6-31G(2df,p) basis set starting with the relaxed geometries from the QM9[24] and G4MP2-GDB9 datasets.[25–27] Each relaxed geometry was then used in computing single point energies in presence of five different solvents using a Solvation Model based on Density (SMD model)[19] at the same level of theory that was used for geometry relaxation. The solvents chosen for the DFT calculations are acetone, ethanol, acetonitrile, dimethyl sulfoxide, and water. The energies of these molecules in the gas phase [E_{gas}] and in the solvent [$E_{solvent}$] were used to calculate the solvation energy [ΔG_{sol}] of the molecules in each of the solvents via the following equation.

$$[\Delta G_{sol}] = [E_{solvent}] - [E_{gas}] \tag{1}$$

The computed solvation energies of GDB9 and selected molecules from the Pedley compilation,[28] reported in kcal/mol, are available on the Materials Data Facility.[20]

Machine learning

We use Message Passing Neural Networks (MPNNs) for most machine learning tasks described in this work. As formalized by Gilmer et al.,[17] MPNNs are a general class of architectures for learning properties of graphs and there exists significant degrees of freedom within MPNN design. The general architecture of MPNNs includes "message" layers that

produce a signal for each node (or, in some architectures, each edge[29]) given its neighboring edges and nodes and a function then updates the features of each node based on its message. The combined "message-then-update" cycle can be repeated many times before the final, "readout" function that produces a single feature vector for the entire graph. The features for the graph are then passed as inputs to another model (e.g., a fully-connected neural network) that maps the graph features to a property of the graph. As with all neural networks, the weights of an MPNN are learned by adjusting them to minimize the error between prediction and true values through gradient descent.

We start with the message, update, and readout functions used by St. John et al.[30] The initial features of the atoms and bonds are determined by looking up a set of features for each type of atom–defined using its element, number of bonds, number of bonded hydrogen atoms, and whether it is aromatic–and for each bond–defined using its type (e.g., single vs. double), whether the bond is conjugated or in a ring, and the elements of the bonded atoms. The features for each type of atom or bond are learned while training the machine learning model. Messages are produced by summing over the product of atom and bond features of each neighboring atom. Features are updated using a gated recurrent unit (GRU) module, which generates the new atom features considering both the new message and all previous feature values for that atom. We apply the message and update functions 6 times. After these layers we generate a set of features for each atom that reflect its properties and those of up to the sixth-nearest neighbors.

As shown in Figure 1, we use two versions of the MPNN that differ based on whether the readout function is applied before or after reducing the atomic features to a single scalar. The first, "Molecular Fingerprint," version uses a sum readout function before reducing the atomic features to a single value, which effectively produces a set of features that describes the entire molecule. The "Atomic Contribution" version reduces the atomic features to a single scalar value before summing over all atoms, which approximates each atom having an additive contribution to solvation energy.

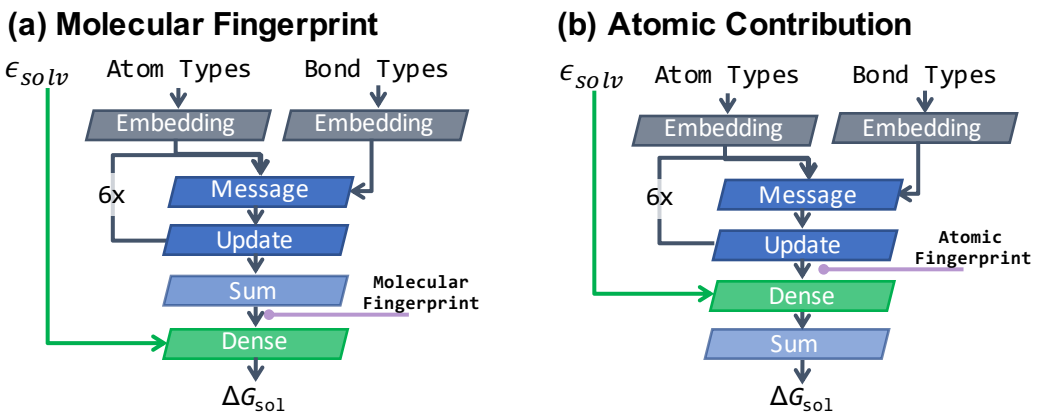

Figure 1: The two message-passing neural network (MPNN) versions used in this work: (a) a model that computes solvation energy based on a molecular fingerprint and (b) a a model that computes solvation energy by summing contributions of each atom in a molecule. Boxes represent key computations in a network, and open yellow arrows illustrate how data is passed between them. Orange closed arrows indicate the model outputs that represent molecular fingerprints for (a) and atomic fingerprints for (b).

Full details of the models are available in the Supplementary Information, on the GitHub page associated with this work and on the Materials Data Facility.[20,21]

Results and discussion

We describe the development of machine learning model that predict the solvation energies of a diverse class of molecules in five solvents. The following sections describe how we accomplished this goal by first establishing a database of solvation energies based on an unbiased sampling of molecular structures and then gradually increasing the complexity of machine learning techniques used to learn correlations from this data.

Database of 650 000 solvation energies of small molecules

We base our training set on the QM9 database of Ramakrishnan et al., which has been used frequently in the molecular machine learning- community, for a number of reasons. First, the database is large: with over 10^5 entries, QM9 provides sufficient training data to support the

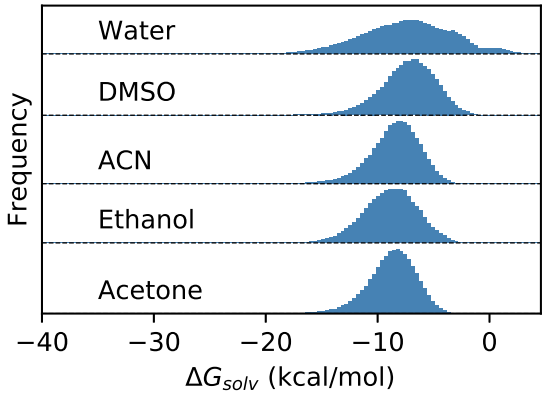

Figure 2: Histogram of solvation energies, ΔG_{solv}, of all 130 258 molecules in our QM9-Solvation dataset in each of the five solvents considered in this work.

development of many machine learning models.[29,31,32] Second, the database is diverse. QM9 is based on an exhaustive sampling of molecules,[5,33] which allows us to train and validate our molecules with minimal concern over whether it is learning patterns related to the sampling of data and not the physics of the problem.

We used the SMD implicit solvation model in Gaussian to compute the solvation energy for 130 258 molecules from QM9. We selected SMD for its comparable accuracy to explicit solvent models and reasonable correlation with experimental solvation energies at a reduced computational cost.[34,35] For each molecule, we computed the solvation energy in five solvents that sample a broad range of dielectric constants: acetone ($\epsilon = 20.49$), ethanol ($\epsilon = 24.85$), acetonitrile (ACN) ($\epsilon = 35.69$), dimethyl sulfoxide (DMSO) ($\epsilon = 46.83$), and water ($\epsilon = 78.36$). We selected these solvents to not only to capture differences among the solvation behavior of different molecules, but also selecting solvents of technological relevance (e.g., ACN and water are common solvents for flow batteries[36]).

As illustrated in Figure 2, the solvation energies for our molecules follow, generally, a single-modal distribution. The notable exception is water, which features multiple peaks near -10, -5 and 0 kcal/mol. The standard deviation in the solvation energy is approximately 2.8 kcal/mol for all solvents except water, which shows a wider distribution with

standard deviation of 4.3 kcal/mol. Some molecules have solvation energies substantially higher or lower than most in the dataset. The solvation energies for these molecules can range from as low as -84 kcal/mol to as high as 5 kcal/mol. For example, there are 543 (0.4%) and 428 (0.3%) molecules with solvation energies below -20 kcal/mol in water and acetone, respectively.

We divide the QM9 molecules into a subset used for model training and a second segment used for model evaluation. Following our previous work,[37] we removed 10% of the full dataset (a total of the 13026 molecules) for use as a test set. These molecules were neither used to train the parameters of any of the architectures nor to determine early-stopping criteria employed in the neural network training. Accordingly, we use them as a test dataset that represents data drawn from the same population as the training examples (i.e., molecules of smaller than nine heavy atoms).

We also included a set of 191 molecules with more than nine heavy atoms to provide further evaluation of the models. These molecules were selected in a previous study,[27] where we identified molecules with between 10 and 14 heavy atoms from the Pedley Compilation[28] that have accurate measurements of formation enthalpy. These molecules capture molecular motifs that are impossible to form with fewer than 10 heavy atoms, such as polycyclic aromatic compounds. We denote these molecules our "Pedley test set." They were not selected randomly but drawn from a set of molecules for which solvation energies have been measured experimentally (unlike the QM9 training set, which were generated procedurally) and, further, were chosen to exhibit qualities desirable for validation (e.g., possessing molecular motifs not included in the QM9-Solvation dataset).

The full dataset, which we denote QM9-Solvation, is available through GitHub and the Materials Data Facility.[20,21]

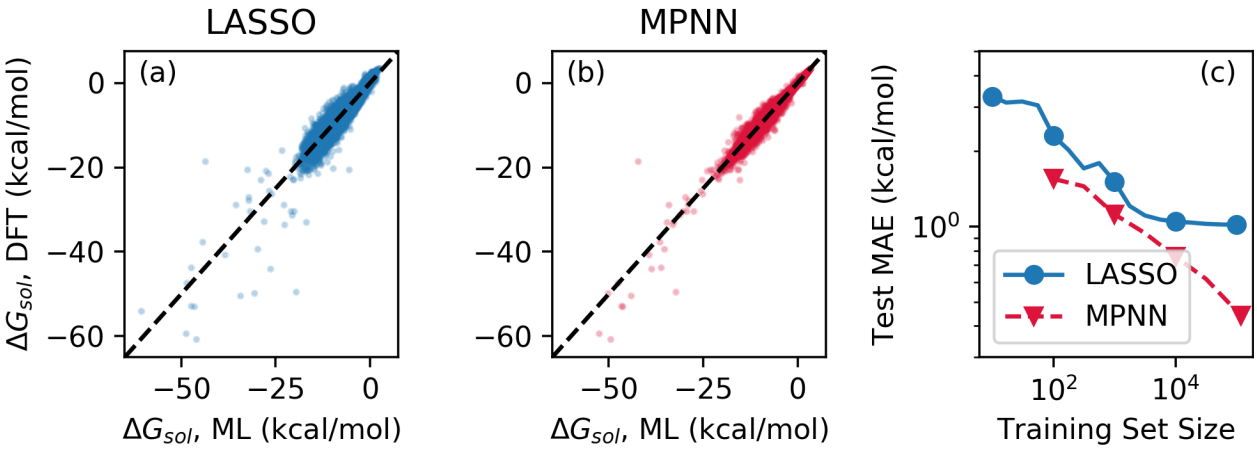

Figure 3: Comparison of the performance of different machine learning models on predicting the solvation energy in water of 13 000 molecules outside of their training set. Each model was trained on the same 117 232 molecules. From left to right: Performance of group contribution model trained using sparse linear regression (LASSO); MPNN performance; LASSO and MPNN vs. training set size. LASSO fails to continue improving in accuracy with training set sizes of $>10^4$ and is half as accurate as MPNN at training set sizes nearing 10^5 entries.

Predicting solvation energies of molecules in water solvent: Message Passing Neural Networks (MPNNs)

Our goal is to link the graph structure of a molecule to solvation energy. Using only the molecular graph means that estimating the solvation energy of a molecule does not require any force field or quantum chemistry computations, which both allows our models to screen new chemicals at high rates and mitigates concerns about tying predicted solvation energy to a specific conformation.

The use of machine learning to link molecular structures to properties is well studied.[7] We choose to evaluate two techniques from the literature due to their simplicity: group contribution with linear regression and message-passing neural networks (MPNNs). Both methods start by labeling atoms in the molecule with type information, such as by element or by functional group, but differ significantly in how these labels are mapped to functional properties.

We modeled our group contribution approach on the widely-used model for water-octanol partition coefficient (LogP) of Wildman and Crippen.[38] We define functional groups for each

non-hydrogen atom in a molecule by first assigning a type based on the element of that atom, whether it is in an aromatic cycle, and its degree. We then complete the description of the atom by computing the types of the neighboring atoms using the same procedure and recording the type of bonding. Our procedure lacks the hand-tuned groups of Wildman and Crippen but distinguishes atomic sites more fully. For example, our method describes primary and secondary aliphatic carbons differently; these are treated as identical by Wildman and Crippen. We believe this further granularity is needed for a dataset as diverse as QM9. We represent each molecule as a vector in which each entry corresponds to the count of that functional group type. We then use L_1 regularized linear regression using the least absolute shrinkage and selection operator (LASSO)[39] to fit a linear model between this vector and solvation energy.

Our second type of model is a Message-Passing Neural Network, which provides an automatic path for learning the similarity between different functional groups and their impact on a material property. A MPNN, as formalized by Gilmer et al.[17] and described in the Methods, learns interactions between atoms/functional groups through "message-passing" steps that update the description of each atom based on the types of neighboring atoms and the types of bonds connecting to the neighboring atoms. In this work, we use an MPNN architecture implemented by St. John et al.[30] Our MPNN describes atoms initially based on their type, degree, number of bonded hydrogen atoms, and whether it is in an aromatic ring; updates the atom description with messages based on sums of those from its nearby atoms; and generates a description of the entire molecule as a sum of all atoms. Full details are available in Methods.

We compared the MPNN and linear regression models for predicting the solvation energy in water using the training and test data sets described in the previous section. As shown in Figure 3, our MPNN model has a Mean Absolute Error (MAE) of 0.44 kcal/mol—2.3× better than the 1.0 kcal/mol MAE of the linear regression model. The increased accuracy of the MPNN can be traced to the limited complexity available to the linear regression model.

The LASSO model only includes 1228 parameters, far below the 897 409 parameters in the MPNN model, which likely explains why the LASSO model does not improve in accuracy when provided with more than 10^4 data points. Some routes to improving the accuracy of this linear regression model could be to add non-linear combinations of existing features or adding more features by differentiating functional groups based on second-nearest neighbors. The elegance of the MPNN is that it learns such non-linear combinations and more complex features automatically. As we demonstrate here, the ability to learn such interactions leads to superior predictive accuracy.

We studied the outliers of the MPNN to gain a further insight on the generalizability of the model. The compounds with the largest absolute errors are are those with particularly high solvation energies, with 15 of the top 25 errors being within the 1% most negative solvation energies. The relatively poor predictive performance on these materials is partly explained by the solvation energy being outliers compared to the rest of the data. Twelve of top 25 (48%) errors belong to a class of molecules, zwitterions, that is rare within the training data with only 396 (0.3%) entries within the training. We conclude that the model's predictions are most reliable for molecules with solvation energies of less negative than -25 kcal/mol and is, for example, inaccurate for zwitterions.

Extending an MPNN solvation model to other solvents

Our next step was to add the ability for our MPNN model to predict solvation energies in other solvents. We explored two different strategies for creating "multi-task" models able to predict solvation energy in different solvents simultaneously. The first approach is a standard multi-task model in which the last layer of the neural network produces multiple outputs, one per solvent. The second approach involves a network that predicts solvation energy given the molecular structure of the solute and the dielectric constant of the solvent. This "dielectric constant" model, unlike the multi-task model, can predict solvation energy in solvents not included in the training set. Here, we test which method produces more accurate models

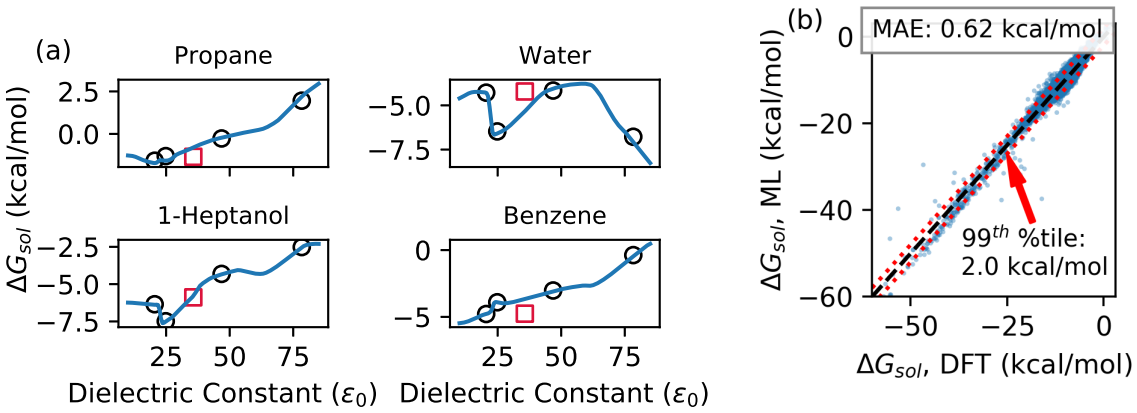

Figure 4: (a) Predictions of solvation energy as a function of the dielectric constant of a solvent for four solute molecules with different solvation behavior. Our Message Passing Neural Network (MPNN) model includes the dielectric constant of the solvent as an input feature. It was trained on the solvation energies of 117 232 molecules. The solid blue line is the solvation energy prediction for each molecule from our MPNN. Black circles represent solvents that were included in the training set (left to right: acetone, ethanol, DMSO, water) and the red square is a solvent withheld from our training set (ACN). (b) Performance of the MPNN model in predicting the solvation energy in DMSO for all 130 258 molecules in our dataset.

and how well the dielectric-constant model computes solvation energy for solvents outside of the training set.

We evaluate our two strategies by training a model with the same 90%/10% test split as our single-solvent model and data from all five solvents. The multi-task and dielectric-constant models have comparable accuracy to the single-solvent model in predicting solvation energy in water. The single-solvent model has an error of 0.44 kcal/mol for solvation energy in water—effectively equal to the errors of 0.43 and 0.45 kcal/mol for the multi-task and dielectric constant models, respectively. We note that we previously observed that training a multi-task model with greatly varied properties (e.g., HOMO energy and atomization energy) can lead to a significant degradation in model accuracy.[37] It appears that the solvation energies of a molecule in different solvents are sufficiently similar that modeling different solvents simultaneously in the same model does not lead to degradation in performance.

We further explored the concept of multi-solvent prediction by withholding data from one of our five solvents, ACN, to use as a test set. We selected ACN for this test case because it

has the median dielectric constant of the five solvents and, thus should be a good test for our machine learning model's ability to interpolate. The MAE of our dielectric-constant model for predicting solvation energy in water is not changed by the exclusion of ACN training data; the MAE decreases by only 5×10^{-4} kcal/mol. Based on these results, we see neither benefit nor disadvantage to training a machine learning model on data from more solvents at least for solvents that exist within the range of dielectric constants in our training set.

The fact that the dielectric constant model uses solvent properties as inputs allows us to predict the properties of solvents not contained within the training set. As shown in Figure 4, our model smoothly interpolates how the solvation energy changes as a function of dielectric constant. The model also appropriately detects that the solvation energy for propane increases for solvents with large dielectric constants and increases for water. The dielectric constant model trained without any data with ACN as a solvent achieves an MAE for the ACN solvation energy of 0.62 kcal/mol across all molecules and an MAE of 0.70 kcal/mol on molecules from the holdout set, for which neither molecule nor solvent are included in the training set. From this, we conclude that the model is able to interpolate between solvents with dielectric constants between 10 and 80 ϵ_0.

Improving performance for molecules larger than the training set

The training data used in the previous sections are drawn from the same population, which provides a useful but limited view of the utility of our model. Specifically, all molecules were taken from the QM9 dataset and, as a result, have only nine or fewer nonhydrogen atoms, limiting the estimates of accuracy discussed above to only these smaller molecules. We therefore studied the accuracy of our model on predicting the solvation energy of 191 molecules with more than nine nonhydrogen atoms (i.e., our Pedeley test set).

The errors for these molecules from the Pedley test set are substantially greater than those observed for the QM9-based test set. For example, the MAE for solvation energy in ACN for the Pedley test set is 3.46 kcal/mol—over 10$\times$ larger than the MAE of 0.32 kcal/mol

for the QM9 test set. The errors tend to increase with molecule size. As shown in Figure 5c, the MAE for molecules in the Pedley test set generally increases as a function of molecular size and reaches over 10 kcal/mol for molecules with 14 heavy atoms. The increase in error with molecule size is not monotonic, with molecules of size 11 having errors around 2× smaller than errors for molecules with 14 nonhydrogen atoms, but is strong enough that we speculate that the poor performance of our models on the Pedley test set has to do with the issues with capturing the effect of molecular size more than the presence of structural motifs absent from the training set.

We explored improving the performance of our model by adopting a "atomic contribution" model for solvation energy. As illustrated in Figure 1, our current model generates a single fingerprint to describe an entire molecule by summing the fingerprint of each atom (the output of the message passing layers) and computing the solvation energy from the resultant "molecular fingerprint." An alternative approach is to predict "atomic contributions" for each atom in the molecule separately and then to express the molecular property as a sum of these contributions—a technique used often in machine learning models of molecular energy.[16,31,40,41] The atomic contribution approach explicitly encodes a linear relationship between a molecular property and the size of the molecule, which is a fair approximation for solvation energies given the known correlation between solvation energy and molecular size.[42]

The atomic contribution model performs worse for the QM9 holdout set but has a less dramatic increase in error with molecule size than our original model. As shown in Figure 5b, the atomic contribution model has a 50% larger error on the QM9 dataset, 0.47 kcal/mol, than the molecular fingerprint model, 0.32 kcal/mol. However, we see a much smoother increase with error with increasing molecular size. The MAE for molecules with 14 nonhydrogen atoms is, for example, at least 5× smaller than our molecular fingerprint model. The drastic differences between the performance on QM9 and our external test sets strongly motivates the need for evaluating machine learning models more closely to how they will be

used for larger molecules and the need for molecular datasets with larger molecules. We also note that the atomic contribution model is not fully free from the effect of molecular size on error. The errors steadily increase for molecules with sizes more distant from nine nonhydrogen atoms, which is the size with the greatest amount of training data. There could be a few explanations for the relationship between error and molecule size. Namely, there could be could be an effect of bias in the training set due to the overrepresentation of data with 9-heavy atoms or a failure in the assumption that atom count and solvation energy should be linearly-related. We therefore propose re-weighting techniques to account for training set bias or different approaches for summing atomic contributions (e.g., generalized means) as a direction for modeling non-linear relationships between atomic size and property.

The final ingredient of the utility of our model is the speed at which we can evaluate new molecules. Evaluating the model requires parsing the SMILES string to an RDKit molecule object, converting the RDKit molecule to a graph representation compatible with Tensorflow, and evaluating the MPNN. The conversion steps can be performed in parallel for each molecule and the individual mathematical operations of the MPNN can also be parallelized, allowing us to exploit highly-parallel CPUs and general-purpose GPUs. Accordingly, we can achieve an evaluation rate of 10^3 molecules per second on two 32-core Intel Xeon CPU E5-2683 v4 CPUs. This evaluation rate makes it possible to evaluate all of QM9 (10^5 molecules) in two minutes and all of ZINC15 in 33 hours using only a single workstation. The extreme evaluation rates noted here can be further improved by running the models on multiple nodes of a supercomputer, making it possible to rapidly screen molecule spaces with billions of candidates.

Quantifying domain of applicability

The empirical nature of machine learning models makes it difficult to identify which predictions are most or least likely to be accurate. In the previous section, we established that molecular size correlates with error—a likely effect of the training set being biased towards

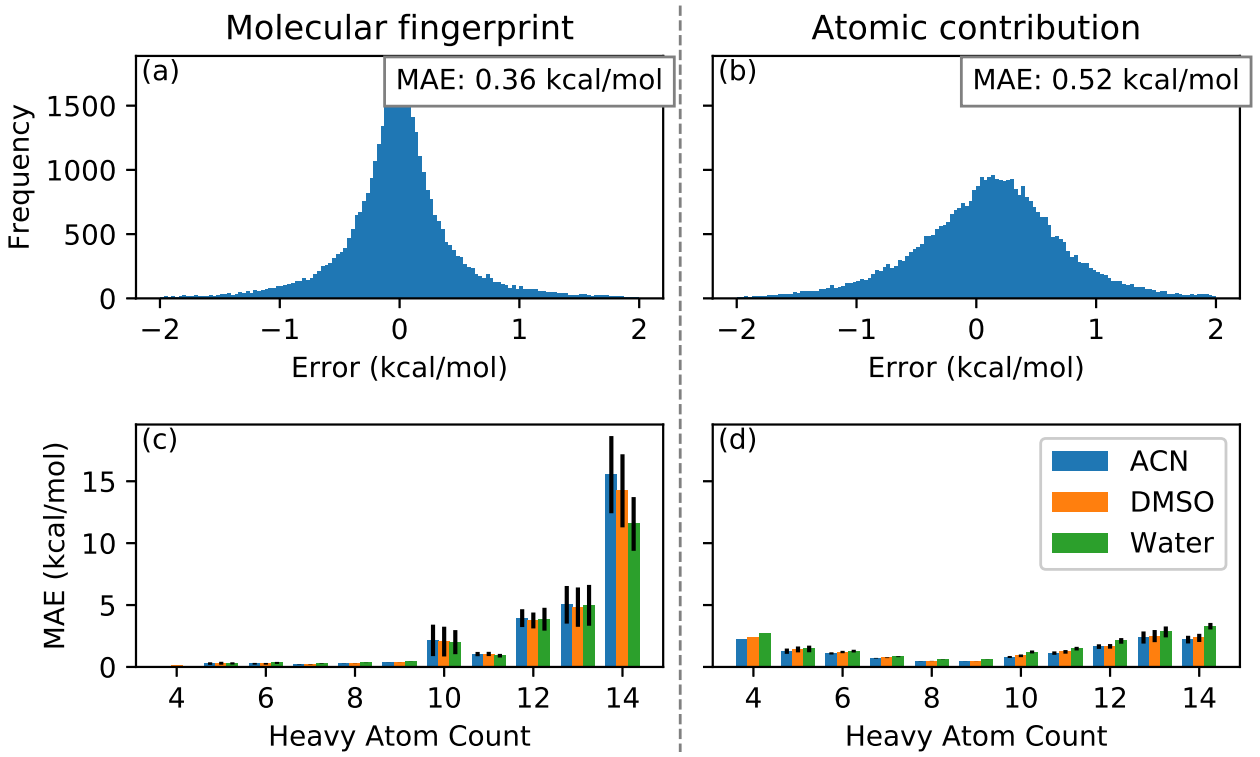

Figure 5: Performance of two different MPNN models trained on 117 232 molecules with nine or fewer nonhydrogen atoms on predicting the solvation energy in acetonitrile (ACN), dimethyl sulfoxide (DMSO), and water on our validation hold-out sets: (a,c) a model which learns the solvation energy from a molecular fingerprint and (b,d) a model which predicts solvation energy by summing over contributions from each atom. (a-b) Histogram of errors in the solvation energy in the ACN solvent for hold-out molecules from the QM9 dataset. (c-d) The mean absolute error for molecules from the QM9 hold-out set and our larger molecule hold-set set binned by the number of nonhydrogen atoms. Error bars represent the standard error of the mean.

molecules with nine nonhydrogen atoms. Here, we explore improved metrics for establishing a quantitative estimate of the size of the error for a prediction.

Our previous analysis of the performance of the models on predicting solvation energies in water give clues on what is important in estimating the performance of the model. We noted, in particular, that zwitterions were overrepresented in the molecules with the largest error (48% of largest errors were zwitterions, compared to only 0.4% in the test set). The poor-performance of the models on a class of molecules that are structurally distinct from most of the training set is reflective on how the MPNN models implicitly learn the behavior of function groups within molecules. Unseen or rare functional groups will have little

training data, which leads can lead to poor performance. Such behavior suggests for us to consider metrics that measure similarity in functional groups between molecules to estimate the performance of our model.

Our metrics are based on common approaches from the chemoinformatics community,[43] such as distances from the training set and variance of an ensemble of machine learning models. We explore three different applicability metrics. We first chose the Tanimoto similarity based on Morgan Fingerprints,[44] which represents a molecular similarity metric that is not tuned for any particular property. We also chose two applicability metrics based on our MPNN models. The first is based on the molecular fingerprint MPNN model, were we compute the mean L_2 distance between the molecular fingerprint (see Fig. 1) for a molecule and those of the 64 closest molecules from the training set. The second is based on the fingerprints for each atom as computed using the atomic contribution model, where we compute the dissimilarity of a molecule from one in the training set as the mean distance between the most-similar pairs of atoms:

$$a(m,n) = \frac{1}{N_m} \sum_i \max_j \|m_i - n_j\|_2 \tag{2}$$

where m is the molecule, n is a molecule from the training set, N_m is the number of atoms in m, the sum $\sum_i$ is over all atoms in m, the maximum $\max_j$ is over all atoms in n, m_i is the atomic fingerprint of atom i in m, n_h is the fingerprint of atom j in n, and $\|...\|_2$ represents the L_2 norm. We note that $a(m,n)$ is not a proper distance metric, as $a(m,n) \neq a(n,m)$. As with the "molecular fingerprint distance," we compute the atomic fingerprint distance as the average distance over the 64 closest molecules. In all cases, we assume larger distances or lower similarities from the training set indicate greater likelihood that a molecule is outside of the training set.

Our final applicability metric is based on the variance of the predictions of an ensemble of MPNNs. We created an ensemble by training 16 atomic contribution MPNNs each with

a different, randomly-selected subset created by randomly sampling the full training set with replacement (i.e., a bootstrapped ensemble). The variance of this training captures the uncertainty of individual predictions.[45,46] We assume that molecules with larger uncertainties are more likely to have larger errors.

We evaluated the quality of our domain of applicability metrics by measuring the correlation between each application metric for each prediction and the size of the observed error. We seek applicability metrics for which molecules with larger applicability metrics have a greater likelihood of having larger errors than those with smaller metrics. The statistical variation in error means that we assume that these changes will be visible in averages and not that all points with larger metrics will have larger errors. Accordingly, we grouped our data into 16 bins based on each of the applicability metric and show the average errors and likelihood of an error being greater than 1 kcal/mol. In all cases, we used the predicted solvation energy in ACN based on the atomic contribution MPNNs for molecules in our Holdout and Pedley test sets. The results are shown in Figure 6.

We find that the applicability based on the molecular fingerprint MPNN yields the best performance according to both quality metrics. We find that there is the strongest correlation between both the magnitude of error and the likelihood of a 1 kcal/mol error. The MPNN-based fingerprint distance strongly outperforms a generic circular fingerprint with Tanimoto similarity, suggesting a benefit to using fingerprints trained to reflect a certain atomic property as a similarity metric. In general, we find significant need for developing good domain of applicability metrics. As illustrated by specialized MPNN fingerprints outperforming conventional Tanimoto fingerprints, the dependence of model error on inputs can require complicated models. Further research on developing complex applicability metrics or, even, measuring the performance of different metrics (e.g., Ref[47]) is greatly needed.

We therefore train a linear regression and logistic classification model on the MPNN molecular fingerprint distance as a means to provide a quantitative estimate of the uncertainty of any prediction made using our model. Users of the model can define their own

cutoff for a domain of applicability depending on how much uncertainty that can tolerate. We can also use the uncertainties of the models to further improve them[48] or accelerate the discovery of molecules with target properties.[49,50]

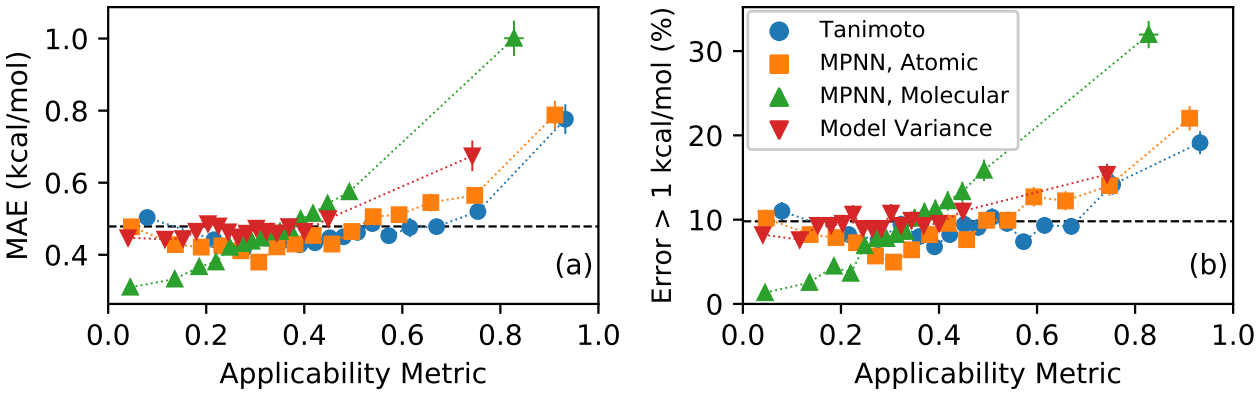

Figure 6: Comparison of different domain of applicability metrics: Tanimoto similarity to the 64 most-similar molecules in the training set, distance based on the atomic and molecular fingerprints from a MPNN model for solvation energy, and variance in a bootstrap ensemble of 16 models. The performance of each metric was assessed using 13217 molecules from our holdout set using ACN as a solvent. (a) The MAE and normalized applicability metric for molecules divided into 16 bins grouped by each metric. Metrics are normalized by applying a linear transformation such that the 1^{st} and 99^{th} percentiles have values of 0 and 1, respectively. (b) Likelihood of a error greater than 1 kcal/mol for 16 bins of molecules grouped by each metric. The dotted lines between data points are intended to guide the eye. The black dashed lines show (a) the MAE on the entire holdout set and (b) the fraction of errors $>$ 1 kcal/mol for the holdout set.

Conclusion

We have presented QM9-Solvation, a new dataset of DFT-computed solvation energies of 130 258 molecules in five solvents, a total of 651 290 values, and used this new dataset to train a suite of machine learning models. Our best-performing machine learning models, based on a message-passing neural network that takes properties of both solute molecules and solvent as inputs, achieve an accuracy of 0.5 kcal/mol, 2× better than group contribution models. We established that our models can generalize to predict solvation energy in solvents outside of the training set and for molecules larger than those used to train our datasets. Finally,

we created a domain-of-applicability metric to provide users an estimate of whether the predictions of our models are trustworthy for a certain molecule. The data, models, and code used in this study are all freely available online, and the machine learning models are accessible through a simple web interface that both invokes the model and provides estimates of model uncertainty for each prediction.

Availability of data, models, and code

The data, models, and code used in this study are freely available via the Materials Data Facility,[51,52] DLHub,[22,22,51] and Github, respectively. Specifically, we have published our training and test datasets;[20] the best trained models (the atomic-contribution models with solvent dependence); and the code used for training and testing our machine learning models.[21]

Supporting Information

The Supporting Information for this manuscript includes a detailed description of the message-passing neural network architecture.

Acknowledgement

LW, ND, BN, RSA, PCR, and LAC were supported to develop the database and ML algorithms as part of the Joint Center for Energy Storage Research (JCESR), an Energy Innovation Hub funded by the US Department of Energy, Office of Science, Basic Energy Sciences. LW and IF were supported to perform ML tasks on High-Performance Computing by the DOE Exascale Computing Project, ExaLearn Co-design Center. BB and work on making the dataset openly available was performed under financial assistance award 70NANB14H012 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Material Design (CHiMaD). BB and IF

were also supported by the National Science Foundation as part of the Midwest Big Data Hub under NSF Award Number: 1636950 "BD Spokes: SPOKE: MIDWEST: Collaborative: Integrative Materials Design (IMaD): Leverage, Innovate, and Disseminate." Work by BB on making machine learning models available through DLHub was supported in part by Laboratory Directed Research and Development funding from Argonne National Laboratory under U.S. Department of Energy under Contract DE-AC02-06CH11357.

We also thank the Argonne Leadership Computing Facility for access to the PetrelKube Kubernetes cluster. This research used resources of the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

References

(1) Cheng, L.; Assary, R. S.; Qu, X.; Jain, A.; Ong, S. P.; Rajput, N. N.; Persson, K.; Curtiss, L. A. Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening. *The Journal of Physical Chemistry Letters* **2015**, *6*, 283–291.

(2) Mennucci, B. Continuum Solvation Models: What Else Can We Learn from Them? *The Journal of Physical Chemistry Letters* **2010**, *1*, 1666–1674.

(3) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

(4) Saadi, A. A. et al. IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads. 2020.

(5) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.

(6) Babuji, Y. et al. Targeting SARS-CoV-2 with AI- and HPC-enabled Lead Generation: A First Data Release. 2020.

(7) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chemical Reviews* **2012**, *112*, 2889–2919.

(8) Mobley, D. L.; Guthrie, J. P. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 711–720.

(9) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689* **2017**,

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

(11) Rauer, C.; Bereau, T. Hydration free energies from kernel-based machine learning: Compound-database bias. *The Journal of Chemical Physics* **2020**, *153*, 014101.

(12) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *Journal of Chemical Information and Modeling* **2017**, *57*, 726–741.

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

(14) Borhani, T. N.; García-Muñoz, S.; Luciani, C. V.; Galindo, A.; Adjiman, C. S. Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics* **2019**, *21*, 13706–13720.

(15) Subramanian, V.; Ratkova, E.; Palmer, D.; Engkvist, O.; Fedorov, M.; Llinas, A. Multi-solvent models for solvation free energy predictions using 3D-RISM Hydration Thermodynamic Descriptors. 2020; doi.org/10.26434/chemrxiv.11791245.v1.

(16) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *Journal of Chemical Physics* **2018**, *148*, 241722.

(17) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. 34th International Conference on Machine Learning. 2017; pp 2053–2070.

(18) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **2017**, *8*, 3192–3203.

(19) 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.

(20) Ward, L.; Dandu, N.; Blaiszik, B.; Narayanan, B.; Assary, R. S.; Redfern, P. C.; Foster, I.; Curtiss, L. A. Dataset: Datasets and Machine Learning Models for Accurate Estimates of Solvation Energy in Multiple Solvents. Materials Data Facility, 2021; `https://petreldata.net/mdf/detail/solv_ml_v1.2`.

(21) `https://github.com/globus-labs/solvation-energy-ml`.

(22) Chard, R.; Li, Z.; Chard, K.; Ward, L.; Babuji, Y.; Woodard, A.; Tuecke, S.; Blaiszik, B.; Franklin, M.; Foster, I. DLHub: Model and data serving for science. International Parallel and Distributed Processing Symposium. 2019; pp 283–292.

(23) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(24) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*.

(25) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate quantum chemical energies for 133 000 organic molecules. *Chemical Science* **2019**, *10*, 7449–7455.

(26) Narayanan, B.; Redfern, P.; Ward, L.; Blaiszik, B.; Foster, I.; Assary, R. S.; Curtiss, L. G4MP2-GDB9 Database. 2019; `https://petreldata.net/mdf/detail/narayananbadri_g4mp2gdb9_database_v1.1`.

(27) Dandu, N.; Ward, L.; Assary, R. S.; Redfern, P. C.; Narayanan, B.; Foster, I. T.; Curtiss, L. A. Quantum-Chemically Informed Machine Learning: Prediction of Energies of Organic Molecules with 10 to 14 Non-hydrogen Atoms. *The Journal of Physical Chemistry A* **2020**, *124*, 5804–5811.

(28) Pedley, J. *Thermochemical data and structures of organic compounds*; CRC Press, 1994; Vol. 1.

(29) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **2019**, *31*, 3564–3572.

(30) St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-passing neural networks for high-throughput polymer screening. *Journal of Chemical Physics* **2019**, *150*, 234111.

(31) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *Journal of Chemical Physics* **2018**, *148*, 241717.

(32) von Lilienfeld, O. A.; Burke, K. Retrospective on a decade of machine learning for chemical discovery. *Nature Communications* **2020**, *11*.

(33) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society* **2009**, *131*, 8732–8733.

(34) Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation* **2017**, *13*, 1034–1043.

(35) Chen, J.; Shao, Y.; Ho, J. Are Explicit Solvent Models More Accurate than Implicit Solvent Models? A Case Study on the Menschutkin Reaction. *The Journal of Physical Chemistry A* **2019**, *123*, 5580–5589.

(36) Shin, S.-H.; Yun, S.-H.; Moon, S.-H. A review of current developments in non-aqueous redox flow batteries: characterization of their membranes for design perspective. *RSC Advances* **2013**, *3*, 9095.

(37) Ward, L.; Blaiszik, B.; Foster, I.; Assary, R. S.; Narayanan, B.; Curtiss, L. Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations. *MRS Communications* **2019**, *9*, 891–899.

(38) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873.

(39) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 267–288.

(40) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters* **2007**, *98*.

(41) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* **2010**, *104*.

(42) Borhani, T. N.; García-Muñoz, S.; Luciani, C. V.; Galindo, A.; Adjiman, C. S. Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics* **2019**, *21*, 13706–13720.

(43) Klingspohn, W.; Mathea, M.; ter Laak, A.; Heinrich, N.; Baumann, K. Efficiency of different measures for defining the applicability domain of classification models. *Journal of Cheminformatics* **2017**, *9*.

(44) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(45) Ling, J.; Hutchinson, M.; Antono, E.; Paradiso, S.; Meredig, B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation* **2017**, *6*, 207–217.

(46) Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics* **2017**, *19*, 10978–10985.

(47) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology* **2020**, *1*, 025006.

(48) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *The Journal of Chemical Physics* **2018**, *148*, 241727.

(49) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. Active learning in materials

science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **2019**, *5*.

(50) Doan, H. A.; Agarwal, G.; Qian, H.; Counihan, M. J.; Rodríguez-López, J.; Moore, J. S.; Assary, R. S. Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials. *Chemistry of Materials* **2020**, *32*, 6338–6346.

(51) Blaiszik, B.; Ward, L.; Schwarting, M.; Gaff, J.; Chard, R.; Pike, D.; Chard, K.; Foster, I. A data ecosystem to support machine learning in materials science. *MRS Communications* **2019**, *9*, 1125–1133.

(52) Blaiszik, B.; Chard, K.; Pruyne, J.; Ananthakrishnan, R.; Tuecke, S.; Foster, I. The Materials Data Facility: Data services to advance materials science research. *JOM* **2016**, *68*, 2045–2052.

Graphical TOC Entry