THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

ConfSolv: Prediction of Solute Conformer-Free Energies across a Range of Solvents

Published as part of The Journal of Physical Chemistry B virtual special issue " Machine Learning in Physical Chemistry Volume 2".

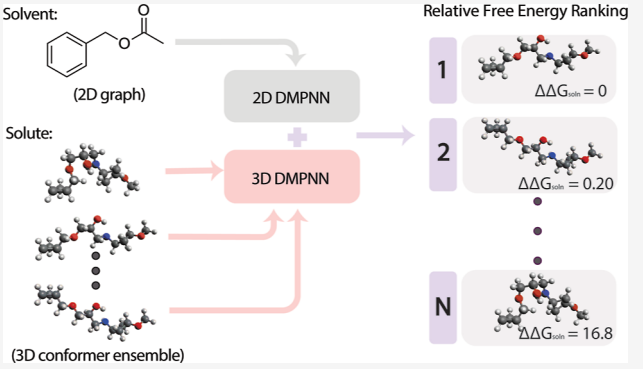
Lagnajit Pattanaik,[‖] Angiras Menon,[‖] Volker Settels, Kevin A. Spiekermann, Zipei Tan, Florence H. Vermeire, Frederik Sandfort, Philipp Eiden, and William H. Green*

Cite This: *J. Phys. Chem. B* 2023, 127, 10151−10170

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Predicting Gibbs free energy of solution is key to understanding the solvent effects on thermodynamics and reaction rates for kinetic modeling. Accurately computing solution free energies requires the enumeration and evaluation of relevant solute conformers in solution. However, even after generation of relevant conformers, determining their free energy of solution requires an expensive workflow consisting of several ab initio computational chemistry calculations. To help address this challenge, we generate a large data set of solution free energies for nearly 44,000 solutes with almost 9 million conformers calculated in 41 different solvents using density functional theory and COSMO-RS and quantify the impact of solute conformers on the solution free energy. We then train a message passing neural network to predict the relative solution free energies of a set of solute conformers, enabling the identification of a small subset of thermodynamically relevant conformers. The model offers substantial computational time savings with predictions usually substantially within 1 kcal/mol of the free energy of the solution calculated by using computational chemical methods.

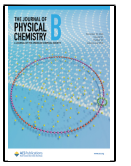


INTRODUCTION

Gibbs free energies are crucial thermodynamic parameters for reaction kinetics as the change in Gibbs free energy determines the equilibrium constant and thus the favorability and direction of a reaction. Nowadays, computation of Gibbs free energies of molecules can be performed using quantum chemistry calculations,[1] but this is typically done in the gas phase. In the case of practical reactions in solution, the free energy of each solvated solute is a crucial property and is dependent on computing both the gas-phase free energies and the solvation free energies for all participating species. As a consequence, calculation of free energies in solution can often guide reaction selection and pathway analysis,[2−4] industrial process optimization,[5] and analysis of protein−ligand binding for bioactivity determination in drug discovery.[6] Typically, these applications require rapid screening of a large number of solutes and solvents, which has resulted in a variety of approaches for computing the solvation free energies.

Accurate computation of solvation and free energies in solution phase can use explicit solvation methods in which solvent molecules are explicitly included. Explicit solvation methods often make use of molecular dynamics, with solute−solvent interactions captured by a force field.[7] However, such approaches are often too expensive, particularly for high-throughput applications. Continuum solvation models are an alternative approach, which have long been integrated in computational quantum mechanical frameworks.[8] Broadly, these approaches treat the solute in a quantum mechanical way, with solvent effects included by placing the solute in a cavity within the reaction field of the solvent. As noted by Tomasi et al.,[8] this largely focuses on electrostatic interactions between the solute and solvent and is applicable primarily for very dilute systems. One of the most widely used continuum solvation model is the polarizable continuum model.[9] Pure continuum solvation models are popular for being computationally inexpensive and, due to their integration with quantum chemistry packages, are applicable to a range of solutes and solvents. However, their accuracy is questionable,[10] so various other methods build on these continuum solvation model

Received: August 31, 2023
Revised: October 25, 2023
Accepted: October 27, 2023
Published: November 15, 2023



approaches to improve performance. This includes the solvent method based on density (SMD),[11] which uses quantum chemistry calculations and parameterization to experiments to improve performance.

Another popular and high-performing model is COSMO-RS.[12,13] COSMO-RS builds on the original COSMO method,[14] which operates by placing a solute molecule within an infinitely strong dielectric and calculating the screened surface charge densities that result from the solute polarizing the dielectric medium. This results in a subsequent back-polarization of the solute, which must be accounted for by introducing the surface charges as external potentials, which then alters the electron density of the solute itself. Thus, the correct surface charge density is obtained in a self-consistent manner. However, this approach, as with all continuum solvation models, neglects temperature and mixture effects. COSMO-RS improves upon COSMO by making use of thermodynamic equilibrium and a statistical mechanical framework to model real solvated systems and compute a variety of properties for general liquid mixtures. COSMO-RS models a liquid as an ensemble of ideally screened molecules. These molecules have different surface charges and screening charge densities, resulting in electrostatic interactions. The contribution of these interactions to macroscopic properties is computed by statistical thermodynamics, while the surface charges themselves are derived mainly from quantum chemistry calculations. Hellweg and Eckert[13] note that parameterization based on these quantum chemistry calculations is conducted such that reasonable predictions can be made for mixtures of small- to medium-sized molecules while using relatively cheap but specific density functional theory (DFT) methods for the quantum chemistry calculations. The COMSO-RS method has already been benchmarked on the prediction of solvated properties, notably for solvation free energy on the SAMPL4 challenge[15] and for logP on the SAMPL6 challenge.[16] The DFT methodology used within COSMO-RS is discussed further in the Methodology section.

Accurate computational determination of experimental observables, such as free energy in solution, requires enumerating relevant conformers of a molecule. Conformers are local minima on the potential energy surface, distinguished primarily by differences in torsion angles of rotatable sigma bonds. If these barriers of rotation are sufficiently low, then conformers will rapidly interconvert between one another, so we must consider Boltzmann averages of low-energy states when calculating free energies in solution. Generation of a representative set of low-energy conformers is a task that has been heavily researched in cheminformatics.[17,18] Many flavors of conformer generation exist, including those based on stochastic searches over interatomic distances,[19] exhaustive searches over torsion angles,[20,21] Bayesian optimization,[22] metadynamics,[23] and even deep learning.[24−26] Literature suggests that a conformational search in COSMO-RS workflows is crucial to accurately calculate solvation free energies. For example, Buggert et al.[27] showed that accurate calculation of partitioning coefficients necessitates low-energy conformers for each phase independently. Hyttinen and Prisle[28] found that COSMO-RS calculations were closer to experimental solubilities when choosing select conformers without H-bond donors, as including these conformers leads to a large overestimation. The typical COSMOconf workflow generates a conformer set using BALLOON and the MMFF94 force field followed by the AM1 semiempirical level of theory to quickly screen low-energy structures, followed by several refinements of the conformer set using more accurate DFT calculations and cluster analysis algorithms.[29] This approach is flexible and can work for a range of molecule sizes, but important conformers can be missed in the initial generation and filtering as small-molecule force fields can sometimes struggle with relative conformer assessments,[30] and AM1 does not capture hydrogen bonding.[31] In fact, the winning SAMPL6 blind challenge for pK_a prediction came from COSMO-RS,[32] which relied on the comprehensive conformer search algorithm developed by Grimme et al.[33] In particular, Pracht et al.[32] noted that the solvation free energy contribution to individual conformers was crucial in determining the relevant conformations necessary for accurate prediction. Others have similarly noted the need for robust workflows to determine low-energy conformers in solution.[34]

Pure ab initio methods to calculate properties in solution have always been followed closely by data-driven methods, and the emergence of machine learning in the physical sciences has accelerated development of the latter.[35] Noteworthy approaches include the use of structural fingerprints to describe solvents and solutes, which are then fed to gradient boosting algorithms[36,37] or neural networks[38] to regress solvation free energy as well as kernel ridge regression methods with an ensemble-based representation of a set of structural configurations of the solute to fit experimental solvation free energies.[39] More recent methods use graph neural networks based on 2D connectivity to feature solutes and solvents.[40] Specifically, Vermeire and Green[41] showed that using a combination of computational and experimental data with a transfer learning strategy resulted in a model with mean absolute error (MAE) predictions near 0.20 kcal/mol. This is excellent performance, but it should be noted that the Vermeire & Green model was tested using relatively small molecules with fewer conformers than the molecules studied in this present work; their model is less accurate for large molecules.[41]

A key issue with chemical deep learning models—especially those that directly predict properties—is that they tend to generalize poorly for data outside the training domain. Chung et al.[42] attempt to address this issue by predicting parameters of physical models which are subsequently used to calculate solvation free energy; however, a model that directly predicts the solvation free energy actually outperforms the other approaches, even for out-of-sample predictions from substructure splits. Vermeire and Green[41] partially overcame the generalization problem by using an initial training set composed of COSMO-RS calculations on a large number of molecules containing many elements, but they did not consider many conformers in those calculations. Similarly, Vermeire et al.[43] used models for solvation free energy to predict the solubility limits of solids through the use of thermodynamic state functions that relate the solubility of gases and solids, again achieving high accuracy across a range of organic solvents, but without considering multiple conformers.

In this work, we computationally generate a data set of thousands of molecules with enumerated conformers and calculate solvation and solution free energies across a range of 41 solvents using COSMO-RS. We analyze this data set and the performance of the COSMO-RS theory with respect to experiment. Similar to other studies, we show that a limited number of conformers is needed to achieve good accuracy in the final solution free energy prediction. Finally, we construct a

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

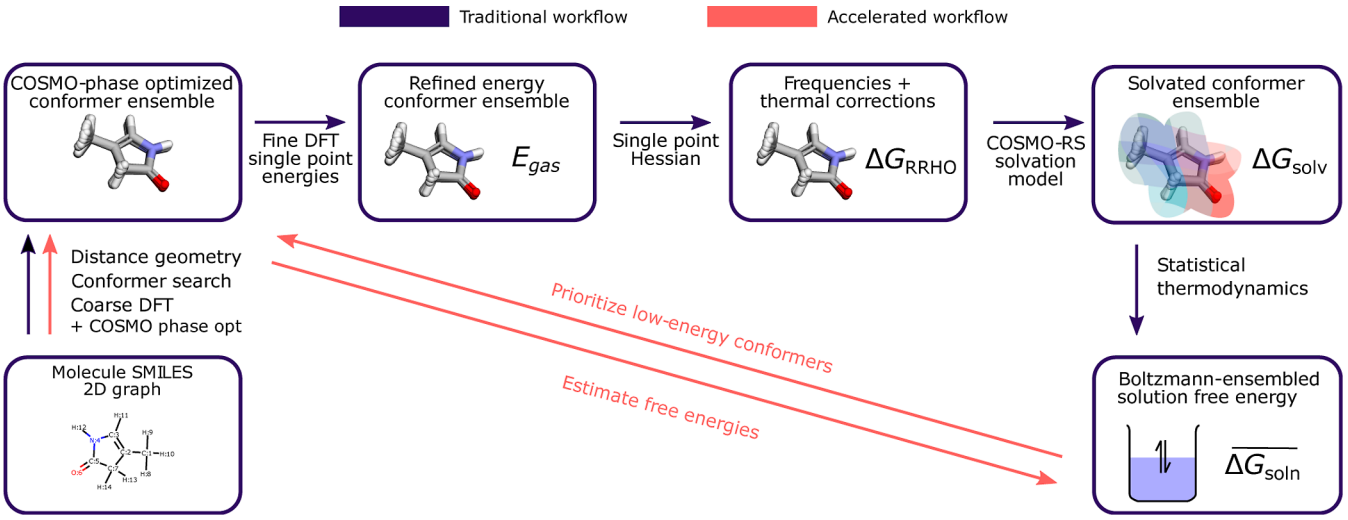


Figure 1. Computational workflow used to calculate the data set using DFT, xTB, and COSMO-RS and proposed acceleration to the workflow using the model developed in this work.

deep learning model based on 3D message passing neural networks (MPNNs) to predict relative solution free energies of 3D conformers to help prioritize low-energy structures for further computation.

METHODOLOGY

Data Set Generation. The key property of interest is ΔG_{soln}, the free energy of an individual solute conformer dissolved in a given solvent, calculated as follows

$$\Delta G_{soln} = E_{gas} + \Delta G_{solv} + \Delta G_{RRHO} \quad (1)$$

Therefore, the data set generation requires computation of three components: the electronic energy at 0 K of each solute conformer in the gas phase, E_{gas}, the change in solvation free energy due to the solute−solvent interactions, ΔG_{solv}, and the change in Gibbs free energy due to thermal contributions, ΔG_{RRHO}. Figure 1 details the full workflow used in this work to construct the data set in blue, with the proposed accelerated workflow developed in this work shown in red. We discuss each step in detail below.

To compute the lowest gas-phase electronic energy at 0 K, E_{gas}, a set of solute conformers must first be generated. In this work, we use a systematic procedure for conformer generation to explore potential low-energy solute structures. First, single bonds, sp^2−sp^2 single bonds, and rotatable double bonds (not in a ring or adjacent to another double bond) in the molecule are systematically rotated in step sizes of 120, 180, and 180°, respectively. This coarse grid for conformational search is chosen as a practical compromise between accuracy and the computational costs given the size of the data set generated in this work. The impact of using a coarse conformer grid is analyzed in subsequent sections by comparing the generated conformer set to that generated using a finer conformer search for a representative set of 28 solutes. The initial conformer set is generated using MMFF94 as implemented in RDKit, with initial xyz structures built from SMILES strings using OpenBabel. Three filtering criteria are applied to narrow the conformer set. The first rejects any generated conformers that undergo a connectivity change upon subsequent optimization compared to the initial structure. To define connectivity, we use a bond distance cutoff between atoms A and B, d_{cutoff}

$$d_{cutoff} = 1.3(r_A + r_B) \quad (2)$$

where r_A and r_B are the covalent radii of atoms A and B, respectively, which in this work are taken from Pyykkö and Atsumi.[44] If the distance between atoms A and B, d_{AB}, is less than d_{cutoff}, then atom A is assumed to be bonded to atom B. Once conformers are pruned for connectivity changes, the conformer set is then checked to find conformers that contain similar torsional angles to one another. If the maximum absolute deviation (MAD) and root-mean-square deviation (RMSD) of the dihedral angles between two conformers are less than 10 and 20°, respectively, then one conformer is removed from the set. The final pruning criterion checks for conformers that have very similar ΔG_{solv} at infinite dilution in a set of selected solvents (ammonia, benzene, chloroform, di-2-butylamine, DMSO, formic acid, water, nonadecanol, perfluoro-hexane, tributylphosphate, and sulfuric acid) as determined by COSMO-RS. If the MAD and RMSD for energy between conformers are less than 1.0 and 0.05 kJ/mol, respectively, and the MAD and RMSD between conformers for rotational constants are less than 0.02 and 0.01 MHz, respectively, the conformer is rejected. This final pruning is performed primarily to retain only one member of each pair of enantiomers. However, to account for the presence of chirality, the weighting of the remaining conformer is increased appropriately in subsequent COSMOtherm and free energy calculations.

We next used DFT to refine conformer geometries. The conformers are optimized using the pure meta-GGA functional TPSS[45] and the double-ζ polarized split-valence basis set, def-SVP.[46] This level of theory has been shown to give good predictions of geometries for organic species as well as transition-metal complexes.[47,48] Potential solvent effects on solute conformer geometries are also included by performing DFT optimizations in the COSMO phase, corresponding to a dielectric constant, ϵ, of infinity. These optimizations are performed using the m3 grid and resolution of identity approximations.[49,50] Single-point energy calculations are then performed at the TPSS-D3(BJ)/def2-TZVP level of theory on these optimized conformer geometries. This makes use of a larger TZVP basis set, Weigend et al.,[51] to improve the estimates of energies of the solute conformers, E_{gas}, and also

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

includes dispersion corrections using the D3 framework with Becke-Johnson damping proposed by Grimme.[52] All of these calculations are performed using the TURBOMOLE software.[53−55]

Next, we compute ΔG_{RRHO}, the thermal contribution to the Gibbs free energy of the solute conformer to go from a 0 K reference to 298 K, using the Rigid Rotor Harmonic Oscillator (RRHO) partition function. This necessitates the calculation of vibrational frequencies for each solute conformer, which involves second derivatives of energies and is thus very computationally expensive with DFT methods. Therefore, we use the single-point Hessian methodology of Spicher and Grimme[56] to expedite this process. In this approach, we reoptimize the DFT geometries computed when calculating E_{gas} with a lower level semiempirical method. To ensure that the geometry is not perturbed substantially, we apply an external biasing potential. This potential enforces DFT-quality geometries but allows some relaxation such that subsequent vibrational frequency calculations do not produce erroneous imaginary values. Vibrational frequencies can then be computed using the same lower level semiempirical method at a fraction of the computational cost. The RRHO partition function and ΔG_{RRHO} can then be computed from statistical thermodynamics. In this work, we use the popular tight binding method, GFN2-xTB, to compute vibrational frequencies as previous benchmark studies have shown that GFN2-xTB offers a very good trade-off between computational cost and accuracy when compared to DFT and other ab initio methods for a range of molecules and complexes.[57,58] Additional studies have benchmarked ΔG_{RRHO} of reaction computed by GFN2-xTB against DFT methods as well as ΔG_{RRHO} for noncovalent interactions and supramolecular structures.[57,59,60] Both these studies found small differences between GFN2-xTB and DFT methods in the range of 5−10%, so we therefore expect relative ΔG_{RRHO} between conformers to be determined reasonably well. We use the very tight convergence criteria for the Hessian computation, as implemented in xTB, with the external potential centered on the position of the optimized DFT structure. The scaling factors for the frequencies, rotor cutoff, and imaginary frequency cutoff are 1, 50 cm^{-1}, and -20^{-1}, respectively, when computing the vibrational partition function. ΔG_{RRHO} is computed at 298 K for all species in this work.

Finally, we calculated ΔG_{solv} using the COSMO-RS methodology. As noted in Hellweg and Eckert,[13] COSMO-RS includes adjustments to specific levels of theory used in quantum mechanics (QM) calculations to improve accuracy and allow the usage of computationally inexpensive DFT methods. To compute ΔG_{solv}, we use the BP-TZVP procedure as implemented in COSMOtherm version 2021 with the BP_TZVP_21.ctd parameterization, which is considered to be the medium-accuracy level of theory implemented in COSMO-RS.[13] This method consists of an initial single-point energy calculation for the solute conformer at the BP86/def-TZVP level of theory and another single-point energy calculation in the COSMO phase with $\epsilon = \infty$, which generates the required.cosmo file. The same procedure is used to generate .cosmo files for each solvent conformer. COSMOtherm uses the solute.cosmo file along with the BP86/def-TZVP electronic energy as the gas-phase energy, the TPSS-D3(BJ)/def2-TZVP electronic energy as the external quantum chemical gas phase energy, and all of the solvent conformer.cosmo files to calculate ΔG_{solv}. Note that this means that solvent conformers are not treated individually as with solutes, but instead their impact is only considered within the implicit solvation framework of COSMOtherm. This is another practical decision, and future work could explore explicit treatment of both solvent and solute conformers together. We use the default COSMOtherm reference state,[61] which computes the free energy change of solvation for transfer of a solute molecule from an ideal gas at 1 M concentration to an ideal solution at the same solute concentration. ΔG_{solv} is also computed at 298 K for all species in this work.

The free energy for each solute conformer in each solvent can then be computed using eq 1, with E_{gas} as the TPSS-D3(BJ)/def2-TZVP gas-phase energy, ΔG_{RRHO} from xTB, and ΔG_{solv} from COSMO-RS. Of the three components, E_{gas} is much larger in magnitude than either ΔG_{solv} or ΔG_{RRHO} and is therefore the main contributing factor to ΔG_{soln}. Achieving accurate values of ΔG_{soln} requires accurate computation of E_{gas}, which DFT methods can struggle with as noted by Hellweg and Eckert.[13] To help mitigate this, we make use of relative ΔG_{soln} values between conformers, which enables us to leverage error cancellation in E_{gas} and should thus be more accurate than the absolute free energy values. While this will not compensate for all of the systematic error from DFT, it is a practical choice given the size of the large data set generated in this work. The range of ΔG_{soln} relative to the lowest energy conformer (LEC) is shown in Figures S1 and S2, decomposed into contributions from E_{gas}, ΔG_{solv}, and ΔG_{RRHO}. This highlights that there will be cases where relative ΔG_{solv} and ΔG_{RRHO} values are important to consider.

Additionally, the range of relative ΔG_{soln} values suggests that multiple conformers may be thermodynamically relevant and warrant consideration. To account for conformers in the ΔG_{soln} calculation, we use the Boltzmann ensemble as described by

$$\begin{aligned}\overline{\Delta G}_{soln} &= -RT\ln(Q_{soln}) \\ &= -RT\ln\left(\sum_{i=0}^{(N_{conf}-1)}\exp\left(\frac{-\Delta G_{soln,i}}{RT}\right)\right)\end{aligned} \tag{3}$$

where Q is the partition function based on the solution free energy, with the summation taken over the N_{conf} conformers located during the conformational search. Using algebraic manipulations (see the Supporting Information), this Boltzmann-ensembled free energy of each solute in each solvent can be written as

$$\begin{aligned}\overline{\Delta G}_{soln} &= \Delta G_{soln,0} - RT\ln\left(1+\sum_{i=1}^{(N_{conf}-1)}\exp\left(\frac{-\Delta\Delta G_{soln,i}}{RT}\right)\right) \\ \Delta\Delta G_{soln,i} &= \Delta G_{soln,i} - \Delta G_{soln,0}\end{aligned} \tag{4}$$

In eq 4, conformers are ordered by ΔG_{soln}, such that $\Delta G_{soln,0}$ is the free energy change of the solution for the LEC. Similarly, one can derive an analogous equation for the gas-phase free energy where ΔG_{gas} is given by the sum of the gas-phase electronic energy E_{gas} and the RRHO contribution ΔG_{RRHO}

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

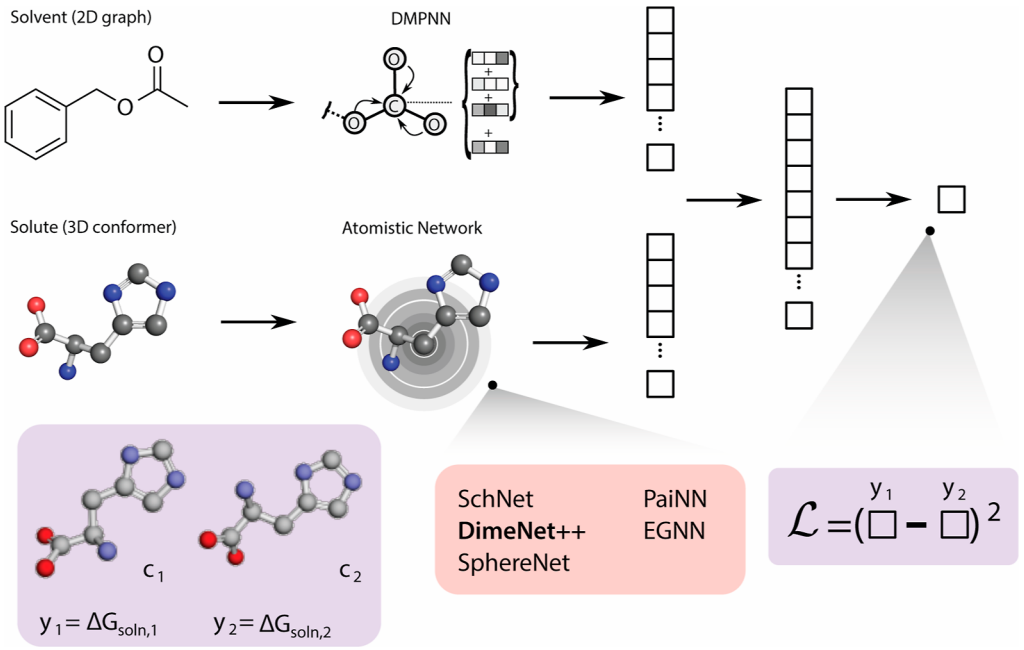


Figure 2. Illustration of machine learning model architecture employed to predict relative free energies of solution of different solute conformers in a given solvent. DimeNet++ is used as for the solutes in this work, but other architectures are also made available as well.

$$\Delta G_{\mathrm{gas},i} = E_{\mathrm{gas},i} + \Delta G_{\mathrm{RRHO},i}$$

$$\overline{\Delta G}_{\mathrm{gas}} = \Delta G_{\mathrm{gas},0} - RT \ln\left(1 + \sum_{i=1}^{(N_{\mathrm{conf}}-1)} \exp\left(\frac{-\Delta\Delta G_{\mathrm{gas},i}}{RT}\right)\right)$$

$$\Delta\Delta G_{\mathrm{gas},i} = \Delta G_{\mathrm{gas},i} - \Delta G_{\mathrm{gas},0} \tag{5}$$

We carry out this workflow for approximately 44,000 solutes and 41 solvents, leading to a large data set of just under 9 million conformer entries and 226 million energy entries; the ConfSolv data set is freely available and can be downloaded from the Supporting Information. We package this data set as two separate zipped pickle files of geometries and energies. The coordinates file `dft_coords.pkl.gz` contains the COSMO-phase optimized geometries as ASE atoms objects[62] within a pandas DataFrame.[63] Each entry is uniquely identified by `mol_id` and a `conf_id`. The energy file `free_energy.pkl.gz` contains absolute and relative values for all quantities in eq 3 in a similar format. It contains the same identifiers along with an additional identifier (`solvent`) for the solvent in which we calculated the energy. The data is made available publicly through Zenodo at https://doi.org/10.5281/zenodo.8292519.

Model Development. To circumvent the calculation of free energies for all generated conformers, we devise a model to estimate their relative free energies in solution ($\Delta\Delta G_{\mathrm{soln}}$). Our model takes as input the individual 3D geometries of the solute conformers in xyz format and the SMILES of the solvent and predicts the relative free energy contribution with respect to the lowest energy solute conformer. That is, if only a single conformer is passed to our model, it will give an output of 0. If multiple conformers are fed to the model, the predictions will be relative to the identified LEC, which will be predicted as 0.

Figure 2 illustrates the primary components of the model. We use a 2D message passing neural network to embed the solvent. Specifically, we use the directed message passing neural network (DMPNN) developed by Yang et al.,[64] which was also used by Vermeire and Green.[41] The model builds on the MPNN framework proposed by Gilmer et al.[65] by considering directed edges, such that messages across a covalent bond (edge) are passed in both directions. To embed 3D conformers, we implemented several 3D MPNN architectures. These include SchNet, devised by Schütt et al.,[66] which relies on continuous filter convolutions to incorporate information regarding atomic distances, DimeNet++, which incorporates angular information through spherical Bessel functions, SphereNet, which conducts message passing across a spherical coordinate system,[67] and PaiNN and EGNN, which are equivariant models, the former achieving equivariance through equivariant tensoral features and the latter through a coordinate update procedure. We observed DimeNet++ to be the most stable, so the results shown below were obtained by using this architecture. We concatenated the outputs of the 2D DMPNN solvent featurization and the 3D MPNN solute featurization together and feed them through a dense layer for prediction. Others[38,68] have suggested the use of advanced aggregation techniques, such as attention mechanisms, to mix feature channels between solvent and solute, but our initial experiments suggested concatenation to be more robust. Vermeire and Green[41] found that using additional features such as polar surface as computed by RDKit can improve performance, and several other studies have also explored augmenting D-MPNNs with a variety of heuristic, quantum mechanical, and chemical descriptors.[69−73] However, the outcome of using descriptors on model performance is not very consistent, and given the large number of solute conformers in this work, we do not use additional features in this work. We do note that the aforementioned studies are for 2D D-MPNNs and that exploring how additional descriptors impact the model architecture adopted here could be a topic of future work.

To train our model, we do not take an MSE loss directly from the energy predictions of the model. Since we are

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

interested in relative free energy predictions, we calculate the loss on the relative energy predictions between conformers C_i and C_j in a set of N conformers

$$\mathcal{L}(C_i, C_j) = \frac{1}{\#\{(i,j) \in N\}} \sum_{(i,j)\in N} (\Delta G_{\text{soln},i} - \Delta G_{\text{soln},j})^2 \quad (6)$$

Thus, we are concerned only with the relative energy predictions rather than the absolute values. Such an approach is often used in relative binding free energy predictions,[74] with the architecture and training approach here akin to Siamese networks,[75] with each "arm" being a 3D solute conformer in a given solution. Other methods exist to incorporate relative differences, such as subtracting the hidden representations of the atoms or molecules[76,77] or the previously mentioned attention-based approaches.[68] While such strategies could lead to slightly improved results, attention significantly impacted our model's runtime and performance using relative loss, and individual representations were observed to be sufficiently stable. So, we decided to simply condition the loss on relative solute conformer energies. During training, we sample K conformers per molecule and tune K through hyperparameter search. During inference, K can be changed depending on the number of conformers present in the molecule in question.

The model is trained using three folds, with each fold split into training, validation, and testing sets in an 88:6:6 ratio with respect to the solute molecules. These splits are created by partitioning the solutes based on their Bemis−Murcko scaffold,[78] computed from their SMILES string using RDKit.[79] Scaffold splits are a more challenging split type and provide a better measure of model generalizability.[64,77,80−82] Since we elect to split on only the solute molecules, the model is trained to make predictions for relative free energies of the solute conformers in all 41 solvents simultaneously. Although future work could explore generalizability to new solvents that are not contained in Table 1, ref 41 has already shown that splitting by solute is a more challenging task and thus a better measure of generalizability. This is an expected result because refs 4283,−8485 show that ΔG_{solv} values typically have narrower distributions across different solvents, which is consistent with our calculated data set.

The source code for all models, installation instructions, and a sample inference notebook are found at https://github.com/PattanaikL/conf_solv.

RESULTS AND DISCUSSION

Data Set Statistics. We first analyze some key molecular features for the solutes in the data set. As can be seen in Figure 3a, the data set is reasonably chemically diverse. The most common atom type is carbon, which is unsurprising given how many solutes are organic. Common heteroatoms are also present in large numbers, such as oxygen, nitrogen, and sulfur, as are fluorine, chlorine, bromine, and iodine. Silicon is also relatively common along with two of its most common dopants, boron and phosphorus.

The period 4 elements of gallium, germanium, arsenic, and selenium are also observed, all of which can form complexes with organics. The rarest atom types are titanium, tin, indium, tellurium, and antimony, corresponding to metal and metalloid elements found in only a few of the solutes.

Figure 3c shows the broad distribution of molecular weights, so the effects of solvation can be studied for a wide range of solute sizes. The most common molecular weight is near 200 Da, which corresponds to approximately 15 heavy atoms, depending on heteroatoms and degree of substitution, suggesting that the majority of solutes are of a moderate size for organics.

Figure 3d presents statistics for the number of hydrogen bond donors present in the solute molecules, computed with RDKit.[79] Hydrogen bond donors typically arise due to the presence of electronegative atoms; therefore, more hydrogen donors can be thought of as an analogue for the degree of substitution or presence of typical heteroatoms in the case of organic solutes. It is also a general indicator of the acidity of such solutes and a key parameter of interest in drug design.[86] Approximately 15,600 molecules, or 36% of the solutes, have no hydrogen bond donors, which means a majority of solutes are expected to undergo hydrogen bonding to some extent. Most of these contain one hydrogen bond donor at a count of

Table 1. Solvents Used in This Study and Dielectric Constants[87]

solvent name	chemical formula	classification	dielectric constant
perfluorohexane	C_6F_{14}	nonpolar	1.57
n-hexane	C_6H_{14}	nonpolar	1.88
isooctane	C_8H_{18}	nonpolar	1.94
dioctyl ether	$C_{15}H_{32}O$	nonpolar	2.00
cyclohexane	C_6H_{12}	nonpolar	2.02
hexafluorobenzene	C_6F_6	nonpolar	2.05
benzene	C_6H_6	nonpolar	2.28
toluene	C_7H_8	nonpolar	2.38
EMC	$C_4H_8O_3$	nonpolar	2.99
nonadecanol	$C_{19}H_{40}O$	fatty alcohol	3.82
diethyl ether	$C_4H_{10}O$	nonpolar	4.33
di-2-butylamine	$C_8H_{19}N$	nonpolar	4.71
chloroform	$CHCl_3$	nonpolar	4.8
benzyl acetate	$C_9H_{10}O_2$	polar aprotic	5.34
ethyl acetate	$C_4H_8O_2$	polar aprotic	6.02
acetic acid	$C_2H_4O_2$	small acid	6.15
triglyme	$C_8H_{18}O_4$	polar aprotic	7.50
THF	C_4H_8O	polar aprotic	7.58
tributyl phosphate	$C_{12}H_{27}O_4P$	polar aprotic	8.29
dichloromethane	CH_2Cl_2	polar aprotic	8.93
o-dichlorobenzene	$C_6H_4Cl_2$	nonpolar	9.93
octanol	$C_8H_{18}O$	fatty alcohol	10.30
ammonia	NH_3	polar aprotic	16.61
isopropanol	C_3H_8O	polar protic	17.9
butanone	C_4H_8O	polar aprotic	18.85
acetone	C_3H_6O	polar aprotic	20.7
ethanol	C_2H_6O	polar protic	24.55
diethanolamine	$C_4H_{11}NO_2$	polar protic	25.75
triethanolamine	$C_6H_{15}NO_3$	polar protic	28.11
propylene glycol	$C_3H_8O_2$	polar protic	32.0
DMF	C_3H_7NO	polar aprotic	36.7
ethylene glycol	$C_2H_6O_2$	polar protic	37.0
acetonitrile	C_2H_3N	polar aprotic	37.5
DMA	C_4H_9NO	polar aprotic	37.8
γ-butyrolactone	$C_4H_6O_2$	polar aprotic	41.68
glycerol	$C_3H_8O_3$	polar protic	46.5
DMSO	C_2H_6OS	polar aprotic	46.7
formic acid	CH_2O_2	small acid	51.1
water	H_2O	water	80.1
EC	$C_3H_4O_3$	polar aprotic	89.78
sulfuric acid	H_2SO_4	small acid	100

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

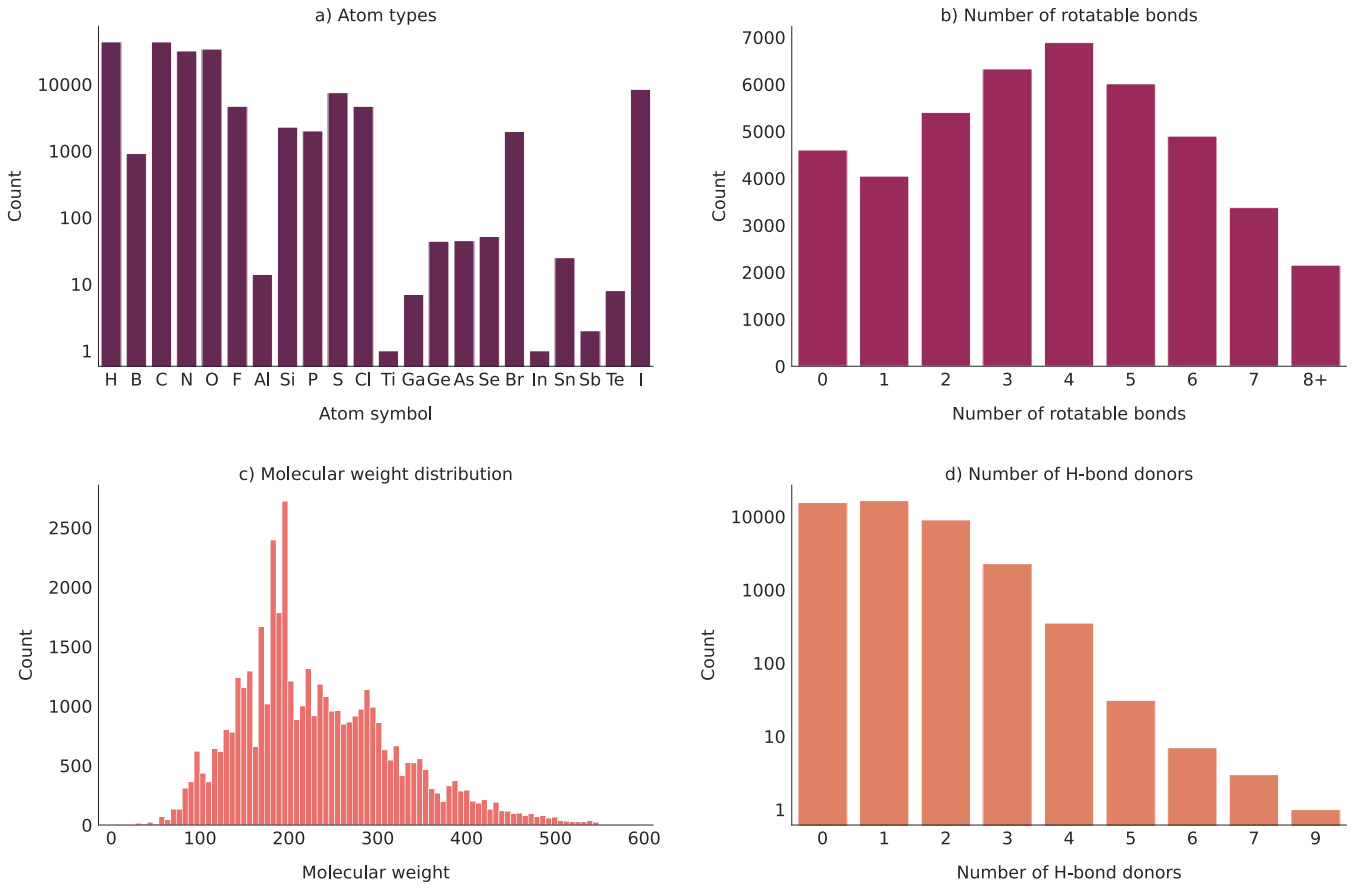


Figure 3. Key molecular features of the solutes studied in this work, including (a) atom types, (b) number of rotatable bonds, (c) molecular weight, and (d) number of hydrogen bond donors.

16,500, suggesting a single polar site, but this does mean that there are approximately 12,000 solutes with multiple hydrogen bond donor sites, which we would expect to have strong propensity for intermolecular and intramolecular hydrogen bonding. Also, a significant fraction of the solute molecules also contains hydrogen bond acceptors such as amine groups.

Figure 3b presents a histogram of the number of rotatable bonds per solute. As these bonds are used to generate the conformer set in the systematic search procedure, the number of rotatable bonds thus determines the number of conformers. Although there are a substantial number of solutes with no rotatable bonds, most solutes will have many thermally accessible conformers. To confirm this, we also display a histogram of the number of conformers for the various solute molecules and how the number of conformers varies with the number of rotatable bonds, shown in Figure 4.

Figure 4a shows the distribution of conformers for the solute molecules in the data set. Approximately 16,000 solutes (36%) have between 0 and 10 conformers. A similar number (~15,000) of solutes have between 10 and 100 conformers, and a further 10,000 solutes have between 100 and 1000 conformers. This leaves approximately 2200 solutes with more than 1000 conformers, corresponding to very flexible molecules. Figure 4b confirms that the average number of conformers generated for a given solute molecule increases with the number of rotatable bonds, as expected of the methodology. However, we do see some solutes with a large number of rotatable bonds but few conformers. This could be due to certain conformers being too similar to the LEC, connectivity changes occurring during the generation process, or because conformers were missed by the coarse grid. Nevertheless, the distribution and trend of number of conformers with number of rotatable bonds suggests that the conformer generation method produces a useful broad data set.

Table 1 enumerates all of the solvents used in this study along with their chemical formulas, broad solvent classifications, and dielectric constants (at 298 K). The last column serves as a proxy of the solvent's polarity, which strongly affects the solvation free energy values computed by COSMO-RS. For polar solvents, we make a distinction between protic and aprotic solvents, classifying on their mode of hydrogen bond donation. We expect that protic solvents readily donate hydrogen bonds to any solute that can accept them, which polar aprotic solvents cannot, thereby providing a point of comparison. We note that polar aprotic solvents can be strong hydrogen bond acceptors, and given that several solutes have one or more hydrogen bond donors or hydrogen acceptors, this should increase the capacity for solute−solvent interactions. The set also includes a few solvents from other common categories such as fatty alcohols and small acids. We denote water with its own category, as it has unique properties. These solvents represent common solvents used in chemical synthesis and manufacturing and should cover a wide range of interactions and ΔG_{soln}.

$\log_{10}P_{OW}$ Validation. We next set out to validate our data set by calculating experimental observables and comparing to existing literature. We extracted a data set of experimental water−octanol partition coefficients ($\log_{10}P_{OW}$) gathered by

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

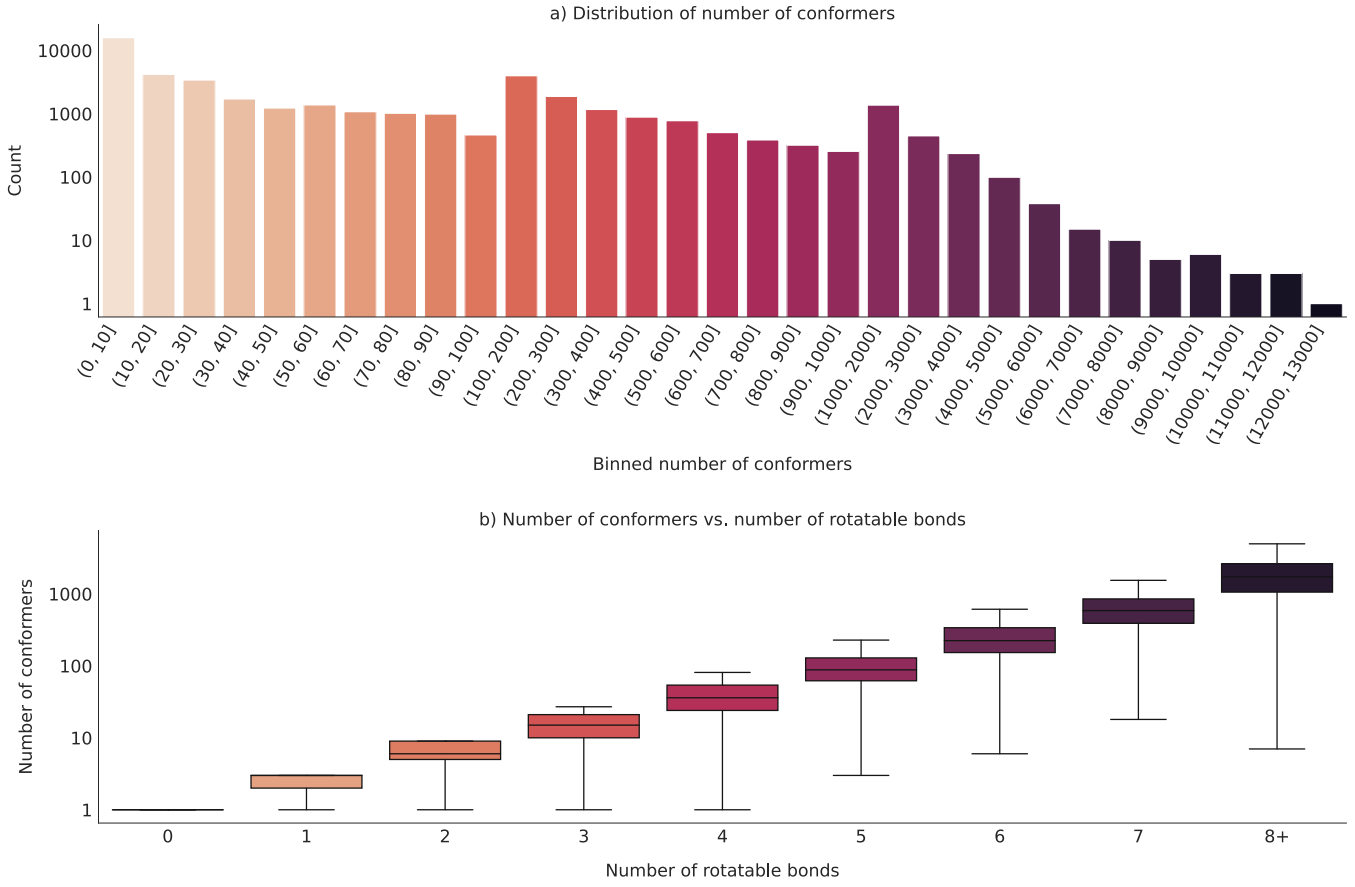


Figure 4. (a) Distribution of number of solutes with number of identified conformers and (b) number of solute conformers plotted against number of rotatable bonds.

Mansouri et al.[88] and compiled by Ulrich et al.[89] From their nearly 13,000 values, 365 solutes matched those in our data set. For these 365 solutes, we calculated $\log_{10}P_{OW}$ and compared it against their experimental values. To investigate the effect of conformers on this calculation, we calculated $\log_{10}P_{OW}$ using several different methods. Note that we use dry octanol and do not consider differences in protonation state or isomers of the solute when in octanol or water. Methods 1 and 2 use the LEC in water and octanol (as determined by ΔG_{soln}), respectively, and are thus single conformer approaches.

Method 3 uses Boltzmann-ensembled corrected values for ΔG_{soln} and ΔG_{gas} computed across the set of conformers generated for the solute in question. ΔG_{solv} is then calculated as the difference between these two values. The Supporting Information provides more details on all methods, with the results for method 3 shown in Figure 5.

From Figure 5, method 3 shows reasonable agreement with the experimental data, especially given the expected errors due to the assumptions made. The difference between predicted and experimental logP values is usually well within 1 log unit, suggesting that the COSMO-RS method is reasonable for relative solvation predictions, as noted by Kundi and Ho.[90] We can also compare this performance to single-conformer methods 1 and 2 and another single-conformer approach in which we take the solute conformer with the lowest gas-phase energy and compute its solution free energy in water and octanol to calculate logP, hereafter referred to as "Gas". We focus on solutes with at least 50 conformers. This is shown in Figure 6.

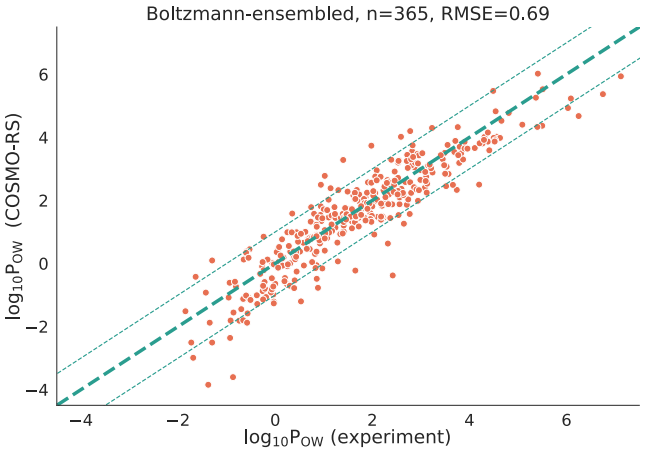


Figure 5. Parity plot comparing true and COSMO-computed values of $\log_{10}P_{OW}$ using a Boltzmann-ensembled method. The dashed lines show 1 log unit above and below the parity line.

Method 3 has the narrowest error distribution, confirming that agreement with experimental logP is generally improved when multiple conformers. However, using the LEC in water (method 1) also performs reasonably well, and using the LEC in octanol (method 2) is only slightly worse. This highlights the ultimate importance of finding a good LEC, but there are notable cases where single-conformer approaches perform very poorly, and method 3 is generally better. In general, using the LEC in the gas phase results in a notably broader and off-center error distribution, suggesting that including solvation

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

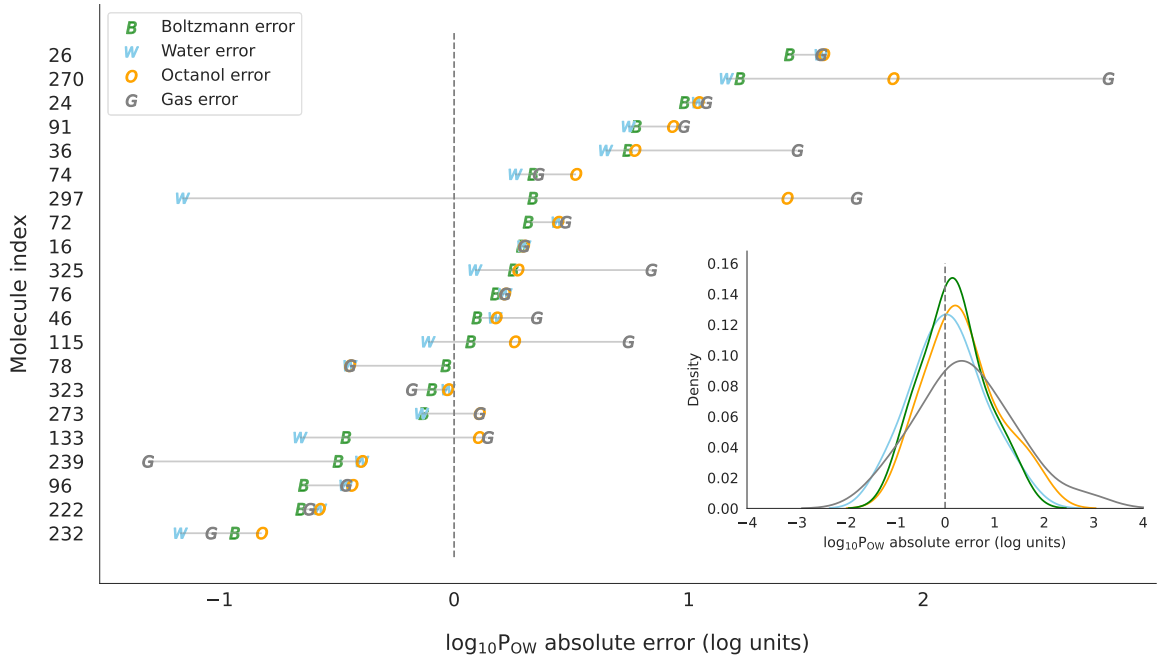


Figure 6. Absolute error in $\log_{10}P_{OW}$ using all conformers (B) and three different single-conformer methods. The "all-conformers" method (B) is the most reliable, but single-conformer methods based on the lowest-energy conformer (LEC) in either water (W) or octanol (O) usually have comparable accuracy. Using the gas-phase LEC (G) is not recommended; it gives poor predictions for almost half of the solutes.

effects is necessary when ranking the conformers. This is further seen in Figure S3, where the error in logP when using the gas-phase LEC increases as the solute has more conformers, and in Figure S4, where we see the difference between the Boltzmann and gas-phase LEC approach is even larger when looking at a nonpolar solvent (cyclohexane), partitioning with water. Using the respective LEC in water and octanol to compute ΔG_{solv} results in a very poor performance, highlighting that a consistent conformer choice is required.

Impact and Identification of Relevant Conformers. The $\log_{10}P_{OW}$ validation suggests that conformers can impact free energy calculations and that it is important to have good estimates for the LEC and low-energy conformers. In this section, we aim to quantify the impact of conformational effects on the overall solution free energy. We do this by analyzing a representative 2200 solute subset of the full data set. As a guideline of when solute conformer contributions are substantial, we look for when the 1 kcal/mol threshold is reached. We do note that for several systems, much lower differences in free energies than 1 kcal/mol can be substantial, but as 1 kcal/mol is often used in computational chemistry for "gold-standard" methods such as CCSD(T),[91] it should still provide some helpful insights here. First, we check whether or not considering solvation free energies is necessary to accurately calculate $\overline{\Delta G}_{soln}$ by considering only gas-phase conformer rankings. Following Gorges et al.,[92] we define an additive correction corresponding to the variation in the free energy introduced by considering multiple conformers

$$\Delta G_{soln,E_0,corr} = \overline{\Delta G}_{soln} - \Delta G_{soln,E_0} \tag{7}$$

Here, $\Delta G_{soln,E_0}$ is the free energy in solution of just one conformer, the conformer that is the LEC in the gas phase. Thus, $\Delta G_{soln,E_0,corr}$ measures the difference between the solvated solute's free energy computed by considering a whole ensemble of conformers, and the much simpler calculation using only the LEC identified only by gas-phase energies. We compute this value for all solute−solvent pairs. Figure S6 shows this analysis across all solvents. Nearly all solvents show mean deviations from $\overline{\Delta G}_{soln}$ near 1 kcal/mol when averaged across all solutes. Notably, the 75th percentile of the correction exceeds 2 kcal/mol for all solvents, and there are a substantial number of solute−solvent combinations that show errors beyond 5 kcal/mol. Four solvents, in particular—di-2-butylamine, ammonia, formic acid, and sulfuric acid—have mean gas-phase correction terms that are notably higher than 1 kcal/mol, and several pairs have corrections in excess of 10 kcal/mol. These solvents correspond to two stronger bases and two stronger acids, highlighting that in strong solvents it is inaccurate to consider only gas-phase rankings of conformers. Overall, the analysis suggests that just using the LEC from the gas phase to estimate $\overline{\Delta G}_{soln}$ is likely to result in substantial errors.

Errors from using just the LEC from the gas phase are often due to neglecting energy changes in other conformers due to solvation. Figure S7 shows that often the LEC is not the same in the gas phase and in solution, which would then impact the one-conformer ΔG_{soln} estimate when compared to the Boltzmann-ensembled $\overline{\Delta G}_{soln}$ value. Given that solvation corrections seemingly impact this, we now aim to quantify this effect. First, we define a correction term analogous to that in eq 7

$$\begin{aligned}\Delta G_{soln,corr} &= \Delta G_{soln,0} - \overline{\Delta G}_{soln} \\ &= RT\ln\left(1+\sum_{i=1}^{(N_{conf}-1)}\exp\left(\frac{-\Delta\Delta G_{soln,i}}{RT}\right)\right)\end{aligned} \tag{8}$$

Here, $\Delta G_{soln,0}$ is the free energy of the LEC in each solvent. Computing this value per solvent across all solutes yields Figure S8, which shows a consistent offset of approximately 0.75 kcal/mol. That is, considering only the LEC in that

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

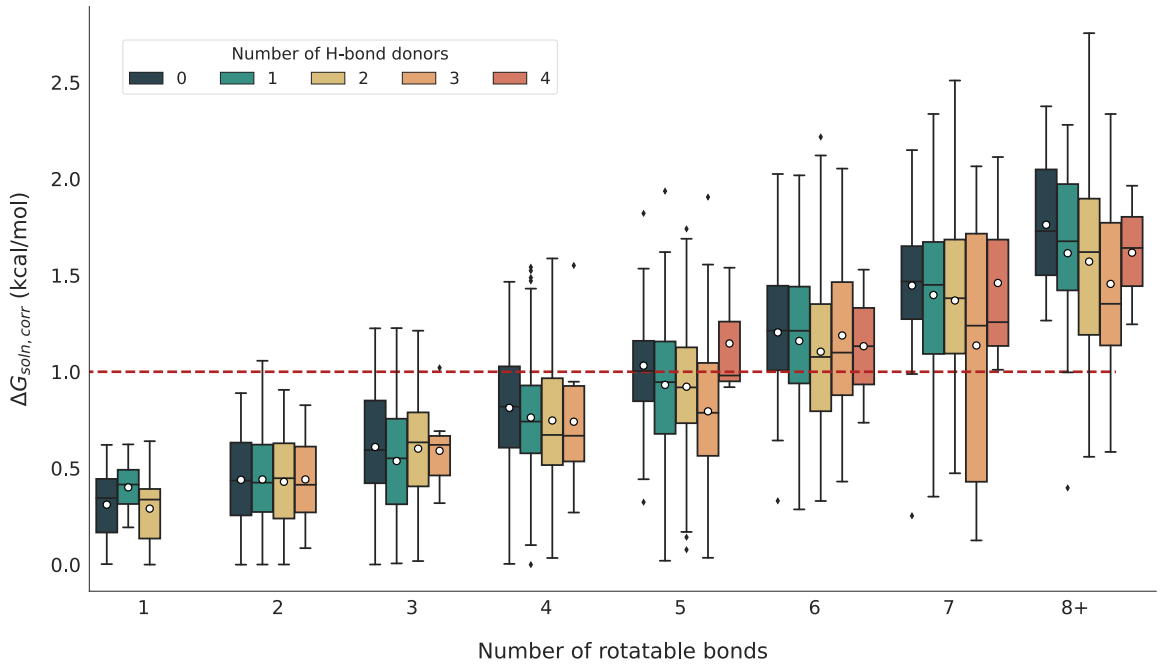


Figure 7. $\Delta G_{soln,corr}$ in water grouped by an increasing number of rotatable bonds and hydrogen bond donors.

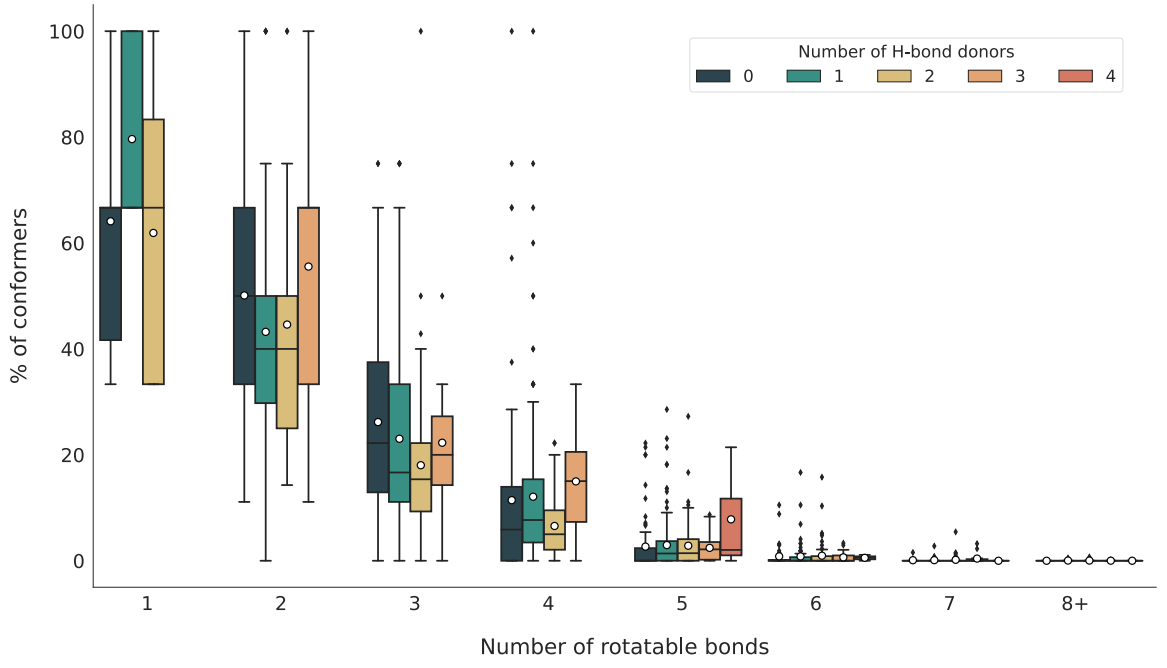


Figure 8. Percent of conformers in water with ΔG_{soln} within 1 kcal/mol of $\overline{\Delta G}_{soln}$ grouped by number of rotatable bonds and H-bond donors.

solvent is usually sufficient to recover the full Boltzmann-ensembled $\overline{\Delta G}_{soln}$ within 0.75 kcal/mol. An offset of 0.75 kcal/mol is fairly small, indicating that considering only the LEC in solution will often be sufficient for an accurate ΔG_{soln} calculation at room temperature. The decrease in error compared to the earlier analysis using the LEC in the gas phase emphasizes the importance of the ΔG_{solv} term to accurately rank conformers in solution phase.

Since Figure S8 shows *average* results across all test solutes, it may obscure some trends in the effect of conformers. We can further analyze the effect of conformers by dividing solutes along known divisions that should influence this correction term. For example, we expect that for flexible solutes, the correction term should be larger, as a greater number of thermally accessible conformers are expected to contribute significantly to the ensemble. . Similarly, for solutes that can hydrogen-bond with either themselves or the solvent, this correction term would be expected to change as well. Figure 7 divides solutes by these two descriptors in order to investigate their impact on the free energy correction term in water.

Figure 7 suggests that for flexible solutes with an increasing number of rotatable bonds, conformers begin to show stronger contributions, and thus, the correction term increases. For example, the conformer correction is often above 1 kcal/mol for molecules with more than four rotatable bonds. However,

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

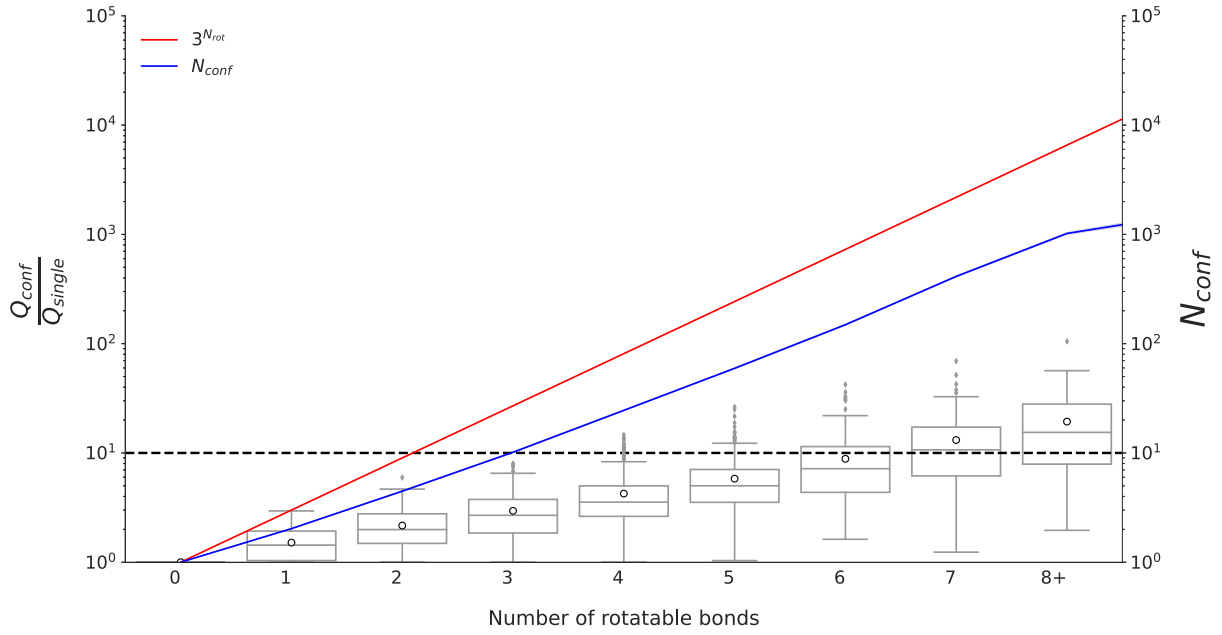


Figure 9. Partition function ratio for solutes in water grouped by the number of rotatable bonds. The blue and red lines show the total number of conformers found in this work, and a theoretical maximum number of conformers using the systematic search method, respectively.

given that the maximum conformer contribution to the solution free energy is around 2.5 kcal/mol, using only the LEC may provide sufficiently accurate estimates of $\overline{\Delta G}_{soln}$ at room temperature in many cases, even for molecules with a substantial number of rotatable bonds and conformers. There is also a slight negative trend with an increasing number of H-bond donors, which becomes more pronounced as the number of rotatable bonds increases. This indicates that the greater propensity to form hydrogen bonds tends to decrease the conformer contribution term. This would suggest that solutes with a higher propensity to form hydrogen bonds have fewer conformers contributing to the free energy change of solution beyond the LEC. This suggests that a small number of conformers can stabilize by hydrogen bonding, and thus a smaller subset of the total conformers are energetically favorable in water. However, this effect is neither very consistent across the number of rotatable bonds nor very substantial in most cases; the LEC is still the most important.

Figure S9 applies the same analysis to a nonpolar solvent (*n*-hexane), which shows very similar magnitudes of corrections to those with water as the solvent. This is consistent with the results when averaging all solutes in each solvent, i.e., the LEC and set of low-energy conformers is crucial in any solvent regardless of polarity. However, there is a slight difference seen for *n*-hexane in that the decrease in correction with an increasing number of hydrogen bond donors is substantially more pronounced than for water. One would expect nonpolar solvents to promote intramolecular hydrogen bonding and solute−solute interactions. The decrease in correction to ΔG_{soln} again suggests that only a very small subset of conformers, including the LEC, are energetically relevant. The promotion of internal hydrogen bonding could result in certain conformers stabilizing by hiding their polar sites, resulting in an even smaller favorable set of conformers than a polar solvent promoting solute−solvent interactions. In any case, whether the solvent is polar or nonpolar, identifying the LEC is crucial for accurate ΔG_{soln} calculations and can even be sufficient in itself for solutes with few rotatable bonds.

While this analysis shows that the LEC is sufficient to reasonably estimate $\overline{\Delta G}_{soln}$, it does not reveal how many other conformers meet this criterion. We subsequently computed the percentage of conformers within our data set that have values of ΔG_{soln} within 1 kcal/mol of the Boltzmann-ensembled $\overline{\Delta G}_{soln}$. Aggregating these distributions across all solvents returns Figure S10. This figure illustrates that on average, 15% of conformers have values of ΔG_{soln} close to $\overline{\Delta G}_{soln}$. Again, this is averaged over all solutes for each solvent, which may obscure more nuanced trends. Hence, Figure 8 again focuses on this analysis in water.

The two selected variables for this case are again the number of rotatable bonds and the number of hydrogen bond donors in the solute molecules. As expected, the percentage of relevant conformers decreases as the number of rotatable bonds increases, essentially indicating that for flexible solutes with a large number of conformers, a lower percentage will be within the 1 kcal/mol threshold. However, this could be offset by the larger number of conformers in these solutes. Figure S11 shows the absolute number of conformers within this 1 kcal/mol threshold. For small solutes with a lower number of rotatable bonds, a larger percentage of the conformers are within the threshold, corresponding to two to three energetically close conformers that need to be taken into account. For solutes with more than four rotatable bonds, the percentage of conformers within the threshold is so low that either only the LEC or no conformers have solution free energy changes close enough to the Boltzmann-ensembled value. This is an interesting result. Effectively, this means that solutes with many flexible bonds have several thermally stable conformers, and many of these conformers meaningfully influence $\overline{\Delta G}_{soln}$. In this case, our analysis demonstrates the need to not only identify the LEC but to identify several low-energy conformers if we require the final prediction to be within 1 kcal/mol of $\overline{\Delta G}_{soln}$.

Overall, the findings emphasize the importance of a conformer search, especially for larger solutes. In general,

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

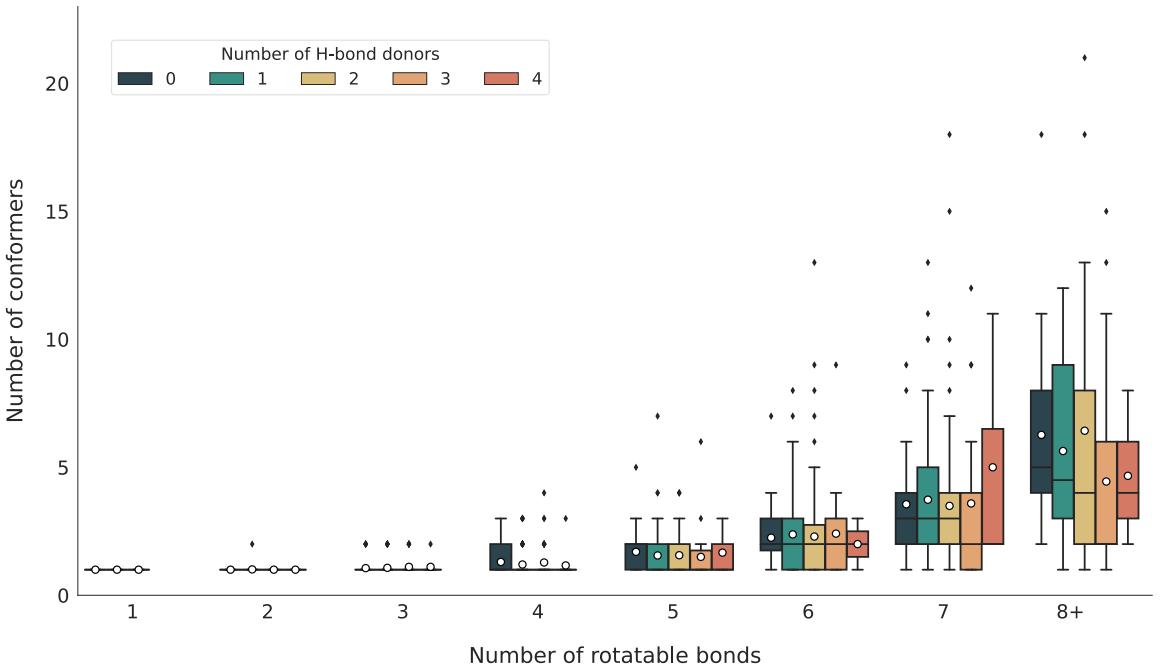


Figure 10. Number of conformers needed to calculate ΔG_{soln} within 1 kcal/mol of the Boltzmann-ensembled $\overline{\Delta G}_{soln}$ in water grouped by number of rotatable bonds and H-bond donors.

locating the LEC or one of the few other low-energy conformers below the threshold should allow the prediction of $\overline{\Delta G}_{soln}$ to be robust. This is especially true for small molecules, where the ΔG_{soln} for the LEC and low-energy conformers is suitably close to $\overline{\Delta G}_{soln}$. For larger and very large solutes with a large number of rotatable bonds, even the LEC is not within 1 kcal/mol of $\overline{\Delta G}_{soln}$, but it is nonetheless the best single-conformer estimate. Identifying the LEC is still key in these cases, as it guides the conformer search and the computation of the free energies. For the select cases where the LEC is not sufficient to determine $\overline{\Delta G}_{soln}$ within 1 kcal/mol, we can compute the partition function ratio $\frac{Q_{conf}}{Q_{single}}$ and compare this to the number of conformers to understand what fraction of conformers is expected to be important to consider. Figure 9 displays this analysis in water.

Figure 9 compares the partition function ratio to both the number of conformers located for the solute by the systematic search and a theoretical maximum number of conformers that could be located, estimated here as $3^{N_{rot}}$, where N_{rot} is the number of rotatable bonds identified for the solute. As expected, all the three values increase with an increasing number of rotatable bonds, with the contribution of conformers above the LEC to the partition function being an order of magnitude above the LEC for solutes with six rotatable bonds and more. As the partition function ratio is relative to the single-conformer case, it effectively represents the contributions of additional conformers above the LEC. Therefore, this ratio being substantially lower than both the number of conformers identified and the theoretical maximum confirms that even when estimating solution thermodynamic properties with a single conformer is insufficient, an accurate estimate should require only a small subset of the total number of conformers identified for the solute.

With this in mind, we can determine how many conformers are actually necessary to compute $\overline{\Delta G}_{soln}$ to within 1 kcal/mol accuracy. To investigate this question, for each molecule and solvent, we ordered conformers by increasing values of ΔG_{soln} and computed a cumulative value of $\overline{\Delta G}_{soln}$. We then subtracted this cumulative value from the true value of $\overline{\Delta G}_{soln}$, akin to the correction term described in eq 8. The point when this difference is under a threshold value of 1 kcal/mol denotes the number of conformers required to compute ΔG_{soln} reasonably accurately. Figure S12 shows this analysis aggregated across solutes for each solvent. For the vast majority of solutes, only 1 or 2 conformers are necessary. However, for select solutes, up to 20 conformers might be required for accurate calculations. Figure 10 again focuses this analysis on water.

The trends from Figure 10 are as expected. For smaller solutes with fewer than four rotatable bonds, the LEC usually suffices for an accuracy of 1 kcal/mol. However, for more flexible solutes, a slightly larger number of low-energy conformers must be computed to achieve this accuracy. Similarly, there is a slight decrease in the number of conformers required with the number of hydrogen bond donors, which is also expected, as this was seen with the correction term as well. Figure S13 shows the results for *n*-hexane. The trends are very similar to water, with generally only 1 or 2 conformers being required. There is a more pronounced decrease in the number of conformers required with the number of hydrogen bond donors, consistent with what we have previously observed for *n*-hexane. As with our previous analyses, these results highlight the need to precisely determine the LEC for small solutes and a small number of low-energy conformers for larger solutes.

Effect of Conformer Search Grid. Given the highlighted importance of finding the LEC when computing thermodynamic properties, it is clear that how the conformer search is performed is also an important factor. As detailed in the

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

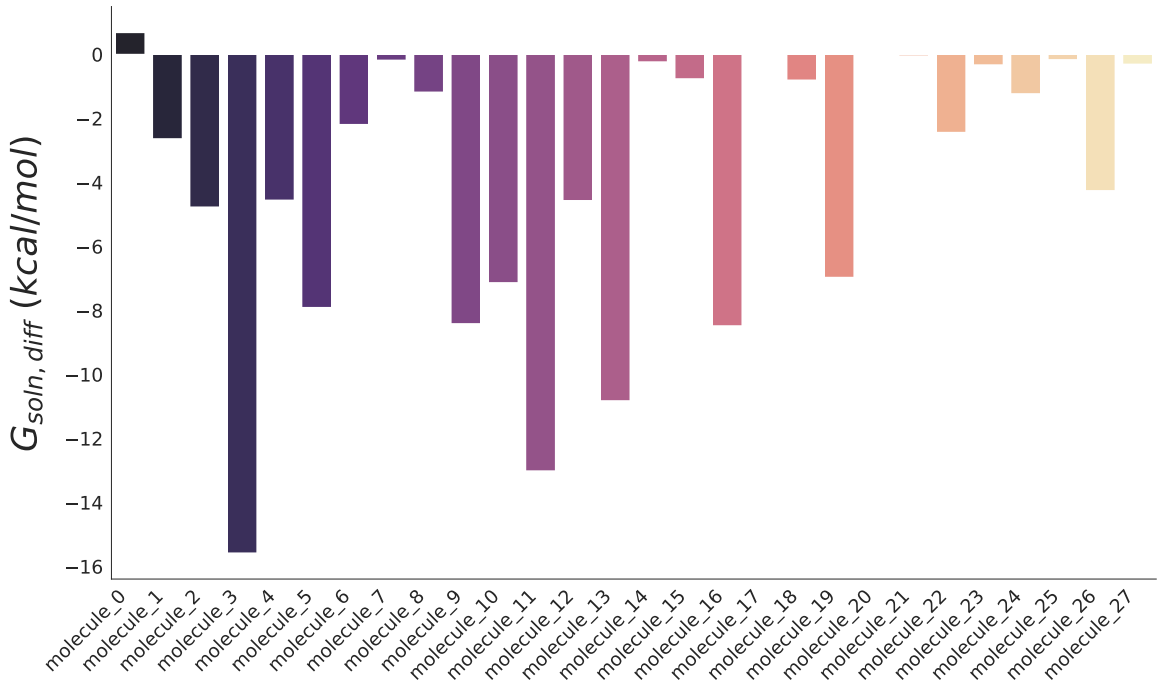


Figure 11. Difference between Boltzmann-ensembled $\overline{\Delta G}_{soln}$ computed using conformer sets derived from the refined and base conformer searches. Results are computed in water. Table S1 shows the SMILES strings for these 28 molecules.

Methodology section, the conformer search here follows a systematic approach but with a coarse grid. To gauge how this would impact the set of conformers discovered and thus the solution free energies and partition functions, we take a small subset of 28 solutes and perform the systematic search with a more refined grid. Specifically, rotatable single bonds, sp^2-sp^2 single bonds, and double bonds in the molecule are systematically rotated in step sizes of 60, 120, and 120°, respectively, in the refined systematic search in comparison to 120, 180, and 180° in the base conformer search. The difference in number of conformers, $\frac{Q_{conf}}{Q_{single}}$, and $\overline{\Delta G}_{soln}$ can then be computed, with the difference defined for all three as $X_{refined} - X_{base}$. The differences in $\overline{\Delta G}_{soln}$ are shown in Figure 11.

Figure 11 shows a substantial difference in the computed solution free energies between the refined and base systematic conformer searches. In general, $\overline{\Delta G}_{soln}$ for the refined conformer search is lower than for the base conformer search. The differences are also usually well above 1 kcal/mol and in several cases even above 10 kcal/mol, clearly highlighting the impact the conformer search can have on the predicted solution free energy. There are several factors contributing to this. The first is due to differences in the number of conformers the refined and base conformer search find, shown in Figure S14. Due to having a smaller step size and thus finer grid, the refined search generally finds many conformers that the base conformer search misses, with it sometimes finding over 1000 additional conformers. Cases where no or very few new conformers were located tend to correlate with lower differences in the solution free energy between the two searches. However, as we noted previously, only a small subset of conformers are thermodynamically relevant, and so we also look at the difference in ΔG_{soln} between the LECs found between the refined and base method, shown in Figure S15. This illustrates that the refined method generally finds a better LEC, with the new LEC being over 1 kcal/mol lower in energy than the base method in some cases, but the differences are noticeably less pronounced than the overall $\overline{\Delta G}_{soln}$. This suggests that the refined search can find multiple relevant low-energy conformers, including a new LEC that is missed by our base method. Therefore, although we train a model on the base conformer search method, we also evaluate the model on the set of conformers located by the refined search for the 28 solutes. While the model will have seen these 28 solutes before, analyzing the conformer searches shows that several conformers and even a new LEC are present, so this can be thought of as a conformer split evaluation.

DMPNN Model Performance and Results. We have shown the importance of identifying the LEC and low-energy conformers in order to accurately predict ΔG_{soln}. We have also shown that low-energy conformers tend to differ between the gas phase and solution phase, and using low-energy gas-phase conformers to predict $\overline{\Delta G}_{soln}$ can result in quite large errors, but using the LEC and a small number of low-energy conformers in solution is a reasonable approximation. Thus, for the trained ML model to be helpful, it should be able to identify the LEC or a suitable subset of low-energy conformers, thereby saving the computational cost of running energy, frequency, and solvation calculations for the unimportant conformers. To check this, we assess the agreement between the model's predictions of the set of low-energy conformers and the true set of low-energy conformers as determined by QM calculations. First, we can compute the model's test set error for $\Delta\Delta G_{soln}$ directly, which can be seen in Figure S16. The consistently low MAE across nearly all solvents suggests good model performance on relative solution free energies. The computed R^2 value between the prediction and ground truth values is 0.85, also suggesting good performance on the scaffold-split test set. However, we also judge model performance for the ensemble of solute conformers and gauge how the

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

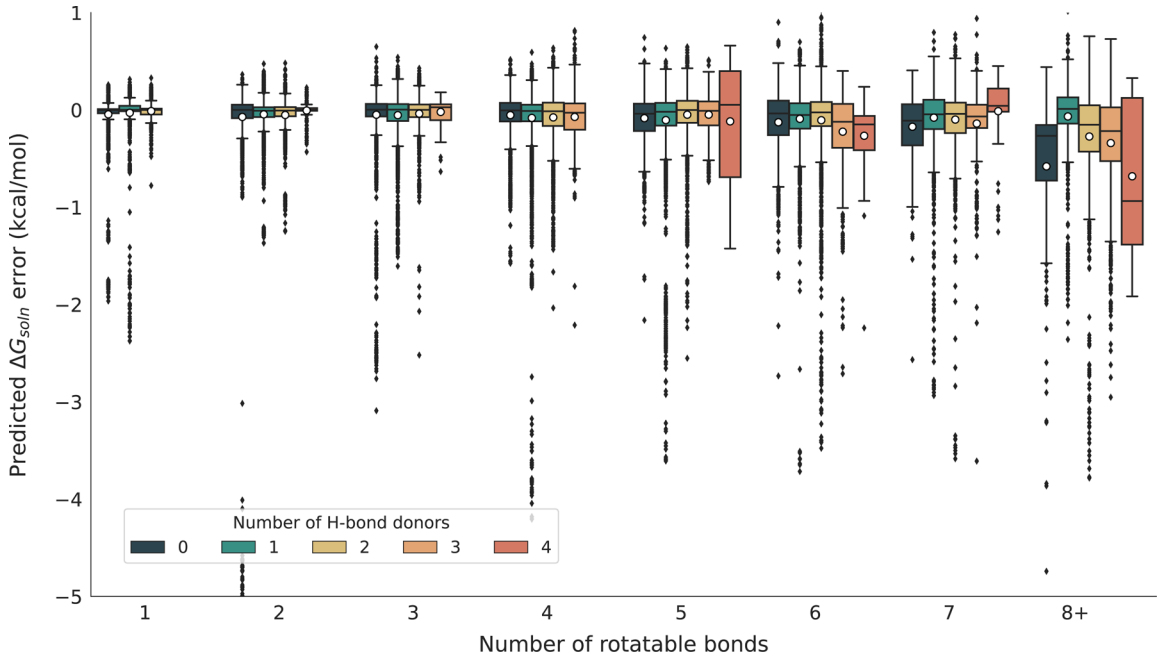


Figure 12. Difference between the true and model-predicted values of ΔG_{soln} separated by the number of rotatable bonds and H-bond donors in the solutes aggregated over all solvents.

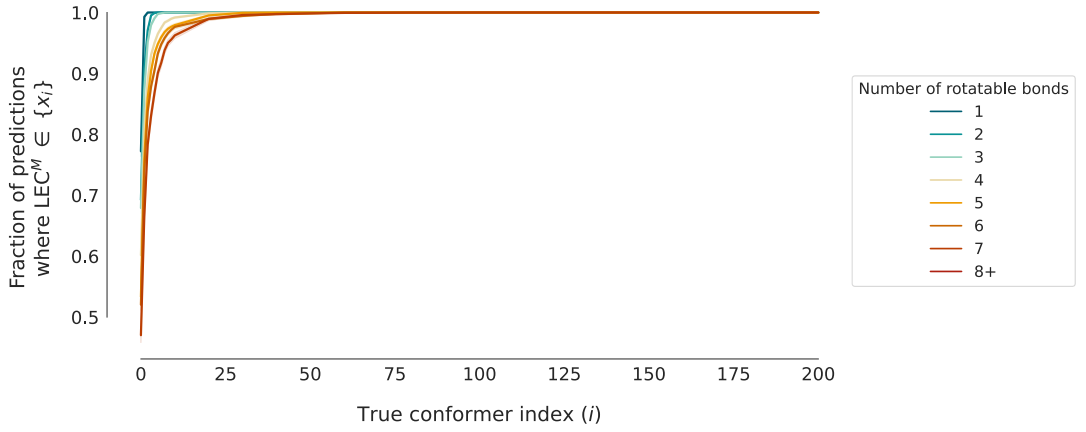


Figure 13. Ranking of the model's predicted LEC aggregated across all solvents separated by the number of solute rotatable bonds.

relative errors for each conformer impact the overall solution property prediction. To do this, we make use of the values of $\Delta\Delta G_{soln}$ predicted by the model to compute a predicted Boltzmann-ensembled $\overline{\Delta G}^{M}_{soln}$. For this prediction, we additionally need the actual computed value of $\Delta G^{M}_{soln,0}$. Note that this is the true free energy change of solution for the conformer, which the model identifies as the LEC, which may be distinct from the true LEC. We can then compute a predicted Boltzmann-ensembled $\overline{\Delta G}^{M}_{soln}$ as follows

$$\overline{\Delta G}^{M}_{soln} = \Delta G^{M}_{soln,0} - RT\ln\left(1+\sum_{i=1}^{(N_{conf}-1)} \exp\left(\frac{-\Delta\Delta G^{M}_{soln,i}}{RT}\right)\right) \quad (9)$$

This value can then be compared to the true Boltzmann-ensembled value, $\overline{\Delta G}_{soln}$, enabling us to assess how the differences in ΔG_{soln} between conformers predicted by the model compare to the calculated values across the whole conformer set. We train the model across three different folds, but as the performance was very similar in each case, we show the results of one fold as a representative result in Figure 12.

Figure 12 confirms that the model predicted $\overline{\Delta G}^{M}_{soln}$ are in good agreement with the true calculated values, with very low differences between the two. The trends are identical to what is seen previously, with the average error being below 0.5 kcal/mol and differences being well below 1 kcal/mol in the majority of cases. Errors and number of outliers broadly increase for solutes with a larger number of rotatable bonds and for those with four hydrogen bond donors. Although there are cases where the errors do exceed 4 kcal/mol, a vast majority of solute−solvent pairs are described well by the model. We can also compute the Spearman correlation coefficient of the predicted values of ΔG_{soln} against the true values, which is displayed in Figure S17. The rank correlation coefficients are close to 1 for a majority of cases with narrow distributions, suggesting that the model is able to capture the relative differences in ΔG_{soln} between solute conformers across the range of solvents. Essentially, this means that the full

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

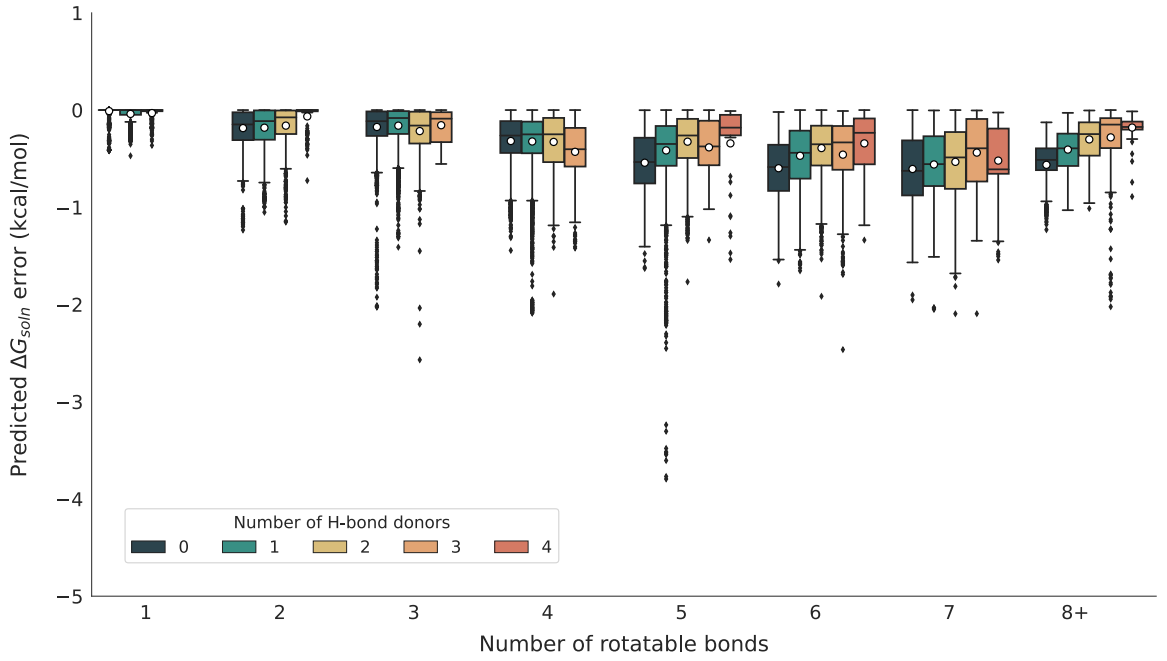


Figure 14. Difference between the true value of ΔG_{soln} and the predicted value of ΔG_{soln} when using only the small subset of conformers identified as relevant by the ML model. The error is shown separated by the number of rotatable bonds and H-bond donors in the solutes aggregated over all solvents.

workflow only needs to be applied to the model predicted LEC for a reasonable estimation of the true Boltzmann-ensembled value, $\overline{\Delta G}_{soln}$.

In some cases, the model predictions lead to errors approaching several kcal/mol. A reason for this discrepancy is the model's inability to identify the LEC in solution. To check this, we can order the conformers by their true values of $\Delta\Delta G_{soln}$, defined in eq 4 and check the index of the predicted LEC. Rather, we can increase the index of ordered conformers and ask whether the model-predicted LEC falls within the set of conformers up to the considered index. Figure 13 shows this analysis.

In most cases, the model-predicted LEC is within the first 25 true LECs, although there is a trend in the quality of the model-predicted LEC with the number of rotatable bonds present in the solute. This is expected, as for a very low number of rotatable bonds, the total number of conformers is low, and thus, the prediction task is simpler. However, as the number of rotatable bonds increases, the model-predicted LEC's true index increases, meaning that the discrepancy between the true and predicted LEC is also larger. For these more flexible solutes, the model-predicted LEC tends to fall within the first 50 conformers instead. These larger discrepancies are likely the cause of the poorer model predictions seen in Figure 12, as we have established that correctly identifying the LEC is crucial to reasonably estimating $\overline{\Delta G}_{soln}$.

Figure 13 initially suggests that the model's applicability may be limited for larger solutes. However, it also motivates us to consider a related statistic—the model-predicted index of the true LEC. This enables us to estimate the size of the subset of model-predicted conformers required for reasonable certainty that the true LEC conformer is included in this subset. In effect, we use the model to down select the number of conformers needed in the full calculation workflow while retaining some confidence that the true LEC is included. This allows computational savings as we can skip computations on all conformers deemed unessential. One can also calculate an estimated Boltzmann-ensembled $\overline{\Delta G}_{soln}$ using just the subset of conformers from the model prediction and compare this to the full Boltzmann-ensembled $\overline{\Delta G}_{soln}$ to gauge improvement.

The statistics for the model-predicted index of the true LEC are shown in Figure S20. One can then use this to define a threshold fraction of 0.9 to identify a subset of conformers that is sufficiently likely to include the true LEC. An estimated Boltzmann-ensembled $\overline{\Delta G}_{soln}^{small}$ over this subset can then be calculated and compared to the true value as discussed above. The results of this analysis are shown in Figure 14.

Figure 14 shows that the agreement between $\overline{\Delta G}_{soln}^{small}$ and $\overline{\Delta G}_{soln}$ is good. The mean error is below 1 kcal/mol for the majority of solute−solvent pairs and below 3 kcal/mol for all pairs. This is a more consistent performance than that directly using the model predictions. This is expected to an extent, as $\overline{\Delta G}_{soln}^{small}$ makes direct use of computed quantities from QM methods and COSMO-RS as opposed to just model predictions. However, this still would suggest that even for cases where the trained model's direct prediction of $\overline{\Delta G}_{soln}$ may not be reliable, like for large solvents, the model's prediction of the set of low-energy conformers is reliable. This means that the model can still help save computational time by identifying a smaller subset of conformers that one has to undergo the full computational workflow for, with the size of these subsets being $O(1)$ for solutes with up to three rotatable bonds, $O(10)$ for solutes with four to six rotatable bonds, and $O(100)$ for solutes with seven or more rotatable bonds. All of these represent a significant reduction in the number of conformers to compute fully when considering the total number of conformers for each of these solutes.

As one final evaluation of the model performance, we also make predictions on the conformer set for 28 selected solutes

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

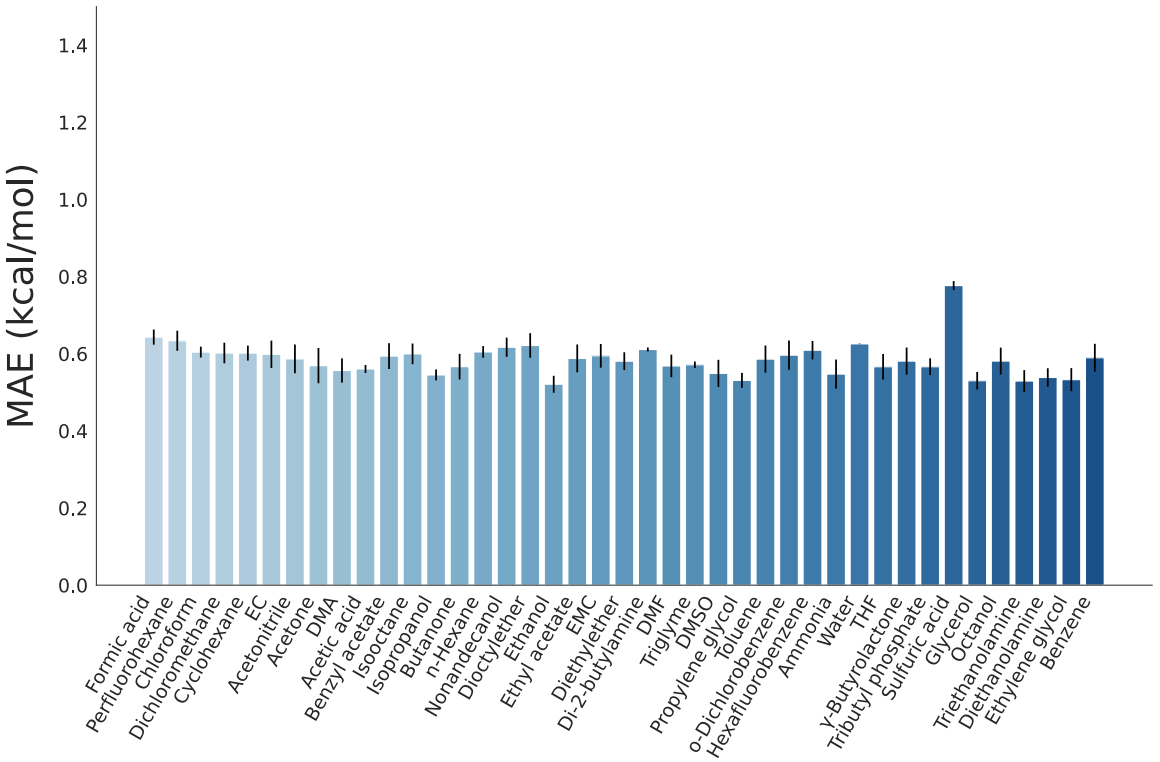


Figure 15. MAE in $\Delta\Delta G_{soln}$ for three models trained on different scaffold split folds evaluated on the refined systematic grid search conformer set. The error is shown for each solvent, and the error bars represent the standard error.

derived by the refined grid search method discussed above. The chemical identifiers for these solutes are given in Table S1. As a comparison between the conformer sets derived from the refined grid search and base grid search suggested that the base grid search finds fewer conformers and can also miss a better LEC, it is important to gauge how the model handles new conformers for the solutes. The performance of three models trained on different folds of scaffold splits is shown in Figure 15.

Figure 15 shows that despite being trained using the base search conformer sets, the model's errors on the conformers found by the refined systematic search are still well under 1 kcal/mol for most solvents. The MAEs are higher than those observed in Figure 12, but they still suggest that the model's evaluation on unseen conformers is reasonable. As a consequence, even if the model is trained on a conformer set generated by a more modest conformer search method, it can still provide reasonable evaluations for conformers generated by a more detailed conformer search. Additionally, Figure 15 includes the full set of conformers, whereas the previous analyses all confirm that the LEC and low-energy conformers are the crucial ones for the model to evaluate accurately. We therefore analyze the model performance on the true set of the 15 LECs for each of the solutes, to evaluate how it handles the conformers expected to contribute most to the thermodynamic properties. The threshold of 15 conformers is chosen on the basis of Figure 10. This is shown in Figure S21. Figure S21 shows that the MAE of the model on the true set of the 15 lowest conformers is approximately 0.2 kcal/mol lower than when every conformer is included. This confirms that the model's evaluation of the key LECs is reasonable, and thus, it is expected that the computed thermodynamic properties from the model predictions will also be good for initial estimates and evaluation even on new conformers not previously seen.

All in all, this gives us confidence that given a set of DFT conformer geometries for the solute, the model can sufficiently identify which conformers are expected to be low energy in solution. This means that a value of ΔG_{soln} could only be computed for one conformer in the averaged to get a reasonable initial estimate, saving computational time in terms of solvation calculations, single-point energy calculations, and frequency calculations for the entire set of conformers. If greater accuracy is required, users can take the model's prediction of important conformers based on performance heuristics and calculate the full workflow for this subset of conformers. This strategy still eliminates a large fraction of the compute time and ensures that the output value of ΔG_{soln} is within a few kcal/mol of the true value. Additionally, if the conformer searches for the solutes are recalculated, the model can give reasonable initial relative free energy predictions on the new conformers, again enabling fast prioritization of which conformers to compute in more detail.

CONCLUSIONS

In this work, we use quantum chemical methods to generate a computational data set consisting of Gibbs free energy change of solution for approximately 44,000 solute molecules across 41 different solvents. This includes enumerating conformers for each solute, applying DFT methods to derive geometries and gas-phase electronic energies, using xTB for vibrations, and computing solvation corrections using COSMO-RS. We then train a deep learning model that uses 3D message passing neural networks to predict the relative free energies of solutions for different solute conformers across the set of solvents.

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

We analyze the data set sequentially, first by comparing the computed solvation free energies from COSMO-RS to experimental octanol−water partition coefficients from the literature. This analysis shows that computing a Boltzmann-corrected $\log P_{OW}$ across conformers improves the agreement with experimental values over single-conformer approaches, especially for larger and more flexible solutes.

We next analyze the contribution of conformers to the value of ΔG_{soln} by comparing the Boltzmann-ensembled value to that of the LEC. Using only gas-phase energies to locate low-energy conformers and using those to compute ΔG_{soln} give a value reasonably close to the Boltzmann-ensembled value on average, but it leads to many solute−solvent pairs with errors above 5 kcal/mol. This is especially true for certain solvents, namely, the more acidic sulfuric acid and formic acid, as well as the more basic ammonia and di-2-butylamine. This suggests that it is important to perform the conformational search with solvation corrections applied for strong solvents and that this can be more crucial for certain solute−solvent pairs.

When using solvation corrections to identify the LEC and comparing the solute free energy of this conformer to the Boltzmann-ensembled value, we observe much smaller deviations, as all differences are below 2.5 kcal/mol, suggesting that the conformer effects were not substantial. A closer analysis for just water shows that the difference in solvation free energies increased for solutes with more rotatable bonds, with a weaker dependence on the number of hydrogen bond donors in the solutes, but the differences are still largely below 1 kcal/mol except for very large solutes with seven to eight rotatable bonds. Thus, identifying the LEC is likely sufficient when estimating the free energy change of solution.

Finally, our trained model is generally able to identify the LEC in solution for solutes with few conformers, but as the number of rotatable bonds and thus the number of conformers present in the solute increase, so does the deviation between the LEC ranking according to the model and the true ranking. However, in most cases, the conformer identified by the model is still a low-energy conformer as the ΔG_{soln} for the model-identified LEC was generally within 1 kcal/mol of the actual LEC. Using the relative differences in ΔG_{soln} for the various solute conformers predicted by the model to compute a predicted Boltzmann-ensembled $\overline{\Delta G}^{M}_{soln}$ generally resulted in values that were close to and correlated well with the true value, confirming that the trained model is capturing the relative differences between solute conformers well across all solvents. This means the model can help save computational time when computing ΔG_{soln} for solute−solvent pairs, as only one conformer needs to be computed fully, and the relative values for all of the other conformers can be derived from the model instead.

Going forward, as the model is currently trained on conformer geometries optimized by using DFT, there is still the requirement of running a geometry optimization using DFT for all of the conformers for each solute. It would be useful to retrain the model on geometries derived from more computationally affordable methods than DFT optimization, such as with GFN2-xTB geometries, as these are expected to provide reliable geometries while requiring a much lower computational cost. Similarly, retraining on 3D conformer geometries predicted by efficient ML models such as Geomol[25] could also improve the model applicability. However, both of these appealing possibilities have the difficulty that during training, the approximate 3D geometry for each conformer predicted by more affordable methods needs to be connected to a high-accuracy conformer energy (e.g., computed using DFT with COSMO), which are typically connected to a corresponding higher accuracy 3D geometry. In some cases, it may be difficult to make a 1-to-1 connection between these geometries. However, so long as the less expensive geometries are sufficiently close to those of DFT, retraining would be substantially helpful and the model could then save a substantial additional amount of computational time.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.3c05904.

Additional data analysis; derivation of Boltzmann-ensembled solution free energies for multiple solute conformers; methodology for computing logP values while including conformational effects; and additional machine learning model performance analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Author

William H. Green − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-2603-9694; Email: whgreen@mit.edu

Authors

Lagnajit Pattanaik − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-7001-6698

Angiras Menon − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Volker Settels − *BASF SE, Scientific Modeling, Group Research, Ludwigshafen am Rhein 67056, Germany*

Kevin A. Spiekermann − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-9484-9253

Zipei Tan − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Florence H. Vermeire − *Department of Chemical Engineering, KU Leuven, Leuven 3001, Belgium; Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Frederik Sandfort − *BASF SE, Scientific Modeling, Group Research, Ludwigshafen am Rhein 67056, Germany*

Philipp Eiden − *BASF SE, Scientific Modeling, Group Research, Ludwigshafen am Rhein 67056, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.3c05904

Author Contributions

‖L.P. and A.M. authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

■ ACKNOWLEDGMENTS

The authors thank the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium for their support. The authors would like to thank Kariana Moreno Sader for her help with the TOC graphic as well as Dr. Sebastian Spicher and Jonathan W. Zheng for useful discussions. The authors also acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper/report.

■ REFERENCES

(1) Deglmann, P.; Schäfer, A.; Lennartz, C. Application of Quantum Calculations in the Chemical Industry—An Overview. *Int. J. Quantum Chem.* **2015**, *115*, 107−136.
(2) Peters, M. A computational approach to solvent selection for biphasic reaction systems. Ph.D. Thesis, Aachen, Techn. Hochsch., Diss., 2008, 2008.
(3) Scheffczyk, J.; Schäfer, P.; Fleitmann, L.; Thien, J.; Redepenning, C.; Leonhard, K.; Marquardt, W.; Bardow, A. COSMO-CAMPD: a framework for integrated design of molecules and processes based on COSMO-RS. *Mol. Syst. Des. Eng.* **2018**, *3*, 645−657.
(4) Chung, Y.; Green, W. H. Computing Kinetic Solvent Effects and Liquid Phase Rate Constants Using Quantum Chemistry and COSMO-RS Methods. *J. Phys. Chem. A* **2023**, *127*, 5637−5651.
(5) Havenith, M. Solvation Science: A New Interdisciplinary Field. *Angew. Chem., Int. Ed.* **2016**, *55*, 1218.
(6) Choi, H.; Kang, H.; Park, H. New Solvation Free Energy Function Comprising Intermolecular Solvation and Intramolecular Self-Solvation Terms. *J. Cheminf.* **2013**, *5*, 8.
(7) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711−720.
(8) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999−3094.
(9) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute with a Continuum. A Direct Utilizaion of AB Initio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55*, 117−129.
(10) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760−768.
(11) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.
(12) Klamt, A. The COSMO and COSMO-RS Solvation Models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 699−709.
(13) Hellweg, A.; Eckert, F. Brick by Brick Computation of the Gibbs Free Energy of Reaction in Solution using Quantum Chemistry and COSMO-RS. *AIChE J.* **2017**, *63*, 3944−3954.
(14) Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Chem. J. Soc. Perkin Trans.* **1993**, *2*, 799−805.
(15) Reinisch, J.; Klamt, A. Prediction of Free Energies of Hydration with COSMO-RS on the SAMPL4 Data Set. *J. Comput. Aided Mol. Des.* **2014**, *28*, 169−173.
(16) Loschen, C.; Reinisch, J.; Klamt, A. COSMO-RS Based Predictions for the SAMPL6 logP Challenge. *J. Comput. Aided Mol. Des.* **2020**, *34*, 385−392.
(17) Hawkins, P. C. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.
(18) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719−2728.
(19) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.
(20) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab-Systematic generation of diverse low-energy conformers. *J. Cheminf.* **2011**, *3*, 8.
(21) Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(22) Chan, L.; Hutchison, G. R.; Morris, G. M. Bayesian Optimization for Conformer Generation. *J. Cheminf.* **2019**, *11*, 32.
(23) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847−2862.
(24) Simm, G. N.; Hernández-Lobato, J. M. A Generative Model for Molecular Distance Geometry. **2019**, arXiv preprint arXiv:1909.11459.
(25) Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; Jaakkola, T. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 2021; Vol. *34*.
(26) Corso, G.; Stark, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *International Conference on Learning Representations*, 2023; p 11.
(27) Buggert, M.; Cadena, C.; Mokrushina, L.; Smirnova, I.; Maginn, E. J.; Arlt, W. COSMO-RS Calculations of Partition Coefficients: Different Tools for Conformation Search. *Chem. Eng. Technol.* **2009**, *32*, 977−986.
(28) Hyttinen, N.; Prisle, N. L. Improving Solubility and Activity Estimates of Multifunctional Atmospheric Organics by Selecting Conformers in COSMO therm. *J. Phys. Chem. A* **2020**, *124*, 4801−4812.
(29) Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of the Free Energy of Hydration of a Challenging Set of Pesticide-Like Compounds. *J. Phys. Chem. B* **2009**, *113*, 4508−4510.
(30) Lim, V. T.; Hahn, D. F.; Tresadern, G.; Bayly, C. I.; Mobley, D. L. Benchmark assessment of molecular geometries and energies from small molecule force fields. *F1000Research* **2020**, *9*, 1390.
(31) Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(32) Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic pKa Values in the context of the SAMPL6 Challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1139−1149.
(33) Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P.; Seibert, J.; Neese, F. Fully Automated Quantum-Chemistry-Based Computation of Spin−Spin-Coupled Nuclear Magnetic Resonance Spectra. *Angew. Chem., Int. Ed.* **2017**, *56*, 14763−14769.
(34) Udvarhelyi, A.; Rodde, S.; Wilcken, R. ReSCoSS: a flexible quantum chemistry workflow identifying relevant solution conformers of drug-like molecules. *J. Comput. Aided Mol. Des.* **2021**, *35*, 399−415.
(35) Skyner, R.; McDonagh, J.; Groom, C.; Van Mourik, T.; Mitchell, J. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174−6191.
(36) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data to Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726−741.
(37) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1338−1346.

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170

(38) Lim, H.; Jung, Y. Delfos Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306−8315.
(39) 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. *J. Chem. Phys.* **2021**, *154*, 134113.
(40) Ansari, R.; Ghorbani, A. *Accurate Prediction of Free Solvation Energy of Organic Molecules via Graph Attention Network and Message Passing Neural Network from Pairwise Atomistic Interactions*, 2021;.
(41) Vermeire, F. H.; Green, W. H. Transfer Learning for Solvation Free Energies: From Quantum Chemistry to Experiments. *Chem. Eng. J.* **2021**, *418*, 129307.
(42) 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. *J. Chem. Inf. Model.* **2022**, *62*, 433−446.
(43) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022**, *144*, 10785−10797.
(44) Pyykkö, P.; Atsumi, M. Molecular Single-Bond Covalent Radii for Elements 1−118. *Eur. J. Chem.* **2009**, *15*, 186−197.
(45) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta−generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.
(46) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571−2577.
(47) Kanai, Y.; Wang, X.; Selloni, A.; Car, R. Testing the TPSS Meta-Generalized-Gradient-Approximation Exchange-Correlation Functional in Calculations of Transition States and Reaction Barriers. *J. Chem. Phys.* **2006**, *125*, 234104.
(48) Bühl, M.; Kabrede, H. Geometries of transition-metal complexes from density-functional theory. *J. Chem. Theory Comput.* **2006**, *2*, 1282−1290.
(49) Bauernschmitt, R.; Häser, M.; Treutler, O.; Ahlrichs, R. Calculation of excitation energies within time-dependent density functional theory using auxiliary basis set expansions. *Chem. Phys. Lett.* **1997**, *264*, 573−578.
(50) Ahlrichs, R. Efficient evaluation of three-center two-electron integrals over Gaussian functions. *Phys. Chem. Chem. Phys.* **2004**, *6*, 5119−5121.
(51) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294*, 143−152.
(52) Grimme, S. Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 211−228.
(53) Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. Turbomole. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 91−100.
(54) Balasubramani, S. G.; Chen, G. P.; Coriani, S.; Diedenhofen, M.; Frank, M. S.; Franzke, Y. J.; Furche, F.; Grotjahn, R.; Harding, M. E.; Hättig, C.; et al. TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations. *J. Chem. Phys.* **2020**, *152*, 184107.
(55) *TURBOMOLE*, V7.5 2020; Development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007, TURBOMOLE GmbH, since, 2007. https://www.turbomole.org.
(56) Spicher, S.; Grimme, S. Single-Point Hessian Calculations for Improved Vibrational Frequencies and Rigid-Rotor-Harmonic-Oscillator Thermodynamics. *J. Chem. Theory Comput.* **2021**, *17*, 1701−1714.
(57) 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. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(58) Pracht, P.; Grant, D. F.; Grimme, S. Comprehensive assessment of GFN tight-binding and composite density functional theory methods for calculating gas-phase infrared spectra. *J. Chem. Theory Comput.* **2020**, *16*, 7044−7060.
(59) Spicher, S.; Grimme, S. Efficient computation of free energy contributions for association reactions of large molecules. *J. Phys. Chem. Lett.* **2020**, *11*, 6606−6611.
(60) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1493.
(61) BIOVIA. *BIOVIA COSMOtherm 2020 Reference Manual*. https://www.3ds.com/fileadmin/Support/rl/BIOVIA-COSMOlogic/BIOVIA_COSMOtherm_2020_Reference_Manual.pdf, 2020; [Online; accessed October 24, 2023].
(62) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The Atomic Simulation Environment-A Python Library for Working with Atoms. *J. Condens. Matter Phys.* **2017**, *29*, 273002.
(63) pandas development team, T.. *pandas-dev/pandas: Pandas*, 2020;.
(64) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(65) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proc. Int. Conf. Mach. Learn.* 2017; pp 1263−1272.
(66) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Adv. Neural Inf. Process Syst.*, 2017; Vol. *30*.
(67) Liu, Y.; Wang, L.; Liu, M.; Zhang, X.; Oztekin, B.; Ji, S. Spherical message passing for 3d graph networks. **2021**, arXiv preprint arXiv:2102.05013.
(68) Francoeur, P. G.; Koes, D. R. SolTranNet−A machine learning tool for fast aqueous solubility prediction. *J. Chem. Inf. Model.* **2021**, *61*, 2530−2536.
(69) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask prediction of site selectivity in aromatic C−H functionalization reactions. *React. Chem. Eng.* **2020**, *5*, 896−902.
(70) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regioselectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **2021**, *12*, 2198−2208.
(71) Stuyver, T.; Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: performance, generalizability, and explainability. *J. Chem. Phys.* **2022**, *156*, 156.
(72) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chem.—Eur. J.* **2023**, *29*, No. e202300387.
(73) Biswas, S.; Chung, Y.; Ramirez, J.; Wu, H.; Green, W. H. Predicting critical properties and acentric factors of fluids using multitask machine learning. *J. Chem. Inf. Model.* **2023**, *63*, 4574−4588.
(74) Jiménez-Luna, J.; Pérez-Benito, L.; Martinez-Rosell, G.; Sciabola, S.; Torella, R.; Tresadern, G.; De Fabritiis, G. DeltaDelta neural networks for lead optimization of small molecule potency. *Chem. Sci.* **2019**, *10*, 10911−10918.
(75) McNutt, A. T.; Koes, D. R. Improving ΔΔg predictions with a multitask convolutional Siamese network. *J. Chem. Inf. Model.* **2022**, *62*, 1819−1829.
(76) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *J. Phys. Chem. Lett.* **2020**, *11*, 2992−2997.

(77) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976−3986.
(78) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1 Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(79) Landrum, G.; et al. *RDKit: Open-Source Cheminformatics Software*, 2006. https://www.rdkit.org.
(80) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2022**, *62*, 2101−2110.
(81) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide Toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954−4965.
(82) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505−508.
(83) Alibakhshi, A.; Hartke, B. Improved Prediction of Solvation Free Energies by Machine-Learning Polarizable Continuum Solvation Model. *Nat. Commun.* **2021**, *12*, 3584.
(84) Lim, H.; Jung, Y.; MLSolv, -A.: A Novel Machine Learning-Based Prediction of Solvation Free Energies from Pairwise Atomistic Interactions. **2020**, arXiv preprint arXiv:2005.06182.
(85) 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.
(86) Coimbra, J. T.; Feghali, R.; Ribeiro, R. P.; Ramos, M. J.; Fernandes, P. A. The Importance of Intramolecular Hydrogen Bonds on the Translocation of the Small Drug Piracetam through a Lipid Bilayer. *RSC Adv.* **2021**, *11*, 899−908.
(87) Rumble, J. R. *CRC Handbook of Chemistry and Physics, (Internet Version 2022)*; CRC Press, 2022.
(88) Mansouri, K.; Grulke, C. M.; Richard, A. M.; Judson, R. S.; Williams, A. J. An Automated Curation Procedure for Addressing Chemical Errors and Inconsistencies in Public Datasets used in QSAR Modelling. *SAR and QSAR in Environ. Res.* **2016**, *27*, 911−937.
(89) Ulrich, N.; Goss, K.-U.; Ebert, A. Exploring the Octanol−Water Partition Coefficient Dataset using Deep Learning Techniques and Data Augmentation. *Commun. Chem.* **2021**, *4*, 90.
(90) Kundi, V.; Ho, J. Predicting Octanol−Water Partition Coefficients: Are Quantum Mechanical Implicit Solvent Models Better than Empirical Fragment-Based Methods? *J. Phys. Chem. B* **2019**, *123*, 6810−6822.
(91) Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K. Breaking a bottleneck: Accurate extrapolation to "gold standard" CCSD (T) energies for large open shell organic radicals at reduced computational cost. *J. Comput. Chem.* **2016**, *37*, 286−295.
(92) Gorges, J.; Grimme, S.; Hansen, A.; Pracht, P. Towards Understanding Solvation Effects on the Conformational Entropy of Non-Rigid Molecules. *Phys. Chem. Chem. Phys.* **2022**, *24*, 12249−12259.

https://doi.org/10.1021/acs.jpcb.3c05904
J. Phys. Chem. B 2023, 127, 10151−10170