THE JOURNAL OF
PHYSICAL CHEMISTRY A

pubs.acs.org/JPCA Article

Quantum-Chemically Informed Machine Learning: Prediction of Energies of Organic Molecules with 10 to 14 Non-hydrogen Atoms

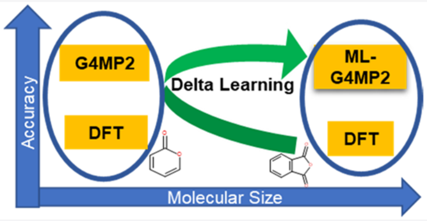
Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Naveen Dandu, Logan Ward, Rajeev S. Assary, Paul C. Redfern, Badri Narayanan, Ian T. Foster, and Larry A. Curtiss*

Cite This: *J. Phys. Chem. A* 2020, 124, 5804−5811

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

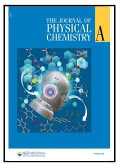


ABSTRACT: High-fidelity quantum-chemical calculations can provide accurate predictions of molecular energies, but their high computational costs limit their utility, especially for larger molecules. We have shown in previous work that machine learning models trained on high-level quantum-chemical calculations (G4MP2) for organic molecules with one to nine non-hydrogen atoms can provide accurate predictions for other molecules of comparable size at much lower costs. Here we demonstrate that such models can also be used to effectively predict energies of molecules larger than those in the training set. To implement this strategy, we first established a set of 191 molecules with 10−14 non-hydrogen atoms having reliable experimental enthalpies of formation. We then assessed the accuracy of computed G4MP2 enthalpies of formation for these 191 molecules. The error in the G4MP2 results was somewhat larger than that for smaller molecules, and the reason for this increase is discussed. Two density functional methods, B3LYP and ωB97X-D, were also used on this set of molecules, with ωB97X-D found to perform better than B3LYP at predicting energies. The G4MP2 energies for the 191 molecules were then predicted using these two functionals with two machine learning methods, the FCHL-Δ and SchNet-Δ models, with the learning done on calculated energies of the one to nine non-hydrogen atom molecules. The better-performing model, FCHL-Δ, gave atomization energies of the 191 organic molecules with 10−14 non-hydrogen atoms within 0.4 kcal/mol of their G4MP2 energies. Thus, this work demonstrates that quantum-chemically informed machine learning can be used to successfully predict the energies of large organic molecules whose size is beyond that in the training set.

INTRODUCTION

Predicting accurate energies is essential for determining reaction energetics and enthalpies and stabilities of molecules in general. Knowledge of precise thermochemical data, particularly of organic molecules, is desired in areas such as battery electrolytes, catalysis, drugs, etc. Specifically, for batteries there is an urgent need to develop organic electrolytes with improved electrochemical performance and longer-term stability.[1] However, screening of the electrolytes used in batteries is challenging because there is a large pool of organic molecules[2] that are of potential interest. Synthesizing and testing each individual molecule would be extremely challenging even with combinatorial high-throughput screening[3] and is not practical. Such screening can be made more feasible by quantum-chemical computations if the computational methods are sufficiently accurate, especially in predicting chemical and electrochemical stabilities. However, one of the challenging tasks is to choose a good computational method that can be applied universally to any set of molecules in order to screen good electrolytes. There are quite a few quantum-chemical methods that can predict molecular energies of better than 1 kcal/mol accuracy, such as the G*n* composite methods,[4−7] the correlation-consistent composite approach,[8−13] coupled cluster methods,[14] Weizmann methods,[15,16] and the Wuhan−Minnesota scaling method.[17] Nevertheless, at present such methods are limited for computing energies of large molecules, as they require substantial amounts of computational power as the number of atoms increases. Composite methods such as G4MP2, one of the G*n* composite methods, have demonstrated better than 1 kcal/mol accuracy on the G3/05 test set[18] and on a dataset of 459 organic molecules with one to nine non-hydrogen atoms.[19] However, applying this method to many large molecules, such as those with more than nine non-hydrogen atoms, is too computationally expensive to be practical.

Received: February 28, 2020
Revised: June 11, 2020
Published: June 15, 2020



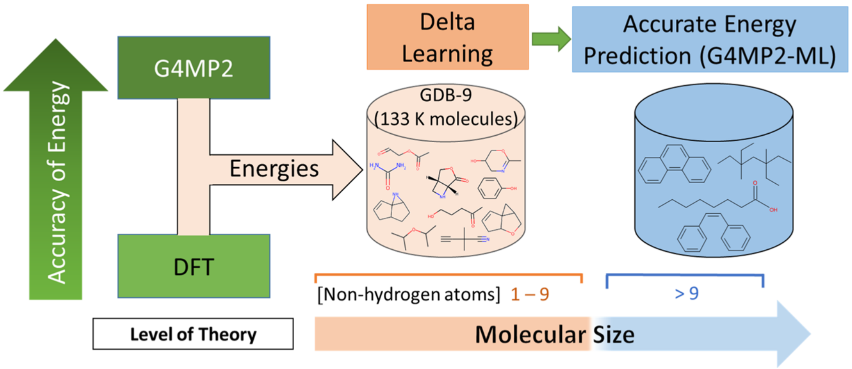


Figure 1. Schematic of the use of Δ learning on G4MP4 and DFT energies on the GDB-9 set of molecules with one to nine non-hydrogen atoms to predict energies of larger molecules with G4MP2 accuracy using only DFT energies.

Machine learning (ML), on the other hand, provides an opportunity to predict molecular energies accurately and fast; in turn, these energies can be used to screen millions of organic molecules of interest on the basis of descriptors derived from energies.[20−23] Several ML models that provide a good trade-off between accuracy and computational cost have been published.[24−27] Other studies have demonstrated success in predicting the difference between low- and high-fidelity calculations by integrating the information from quantum-chemistry methods (also known as Δ learning)[28] and training a model on multiple properties (also known as transfer learning).[29] There have also been some previous reports on predicting the energies of large organic molecules using machine learning.[30−32] However, the efficacy of these methods for predicting energies of large organic molecules has not been assessed systematically.

Here we report an assessment of quantum-chemically informed machine learning for the prediction of energies of organic molecules *larger than those in the training set* used for learning, as illustrated in the schematic in Figure 1. The objective of this work is to find an ML method that can reproduce the accuracy of G4MP2 calculations for large organic molecules with more than nine non-hydrogen atoms, but with much less computational expense. Previously, we used quantum-chemical energies of a set of 130 258 molecules with one to nine non-hydrogen atoms to train and assess the performance of several ML methods. The assessments of the ML methods were done on a holdout set of the dataset, i.e., a set of molecules not used in the training of the method. The quantum-chemical methods used for the ML were the G4MP2 composite method[5] and the B3LYP density functional method.[33−35] In this work, we used the two best-performing ML methods found in that work.[36] In addition, in this study we assessed the suitability of the ωB97X-D density functional method for learning. The first part of this work involved the development of a dataset of accurate enthalpies of formation for molecules with 10 to 14 non-hydrogen atoms. This dataset was derived from accurate experimental data in the Pedley compilation[37] along with the use of an isodesmic computational scheme to check for inaccuracies. Deficiencies found in the quantum-chemical methods on certain types of the larger molecules are also discussed. We note that in our previous study[36] the ML methods were tested on the energies of 66 sugar molecules with more than nine non-hydrogen molecules, but these larger molecules did not have accurate experimental energies. Having set of larger molecules with accurate experimental energies is important for assessing the ML methods for larger molecules. The second part of this work involved an evaluation of the performance of previously reported ML models for prediction of molecular energies. Two ML methods were used: kernel-based ridge regression (FCHL)[38] and a continuous filter convolutional neural network (SchNet),[39] both combined with Δ learning on the difference between the DFT and G4MP2 energies. These methods were previously reported to predict accurate atomization energies of the GDB-9 dataset, which was also used for training.[36] The overall objective of this work is to assess these Δ-learning ML models for accurate prediction of the energies of a set of molecules with a larger number of heteroatoms than in the training set.

COMPUTATIONAL DETAILS

Quantum-Chemical Methods. Enthalpies of formation for the dataset of large organic molecules described below were calculated with the G4MP2 method,[5] as implemented within the Gaussian 16 package[40] using the standard settings. The G4MP2 method is a composite one that replaces the fourth-order perturbation methods in G4 theory[6] with reduced perturbation theory levels to lower the computational cost. Other parts of the G4MP2 method remain the same as in G4 theory. It has been assessed on the G3/05 test set of accurate experimental data,[18] on which it has a mean absolute deviation of 0.77 kcal/mol for the 138 hydrocarbons and substituted hydrocarbons in the test set.

This investigation is based on a dataset of large molecules with 10 to 14 non-hydrogen atoms (C, N, O, and F). They were chosen from the molecules with 10 to 14 non-hydrogen atoms in the Pedley compilation of experimental gas-phase enthalpy of formation data for organic molecules.[22] Two criteria were used in choosing the molecules. First, they had to have a stated experimental uncertainty of ±1 kcal/mol in the compilation, which was the case for 225 molecules. The G4MP2 enthalpies of formation at 298 K ($\Delta_f H°(298\ K)$) were calculated for this set of molecules in a similar manner as in previous work,[41,42] where enthalpies of formation at 0 K ($\Delta_f H°(0\ K)$) were calculated by taking the difference between the known enthalpies of formation of the corresponding atoms and the calculated atomization energies. The $\Delta_f H°(298\ K)$ values were then obtained by adding heat capacity corrections to the $\Delta_f H°(0\ K)$ values.[41] Second, any experimental value differing by more than 2 kcal/mol from the G4MP2 value was examined using an isodesmic scheme[43] that is generally accurate to 0.5 kcal/mol, and if the difference was also more than 2 kcal/mol with the isodesmic scheme, we dropped that experimental value as likely to be inaccurate for this investigation. Since the geometries were generated from PubChem,[44] we examined other possible conformers to make sure that the source of error was not the wrong conformer. In some cases we found that the geometries generated from PubChem had incorrect conformers, but this was not the cause of the errors greater than 2 kcal/mol. On the basis of the second criterion, 34 molecules were dropped, resulting in a dataset of 191 molecules, which is called the PDS(10-14) dataset throughout this paper. The resulting dataset had 71 hydrocarbons and 120 substituted hydrocarbons. The 34 molecules that were dropped because it was likely that they had large experimental errors are listed in Table S1. An Excel file containing the geometries, energies, and zero-point energies is provided in the Supporting Information.

Furthermore, we also used two density functional methods, B3LYP[33−35] and ωB97X-D,[45,46] for both assessment of the performance on the larger molecules in the PDS(10-14) set and use in the ML. The B3LYP functional was chosen because it is a widely used functional, while ωB97X-D was chosen because it is one of the most accurate for energies.[46] The B3LYP method used the 6-31G(2df,p) basis set for both geometry optimization and energy evaluation. The ωB97X-D energy calculations were done with the 6-311+G(3df,2p) basis set at the B3LYP/6-31G(2df,p)-optimized geometries. For both methods unscaled B3LYP/6-31G(2df,p) zero-point energies were used. The larger basis set was used in computing ωB97X-D energies because it gave significantly better performance for enthalpies of formation in a study of the GDB-9 dataset in our earlier published results, whereas the smaller basis set did better for B3LYP.[19] We computed both standard enthalpies of formation at 298 K and atomization energies at the G4MP2, B3LYP, and ωB97X-D levels of theory. The generated dataset of atomization energies of the PDS(10-14) molecules was used in assessing ML to predict energies of larger molecules.

Machine Learning Methods. We used two ML approaches from our previous study,[36] SchNet-Δ and FCHL-Δ, in this investigation to determine how well ML can be used for accurate prediction of a set of molecules with a larger number of heteroatoms than in the training set. Both use Δ learning on the difference between the DFT and G4MP2 energies. Establishing the accuracies of the DFT and G4MP2 methods for the larger molecules in the Results and Discussion is critical for the assessment of the ML methods.

The SchNet ML method is a continuous-filter convolutional neural network method[39] designed to predict molecular energies and atomic forces. The SchNet architecture consists of multiple continuous-filter convolutional layers. At each layer, molecules are represented atomwise, similar to the concept of pixels in an image. Each atomic number is mapped to a vector to generate the initial representation of each atom. Interactions between atoms are updated on the basis of distances between nearby neighbors. The atomwise updates of these representations produce the contribution of each atom. A molecular property is then generated by the summation of all atomic contributions. We used an open-source code available in SchNetPack.[47] SchNet-Δ is a variant of SchNet that learns the difference between properties computed with high- and low-fidelity methods. In this work, we trained the SchNet-Δ model using the difference between the G4MP2 and ωB97X-D (or B3LYP) atomization energies.

The FCHL ML method learns molecular properties using a kernel-based ridge regression model.[38] Each atom is represented as a weighted sum of Gaussian distributions using a many-body approach. The approach measures the similarities between the local environments of two atoms as well as molecules. The distance between two atomic environments is measured using a weighted sum of the distances between each many-body expansion. FCHL-Δ is a variant of FCHL that learns the difference between properties computed with high- and low-fidelity quantum-chemical methods. In this work, we trained the FCHL-Δ model using the difference between the G4MP2 and ωB97X-D (or B3LYP) atomization energies.

The FCHL- and SchNet-trained machine learning models used in this work are available on DLHub,[48] and the scripts used to perform parameter optimization and their outputs have been posted on GitHub.[49]

RESULTS AND DISCUSSION

Accuracies of Quantum-Chemical Methods for the PDS(10-14) Dataset. Table 1 gives the mean absolute deviations (MADs) of the standard enthalpies of formation at 298 K for the 191 PDS(10-14) molecules computed at the G4MP2, B3LYP, and ωB97X-D levels of theory from their experimental values. The standard enthalpies of formation for the 191 molecules at these three levels of theory along with their experimental values are given in Tables S2 and S3. The overall MADs of the G4MP2, B3LYP, and ωB97X-D methods relative to the experimental values are 1.41, 4.81, and 3.13 kcal/mol, respectively. Among these 191 molecules, there are 71 hydrocarbons and 120 substituted hydrocarbons. The MADs of the 71 hydrocarbons at the G4MP2, B3LYP, and ωB97X-D levels are 1.75, 4.04, and 3.47 kcal/mol, respectively, whereas the MADs of the substituted hydrocarbons at the G4MP2, B3LYP, and ω-B97X-D levels are 1.21, 5.26, and 2.92 kcal/mol, respectively. These results indicate that the ω-B97XD functional performs better than B3LYP. Nevertheless, the better-performing functional, ω-B97X-D, has an overall MAD that is about twice as large as that of G4MP2.

Table 1. Mean Absolute Deviations (MADs) from Experiment of the Enthalpies of Formation at 298 K for the PDS(10-14) Dataset for the G4MP2 and DFT Methods (in kcal/mol)

	MAD of $\Delta_f H°$(298 K) (kcal/mol)		
molecule type	G4MP2	B3LYP/ 6-31G(2df,p)[a]	ωB97X-D/ 6-311+G(3df,2p)[a]
hydrocarbons (71)	1.75	4.04	3.47
non-hydrocarbons (120)	1.21	5.26	2.92
total (191)	1.41	4.81	3.13

[a] Calculations at the ω-B97XD/6-311+G(3df,2p) level were done using the B3LYP/6-31G(2df,p) geometries. The B3LYP energies were calculated using a smaller basis set because the performance with a larger basis set was worse (see the text).

The MADs for the three methods as functions of the number of non-hydrogen atoms are plotted in Figure 2; the values are given in Table 2. Figure 2 also includes previously reported data[19] for 459 molecules with two to nine non-hydrogen atoms. The B3LYP method tends to show larger MADs compared with the other methods, with ωB97X-D in between B3LYP and G4MP2, except in the case of 14 non-

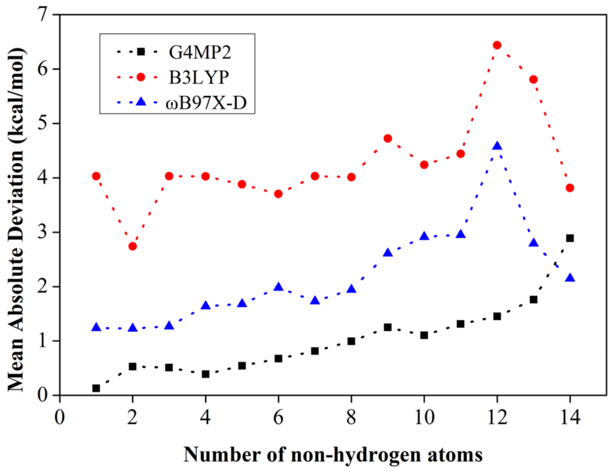


Figure 2. Mean absolute deviations (MADs) from experiment for enthalpies of formation at 298 K computed using G4MP2, B3LYP, and *ω*B97X-D for the PDS(10-14) molecules as functions of the number of non-hydrogen atoms. Results for the 459 molecule test set (one to nine non-hydrogen atoms) from ref 19 are included for comparison.

Table 2. Mean Absolute Deviations (MADs) from Experiment of Enthalpies of Formation at 298 K for the PDS(10-14) Dataset for the G4MP2 and DFT Methods (in kcal/mol) as Functions of Number of Non-hydrogen Atoms

		MAD of $\Delta_f H°$(298 K) (kcal/mol)		
no. of non-hydrogen atoms	no. of molecules	G4MP2	B3LYP	*ω*B97X-D
10	80	1.10	4.24	2.91
11	43	1.31	4.44	2.95
12	31	1.45	6.44	4.58
13	24	1.76	5.81	2.80
14	13	2.89	3.81	2.15

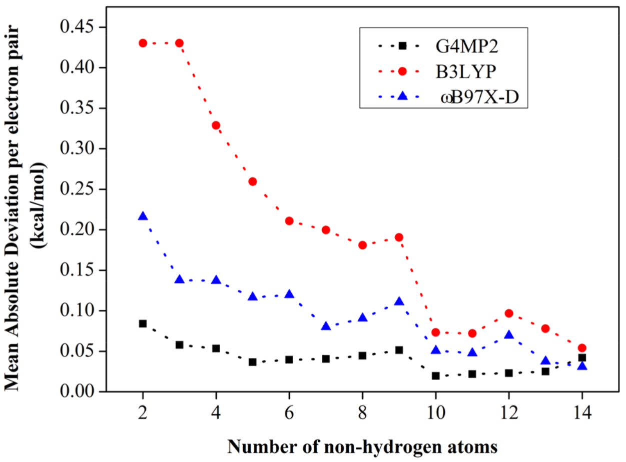


Figure 3. Mean absolute deviations (MADs) from experiment per electron pair for enthalpies of formation at 298 K computed using G4MP2, B3LYP, and *ω*B97X-D for the PDS(10-14) molecules as functions of the number of non-hydrogen atoms. Results for the 459 molecule test set (two to nine non-hydrogen atoms) from ref 19 are included for comparison.

hydrogen atoms, where *ω*B97X-D actually performs better than G4MP2. This is due to the larger percentage of aromatic molecules in the set with 14 non-hydrogen atoms, i.e., 11 of 13 (85%) as opposed to 88 of 191 (46%) for the whole PDS(10-14) set. The failure of G4MP2 for aromatics has not been noted before as far as we are aware, probably because of the smaller sizes and numbers in previous studies. We carried out an evaluation of 54 aromatics in the PDS(10-14) set with G4 theory, which has fewer approximations than G4MP2 theory, and found a MAD of 1.15 kcal/mol compared with 1.84 kcal/mol for G4MP2 for this same set. For the six aromatics with 14 non-hydrogen atoms in this set of 54 aromatics, the MAD for G4MP2 is 3.63 kcal/mol, compared with 1.65 kcal/mol for G4 theory. Thus, much of the increase in error found for G4MP2 is due to the many aromatics in the set of larger molecules PDS(10-14) compared with previous studies on smaller molecules. Table S4 contains a summary of the G4 energies for the 54 selected aromatic molecules in the PDS(10-14) test set. The failure of G4MP2 for aromatics will be considered in more detail in a future study.

As can be seen from Figure 2, the MAD for G4MP2 generally increases with the number of non-hydrogen atoms. This result can be attributed to the errors due to the increase in the number of electron pairs in the molecule with size, as shown in Figure 3, where the MADs per electron pair are plotted as functions of the number of non-hydrogen atoms. Despite the increase in error with increasing size of the molecule, we observe that the G4MP2 errors per electron pair for PDS(10-14) remain approximately the same with increased molecule size, with the exception of the set with 14 non-hydrogen atoms, due to the many aromatics in that set. By comparison, the B3LYP and *ω*B97X-D density functionals show an irregular trend with increasing molecule size.

To further analyze the error in the various methods, we categorized the molecules with respect to elemental constituents, as shown in Table 3. The largest MADs at the G4MP2 level are for molecules that contain "HCN". Among those, specifically, the molecules that contain N−N bonds, [e.g., *cis*-azobenzene (3.69 kcal/mol), *trans*-azobenzene (2.13 kcal/mol), and benzo[*c*]cinnoline (3.31 kcal/mol)] and those that contain multiple nitrile bonds [e.g., 1,2-benzenedicarbonitrile, 1,3-benzenedicarbonitrile, and 1,4-benzenedicarbonitrile] have the largest deviations. Similar large errors for these types of molecules for G4 theory were reported previously by Dorofeeva et al.[50] The second-largest MADs at the G4MP2 level were for the hydrocarbons, especially for aromatics as noted previously [e.g., naphthalene (3.17 kcal/mol), acenaph-

Table 3. Mean Absolute Deviations (MADs) from Experiment of Enthalpies of Formation at 298 K for the PDS(10-14) Dataset for the G4MP2 and DFT Methods (in kcal/mol) as Functions of Type of Atoms in the Molecules

		MAD of $\Delta_f H°$(298 K) (kcal/mol)		
element constituency	no. of molecules	G4MP2	B3LYP	*ω*B97X-D
CF	1	1.12	25.07	14.44
HCNF	1	1.52	25.20	8.00
HCF	4	0.62	15.99	7.60
HCOF	4	0.93	15.86	8.48
HCN	14	1.97	3.37	2.22
HCON	21	1.15	7.19	5.04
HC	71	1.75	4.02	3.47
HCO	75	1.13	3.41	1.69

thylene (3.07 kcal/mol), and anthracene (4.84 kcal/mol)] as well as some polycyclic bridged systems such as adamantane (2.04 kcal/mol). There are different trends with DFT functionals. Larger MADs occur in the case of fluoride-containing molecules (CF, HCNF, HCF, HCOF), although there are few such entries. This observation is important because fluorinated systems are dominant in energy-storage chemistry. Of the groups with a significant number of entries, molecules with constituents “HCON” have the largest MADs (7.0 kcal/mol for B3LYP and 5.06 for ωB97X-D).

Performance of Machine Learning for the PDS(10-14) Dataset. In this work, we used the two best ML approaches (SchNet-Δ and FCHL-Δ) investigated in our previous work,[36] where we assessed multiple ML approaches for predicting the atomization energies of organic molecules using the GDB-9 set of 130 258 molecules. That work applied ML on molecules with nine or fewer non-hydrogen atoms (training set of 117 232 entries) and then tested on a holdout set (13 026 entries). The two best ML models, SchNet-Δ and FCHL-Δ, have different accuracy/speed trade-offs and enabled the efficient prediction of G4MP2-level energies with an accuracy of up to 0.1 kcal/mol for the holdout set. These models were based on energies from the B3LYP functional of the holdout set and the G4MP2 and B3LYP energies of the GDB-9 training set. Furthermore, the trained FCHL-Δ and SchNet-Δ methods were assessed on 66 sugar molecules with more than nine non-hydrogen atoms but without accurate experimental values. The predictions of G4MP2-level energies for the 66 molecules had MADs of 0.29 and 0.91 kcal/mol for the two methods, respectively.

In this work, we determined how the two best models, SchNet-Δ and FCHL-Δ, performed on the 191 larger molecules in the PDS(10-14) dataset when trained on the G4MP2 and DFT energies of the set of smaller molecules, GDB-9. In the Δ-learning results presented here, we have also used the better-performing functional ωB97X-D in addition to B3LYP. We used a training set of 13 026 molecules chosen randomly from the GDB-9 set of 130 258 molecules. We used previously computed G4MP2 and B3LYP energies[19] for this training set as well as ωB97X-D energies computed in this work. These energies were then used to perform Δ learning. We randomly selected 10% (i.e., 1303 molecules) as a holdout set to test the performance of the two ML methods as well as the two DFT methods. The methods were then assessed on the PDS(10-14) molecules.

Table 4 summarizes the ML results for the holdout set of 1303 molecules from the GDB-9 dataset and the 191 molecules in the PDS(10-14) dataset. For comparison, the table also includes results from a larger holdout set of 13 026 molecules (and larger training set) done in our previous work using B3LYP energies. The B3LYP FCHL-Δ results indicate that the MAD for the smaller holdout set of 1303 molecules increases somewhat compared with the larger holdout set (0.18 vs 0.12 kcal/mol).[36] Similar results were obtained for SchNet-Δ (i.e., 0.40 vs 0.36 kcal/mol).[36] This result is expected, as a much larger number of molecules was used for training in the latter cases.

The B3LYP-based FCHL-Δ results for the PDS(10-14) set given in Table 4 for the small training set (13026) indicate that the MAD for the predicted energies relative to the actual G4MP2 energies is 0.47 kcal/mol, which is larger than that for the GDB-9 molecules in the holdout set (0.18 kcal/mol). The ωB97X-D-based FCHL-Δ method had a MAD of 0.37 kcal/mol for the PDS(10-14) set, which is slightly improved compared with the B3LYP-based FCHL-Δ results. This result is expected on the basis of the improved performance of ωB97X-D compared with B3LYP. The SchNet-Δ MADs for the holdout set of 1303 molecules were 0.40 and 0.31 kcal/mol for B3LYP and ωB97X-D, respectively, compared with 1.71 and 0.94 kcal/mol for the PDS(10-14) set. The poorer performance of SchNet-Δ relative to FCHL-Δ is similar to what we found previously for smaller molecules.[36] Finally, we note that when the B3LYP-based FCHL-Δ method from the training set (117232) was used, the MAD was only slightly smaller (0.39 kcal/mol).

Table 4. Mean Absolute Deviations (MADs) from G4MP2 Atomization Energies for two ML Models (Each Using Either B3LYP or ωB97X-D for Δ Learning); The Results Are Given for the PDS(10-14) Dataset and Two Holdout Sets from Different Sized Training Sets for Comparison

		MAD (kcal/mol)			
		B3LYP		ωB97X-D	
molecules on which MAD is based	training set size	FCHL-Δ	SchNet-Δ	FCHL-Δ	SchNet-Δ
1303 holdout molecules	13026[a]	0.18	0.40	0.12	0.31
13 026 holdout molecules[b]	117232[c]	0.12	0.36	–	–
191 molecules of the PDS(10−14) test set	13026[a]	0.47	1.76	0.37	0.98
	117232[c]	0.39	0.88	–	–

[a] The training set was 10% (randomly chosen) of the GDB-9 set of 130 258 molecules. [b] Results from ref 36 for comparison. [c] The training set was the GDB-9 set of 130 258 molecules with 10% (randomly chosen) held out.

Overall, the FCHL-Δ model based on ωB97X-D energies performed the best for the PDS(10-14) molecules, which are larger molecules than used in the training set (one to nine non-hydrogen atoms). As FCHL-Δ is based on the kernel ridge regression model, it requires comparison of each test entry to all available training entries when predicting a molecular property. Thus, its computational cost per prediction is at least 100 times greater than that of SchNet-Δ but orders of magnitude faster than that of G4MP2. Hence, we recommend the use of FCHL-Δ for its greater accuracy. Both codes are available at DLHub.[48] Despite the smaller number of molecules used for training of the larger molecules, the results in Table 4 and illustrated in Figure 4 show that the SchNet-Δ and FCHL-Δ methods have larger errors relative to G4MP2 as the size of the molecule increases, although the accuracy for FCHL-Δ of better than 0.4 kcal/mol is still good. The molecules contributing to the increase in this error correlate with the failure of G4MP2 for larger molecules such as aromatics and molecules containing N−N bonds. Thus, an improved quantum-chemical method for predicting energies could improve the ML performance.

■ CONCLUSIONS

The use of quantum-chemically informed machine learning to predict the energies of molecules larger than ones used in the training set was investigated. This work extends a previously reported study[36] in which machine learning was used to predict the energies of molecules with one to nine non-

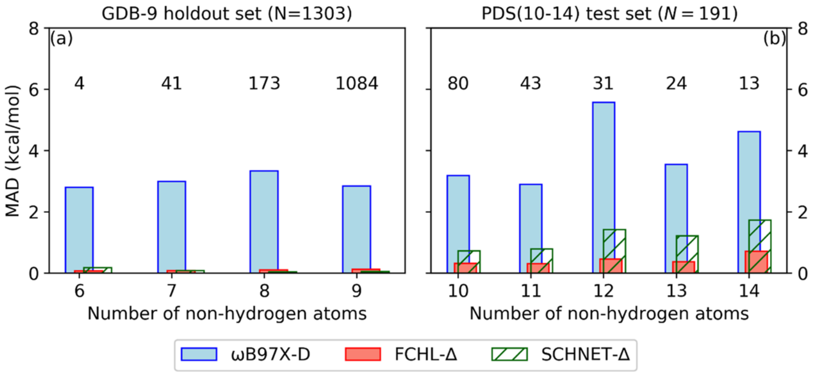


Figure 4. MADs between G4MP2 atomization energies and FCHL-Δ and SchNet-Δ predictions based on *ω*B97X-D energy calculations. The ML model was trained using only molecules with nine or fewer non-hydrogen atoms to predict the MADs between *ω*B97X-D and G4MP2 (i.e., *ω*B97X-D-Δ) for the 191 molecules with 10−14 non-hydrogen atoms.

hydrogen atoms on the basis of training on a set of molecules of that size using DFT and G4MP2 energies. In that study, the ML methods were tested on energies of 66 sugar molecules with more than nine non-hydrogen atoms, but these larger molecules did not have accurate experimental energies. The current work is based on a set of 191 molecules with 10−14 non-hydrogen atoms (i.e., larger than those in the training set) having accurate experimental enthalpies of formation. The following conclusions can be drawn from this study:

1. The better-performing ML method investigated in this paper, FCHL-Δ, gave atomization energies for the 191 organic molecules that were within about 0.4 kcal/mol of the accurate quantum-chemical energies calculated by the G4MP2 method. Although this level of accuracy is less than that obtained with FCHL-Δ for the holdout set of the smaller molecules, it is still promising for the use of ML methods for molecules larger than the training set, especially considering that it is at least 3 orders of magnitude faster per molecule than the G4MP2 method for predicting the large molecules considered here.
2. The SchNet-Δ method gave atomization energies for the 191 organic molecules that were within about 0.9 kcal/mol of the accurate quantum-chemical energies calculated by the G4MP2 method. Although this accuracy is not as good as that seen for the FCHL-Δ method, SchNet-Δ is a faster method.
3. The analysis of G4MP2 enthalpies of formation of the 191 molecules compared with accurate experimental data indicated that G4MP2 has a somewhat larger MAD (1.41 kcal/mol) for the molecules with 10−14 non-hydrogen atoms than for molecules having nine or fewer non-hydrogen atoms (0.79 kcal/mol). This is the case because G4MP2 does poorly for aromatic molecules and for molecules with multiple nitrogen atoms. Full G4 theory does much better but would take too much computer resources for generation of a large enough dataset.
4. Of the two density functional methods investigated on the 191 molecules, the *ω*B97X-D functional was found to perform better, with a MAD of 3.13 kcal/mol with respect to experiment, compared with B3LYP, with a MAD of 4.81 kcal/mol. Both methods have significantly larger errors than G4MP2. The use of *ω*B97X-D for ML Δ learning gave slightly better results for predicting the G4MP2 energies of the 191 molecules than did B3LYP.

Thus, this work demonstrates that quantum-chemically informed machine learning can be used to successfully predict energies of large organic molecules with sizes beyond those in the training set at a much lower cost in computer time.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.0c01777.

Deviations of G4MP2, B3LYP, and *ω*B97X-D calculated enthalpies of formation from experiment; list of 34 molecules of PDS(10-14) eliminated by isodesmic reactions; G4 energies for selected aromatic hydrocarbon and non-hydrocarbon molecules in the PDS(10-14) test set (PDF)

Geometries, energies, and zero-point vibrational energies of the PDS(10-14) test set (XLS)

■ AUTHOR INFORMATION

Corresponding Author

Larry A. Curtiss − *Materials Science Division and Joint Center for Energy Storage Research (JCESR), Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0001-8855-8006; Email: curtiss@anl.gov

Authors

Naveen Dandu − *Materials Science Division and Joint Center for Energy Storage Research (JCESR), Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0001-7122-8537

Logan Ward − *Joint Center for Energy Storage Research (JCESR) and Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-1323-5939

Rajeev S. Assary − *Materials Science Division and Joint Center for Energy Storage Research (JCESR), Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-9571-3307

Paul C. Redfern − *Materials Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Badri Narayanan − *Department of Mechanical Engineering, University of Louisville, Louisville, Kentucky 40292, United States;* orcid.org/0000-0001-8147-1047

Ian T. Foster − *Joint Center for Energy Storage Research (JCESR) and Data Science and Learning Division, Argonne*

National Laboratory, Lemont, Illinois 60439, United States; Department of Computer Science, University of Chicago, Chicago, Illinois 60637, United States

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpca.0c01777

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Joint Center for Energy Storage Research (JCESR), an Energy Innovation Hub funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences. We acknowledge a generous grant of computer time from the ANL Laboratory Computing Resource Center (Bebop).

■ REFERENCES

(1) Cheng, L.; Assary, R. S.; Qu, X. H.; Jain, A.; Ong, S. P.; Rajput, N. N.; Persson, K.; Curtiss, L. A. Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 283−291.

(2) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432*, 823−823.

(3) Ellman, J.; Stoddard, B.; Wells, J. Combinatorial Thinking in Chemistry and Biology. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 2779−2782.

(4) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gn Theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 810−825.

(5) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory Using Reduced Order Perturbation Theory. *J. Chem. Phys.* **2007**, *127*, 124105.

(6) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126*, 084108.

(7) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) Theory for Molecules Containing First and Second-Row Atoms. *J. Chem. Phys.* **1998**, *109*, 7764−7776.

(8) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. The Correlation Consistent Composite Approach (ccCA): An Alternative to the Gaussian-N Methods. *J. Chem. Phys.* **2006**, *124*, 114104.

(9) DeYonker, N. J.; Wilson, B. R.; Pierpont, A. W.; Cundari, T. R.; Wilson, A. K. Towards the Intrinsic Error of the Correlation Consistent Composite Approach (Ccca). *Mol. Phys.* **2009**, *107*, 1107−1121.

(10) Das, S. R.; Williams, T. G.; Drummond, M. L.; Wilson, A. K. A Qm/Qm Multilayer Composite Methodology: The Oniom Correlation Consistent Composite Approach (Oniom-Ccca). *J. Phys. Chem. A* **2010**, *114*, 9394−9397.

(11) Lynch, B. J.; Truhlar, D. G. Robust and Affordable Multicoefficient Methods for Thermochemistry and Thermochemical Kinetics: The Mccm/3 Suite and Sac/3. *J. Phys. Chem. A* **2003**, *107*, 3898−3906.

(12) Oyedepo, G. A.; Wilson, A. K. Multireference Correlation Consistent Composite Approach [Mr-Ccca]: Toward Accurate Prediction of the Energetics of Excited and Transition State Chemistry. *J. Phys. Chem. A* **2010**, *114*, 8806−8816.

(13) Omary, M. A.; Sinha, P.; Bagus, P. S.; Wilson, A. K. Electronic Structure of Mercury Oligomers and Exciplexes: Models for Long-Range/Multicenter Bonding in Phosphorescent Transition-Metal Compounds. *J. Phys. Chem. A* **2005**, *109*, 690−702.

(14) Feller, D.; Dixon, D. A. Extended Benchmark Studies of Coupled Cluster Theory through Triple Excitations. *J. Chem. Phys.* **2001**, *115*, 3484−3496.

(15) Chan, B.; Radom, L. W3x: A Cost-Effective Post-Ccsd(T) Composite Procedure. *J. Chem. Theory Comput.* **2013**, *9*, 4769−4778.

(16) Karton, A.; Martin, J. M. L. Explicitly Correlated Wn Theory: W1-F12 and W2-F12. *J. Chem. Phys.* **2012**, *136*, 124114.

(17) Zhao, Y.; Xia, L. X.; Liao, X. B.; He, Q.; Zhao, M. X.; Truhlar, D. G. Extrapolation of High-Order Correlation Energies: The Wms Model. *Phys. Chem. Chem. Phys.* **2018**, *20*, 27375−27384.

(18) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Assessment of Gaussian-3 and Density-Functional Theories on the G3/05 Test Set of Experimental Energies. *J. Chem. Phys.* **2005**, *123*, 124107.

(19) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate Quantum Chemical Energies for 133 000 Organic Molecules. *Chem. Sci.* **2019**, *10*, 7449−7455.

(20) Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.

(21) Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, *9*, 3543−3556.

(22) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(23) Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *arXiv (Physics.Chemical Physics)*, November 28, 2019, 1911.10084, ver. 2. https://arxiv.org/abs/1911.10084 (accessed 2020-02-28).

(24) Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A. Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546−1559.

(25) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.

(26) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(27) von Rudorff, G. F.; von Lilienfeld, A. Atoms in Molecules from Alchemical Perturbation Density Functional Theory. *J. Phys. Chem. B* **2019**, *123*, 10073−10082.

(28) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.

(29) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. Outsmarting Quantum Chemistry through Transfer Learning. *ChemRxiv* **2018**, DOI: 10.26434/chemrxiv.6744440.v1.

(30) Smith, J. S.; Isayev, O.; Roitberg, A. E. Ani-1: An Extensible Neural Network Potential with Dft Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.

(31) Unke, O. T.; Meuwly, M. A Reactive, Scalable, and Transferable Model for Molecular Energies from a Neural Network Approach Based on Local Information. *J. Chem. Phys.* **2018**, *148*, 241708.

(32) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecules Neural Network. *Sci. Adv.* **2019**, *5*, No. eaav6490.

(33) Becke, A. D. Density-Functional Thermochemistry. Iii. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(34) Becke, A. D. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(35) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(36) Ward, L.; Blaiszik, B.; Foster, I.; Assary, R. S.; Narayanan, B.; Curtiss, L. Machine Learning Prediction of Accurate Atomization Energies of Organic Molecules from Low-Fidelity Quantum Chemical Calculations. *MRS Commun.* **2019**, *9*, 891−899.

(37) Pedley, J. B. *Thermochemical Data and Structures of Organic Compounds*; Thermodynamics Research Center: College Station, TX, 1994.

(38) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning. *J. Chem. Phys.* **2018**, *148*, 241717.

(39) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. *Gaussian 16*, rev. C.01; Gaussian, Inc.: Wallingford, CT, 2016.

(41) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106*, 1063−1079.

(42) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 Theory for Molecular Energies of First- and Second-Row Compounds. *J. Chem. Phys.* **1991**, *94*, 7221−7230.

(43) Raghavachari, K.; Stefanov, B. B.; Curtiss, L. A. Accurate Thermochemistry for Larger Molecules: Gaussian-2 Theory with Bond Separation Energies. *J. Chem. Phys.* **1997**, *106*, 6764−6767.

(44) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. Pubchem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.

(45) Grimme, S. Semiempirical Gga-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27*, 1787−1799.

(46) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.

(47) Schutt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Muller, K. R. Schnetpack: A Deep Learning Toolbox for Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.

(48) https://github.com/globus-labs/g4mp2-atomization-energy/tree/master/dlhub.

(49) https://github.com/globus-labs/g4mp2-atomization-energy.

(50) Dorofeeva, O. V.; Kolesnikova, I. N.; Marochkin, I. I.; Ryzhova, O. N. Assessment of Gaussian-4 Theory for the Computation of Enthalpies of Formation of Large Organic Molecules. *Struct. Chem.* **2011**, *22*, 1303−1314.