JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 4687−4698

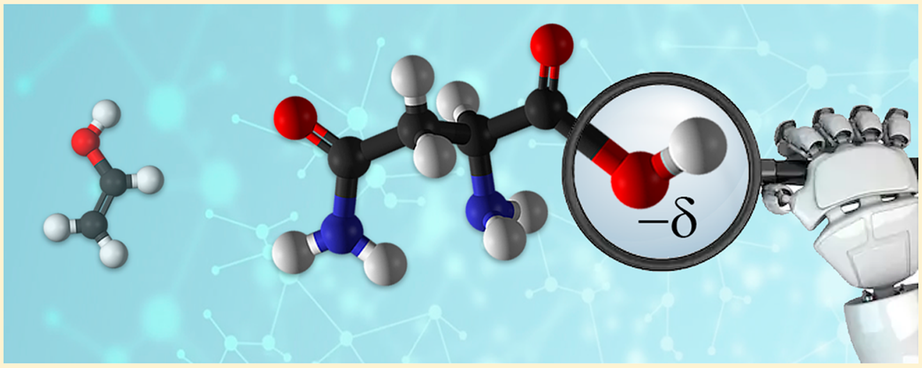
pubs.acs.org/JCTC

Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks

Benjamin Nebgen,[†] Nicholas Lubbers,[†] Justin S. Smith,[†,‡] Andrew E. Sifain,[†,§] Andrey Lokhov,[†] Olexandr Isayev,[∥] Adrian E. Roitberg,[‡] Kipton Barros,[†] and Sergei Tretiak*,[†]

[†]Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States
[‡]Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States
[§]Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90037, United States
[∥]UNC Eshelman School of Pharmacy University of North Carolina Chapel Hill, Chapel Hill, North Carolina 27599, United States

Supporting Information



ABSTRACT: The ability to accurately and efficiently compute quantum-mechanical partial atomistic charges has many practical applications, such as calculations of IR spectra, analysis of chemical bonding, and classical force field parametrization. Machine learning (ML) techniques provide a possible avenue for the efficient prediction of atomic partial charges. Modern ML advances in the prediction of molecular energies [i.e., the hierarchical interacting particle neural network (HIP-NN)] has provided the necessary model framework and architecture to predict transferable, extensible, and conformationally dynamic atomic partial charges based on reference density functional theory (DFT) simulations. Utilizing HIP-NN, we show that ML charge prediction can be highly accurate over a wide range of molecules (both small and large) across a variety of charge partitioning schemes such as the Hirshfeld, CM5, MSK, and NBO methods. To demonstrate transferability and size extensibility, we compare ML results with reference DFT calculations on the COMP6 benchmark, achieving errors of 0.004e$^-$ (elementary charge). This is remarkable since this benchmark contains two proteins that are multiple times larger than the largest molecules in the training set. An application of our atomic charge predictions on nonequilibrium geometries is the generation of IR spectra for organic molecules from dynamical trajectories on a variety of organic molecules, which show good agreement with calculated IR spectra with reference method. Critically, HIP-NN charge predictions are many orders of magnitude faster than direct DFT calculations. These combined results provide further evidence that ML (specifically HIP-NN) provides a pathway to greatly increase the range of feasible simulations while retaining quantum-level accuracy.

INTRODUCTION

The solution to the Schrödinger equation of quantum mechanics (QM) provides, in principle, a complete description of all chemical phenomena. However, exact solutions are infeasible for systems of practical interest due to the exponential scaling of computational cost with molecular size (i.e., the number of constituent atoms). Modern theoretical chemistry has turned to a wide range of approximate methods for the simulation of chemical systems with an important trade-off between accuracy and computational cost. For example, one of the most popular theoretical methods (with over 15000 publications per year)[1,2] is density functional theory (DFT), where the computational effort is reduced to roughly cubic scaling with molecular size. Yet, even cubic scaling is prohibitive for many applications. Physically relevant length and time scales are often completely inaccessible to DFT and higher-order QM methods. The development of more efficient methods that can retain quantum accuracy will open the door to a wealth of useful studies for physics, chemistry, biophysics, and materials science.

Received: May 29, 2018
Published: July 31, 2018

DOI: 10.1021/acs.jctc.8b00524
J. Chem. Theory Comput. 2018, 14, 4687−4698

A conventional approach to address these challenges is the formulation of better numerical techniques (e.g., to enable a more efficient solution of DFT and related theories). An alternative and emerging approach is to use Machine Learning (ML) models, which are poised to revolutionize the state of electronic structure methods. ML offers the accuracy of high fidelity quantum mechanical calculations at a fraction of the cost. ML methods commonly scale linearly with the number of atoms.[3,4] A typical ML approach utilizes a reference data set and supervised learning algorithm which constructs a robust model of the data. Since ML is data-dependent, it is not a replacement for, but a complement to existing electronic structure theory by enabling multistep modeling. It is well-known that what drives the accuracy and generality of a model is the data used to train it.[3,5−8] It was shown by many groups that ML models which utilize a proper descriptor and a reference database of small and diverse molecules or fragments, which represent "building blocks" for extended systems, can lead to size extensible potential energy predictions.[3,4,9,10] Prediction of molecular potential energies using ML has been explosively developing in recent years.[4,8,11−18] Modern ML models commonly fit DFT energy data sets to within 0.5 kcal/mol (1 kcal/mol is frequently quoted as a gold standard for chemical energy accuracy). Remarkably, this is closer to DFT than DFT is to the exact energy. Moreover, once trained, a ML prediction is usually thousands of times faster than a conventional DFT calculation. Many published ML property prediction models scale between $O(N)$ and $O(N^2)$.[4,11] Additionally, reported examples of ML in quantum chemistry beyond energy prediction include ML of band gap calculations,[19,20] high-throughput screening for materials discovery,[21,22] crystal structure prediction,[23] and excited state dynamics.[24−26]

Atomic charges constitute an important quantity allowing chemists to rationalize and extend their chemical intuition and to calculate a variety of molecular properties. Most charge assignment schemes require complete quantum mechanical calculations of the electronic density with subsequent partitioning of the latter into atom-centered point charges.[27] Recent work on electrostatic equilibration method (EEM) has demonstrated that accurate charges (root-mean-square error ≈ $0.01e^-$)[28] can be obtained with parametrized models, bypassing the Schrödinger Equation.[29,30] Still, these simulations required the solution of an $O(N^3)$ set of equations, which can be reduced to $O(N \log(N))$ using locality approximations and batching.[29] This illustrates how charge assignment is critical to broad areas in theoretical chemistry, ranging from modeling of infrared (IR) spectra,[31] evaluation of solvation free energies,[32,33] classical force field parametrization,[34,35] etc. Recent work has demonstrated the ability of ML to predict charges[36−39] and IR spectra[40,41] to high accuracy on a single system at a time. Other work has produced accurate results learning a variety of properties such as energies, forces, monopoles, dipoles, and quadrupoles,[42,43] the most general of these being Parkhill and co-workers' TensorMol.[3] Yet, many questions remain unexplored, including extensibility (i.e., how well can we predict charges on systems much larger than those included in the reference data set?), charge partitioning reliability (i.e., can all charge assignment methods be well-learned?), and transferability (i.e., can systems external to the reference data set be accurately predicted?).

In this article, we use the hierarchically interacting particle neural network[11] (HIP-NN) ML model to predict atomic charges from a variety of charge assignment schemes. Our goals are three-fold: first, we demonstrate that HIP-NN can accurately and efficiently predict charges for a wide range of small molecules, which is critical for classical force field parametrization. Second, we show that HIP-NN can effectively predict many different charge assignment schemes, allowing user customization for a variety of applications. Finally, we validate the extensibility of ML models validating them against significantly larger systems (such as druglike molecules, short peptides, and small proteins) than those included in the reference data set. This demonstrates conclusively that a carefully constructed reference data set combined with a properly optimized network can make reliable predictions on never seen before systems.

■ METHODS

Charge Partitioning Schemes. From a QM perspective, charge is distributed as a continuous density across the entire molecule. Unfortunately, there is no unique or exact way to condense charge density onto individual constituent atoms. However, the ability to describe charges as points on atoms leads to vastly simplified physics in a myriad of applications. There are three broad classes of atomic charge assignment schemes that can be derived from ab initio simulations.[27] First, the electron density obtained from ab initio calculations can be partitioned directly onto each atom. In this category, Mulliken charges[44] are the simplest, cheapest, and least accurate. Hirshfeld charges[45] are a more sophisticated and accurate version. Also, the natural bond orbital (NBO)[46,47] scheme assigns charges according to the natural atomic orbitals of a molecule, which are orbitals designed to capture the behavior of atomic orbitals in a given molecular environment. The second family of approaches is to use ab initio calculations to produce molecular properties, which guide charge assignment. Popular choices include charge multipoles and the electrostatic field at specific points surrounding the molecule. Then, the charges are fit to best recover these properties. Merz−Singh−Kollman (MSK)[48] charges are restrained to exactly reproduce the molecular dipole calculated from the continuous charge distribution. In the last family, one combines ab initio electronic density with gas-phase experimental measurements using a simple correction. An example is Charge Model 5 (CM5),[27] in which the correction is parametrized to replicate dipole moments for electrostatic energy calculations. Unlike MSK charges, CM5 charges are not constrained to exactly replicate the quantum molecular dipole. Rather, one expects CM5 charges to produce an approximate dipole moment. Section 1 of the Supporting Information contains further details on these charge schemes. In this work, we test the HIP-NN model trained against commonly used Hirshfeld, NBO, MSK, and CM5 charge schemes.

To illustrate these distinct charge schemes and to establish a baseline accuracy for our ML model, we consider charge assignment in the methanol molecule (CH_3OH). We select methanol because it has a strong molecular dipole moment and is simple enough for chemical intuition to be applicable. In Table 1, we report the charge on the carbon and oxygen atoms in methanol on a variety of charge schemes. The charges were generated using DFT calculations with wB97x functional and two different basis set sizes (additionally, results obtained with a simpler 6-31G basis are reported in the Supporting Information). There is a significant variation between the methods. Of the four charge schemes, only MSK assigns a

Table 1. Charges on Carbon and Oxygen Atoms in Methanol (CH_3OH) According to Various Charge Partitioning Schemes and Typical Basis Sets

atom	basis set	Hirshfeld	CM5	MSK	NBO
oxygen	6-31G*	−0.255	−0.475	−0.618	−0.751
oxygen	6-31+G*	−0.252	−0.472	−0.699	−0.777
carbon	6-31G*	−0.007	−0.132	0.142	−0.317
carbon	6-31+G*	−0.010	−0.134	0.234	−0.334

positive charge to the carbon atom. The Hirshfeld method attributes a nearly neutral charge to the carbon, while the remaining schemes assign various negative charge values to this atom. This demonstrates how differing charge schemes may operate in fundamentally disparate ways, each posing an independent challenge for ML modeling.

Both CM5 and Hirshfeld charges are nearly independent of basis set, with the difference between assigned charges under 0.01 e^- (elementary charge). NBO has a larger basis set dependence, with the charge variation within 0.03 e^-. Here, the local atomic charge densities that are used to partition the molecular charge density are nearly independent of basis set size. In contrast, MSK charges vary greatly with the basis set size. This is because the global constraint to reproduce the dipole moment introduces nonlocality into the charge partitioning. It is not obvious whether larger basis sets result in more practical charge assignments.[49] As basis sets get larger and orbitals are more diffuse, associating charge with specific atoms becomes more ambiguous. Within a single charge assignment scheme, it is unreasonable to expect errors smaller than 0.01 e^-. Thus, in the following analysis, we aim to build ML models of charge assignment that are accurate to within 0.01 e^-.

HIP-NN Approach. Here we briefly describe deep neural networks for chemical property prediction,[50,51] and, in particular, the hierarchically interacting particle neural network (HIP-NN) used in this work. For details, see Section 2 of the Supporting Information and ref 11.

To begin, a molecule is converted into a feature representation that can be parsed by a neural network. A variety of schemes are available,[52−57] such as the Coulomb matrix[58] or atom-centered symmetry functions suggested by Behler and used in ANI-1.[16] The feature representation used for HIP-NN is minimal: we use the atomic number of each atom and the pairwise distances between atoms. This simple representation ensures the network predictions satisfy translational, rotational, and reflection invariances. The features are passed through many hidden layers to produce a vector of new internal features or activations at each layer. Each layer's activations are produced using matrix-vector products and an element-wise nonlinearity. Finally, output layers form linear combinations of the activations to produce predictions at each atom. The internal weights used in these computations constitute the learnable parameters of the network. These parameters are fit to match the training data set using an iterative optimization process. In this work, each network has about 10^4 parameters.

The structure of HIP-NN is depicted in Figure 1. The network variables reside on each atom in the molecule, whose initial features encode the species of the atom. Green bars illustrate interaction layers, which allow sharing of information between nearby atoms. Interaction layers are comprised of many sensitivity functions, which allow atoms at different distances to interact in different ways. The sensitivity functions are constrained by two cutoff distances: the soft cutoff is the distance at which all interactions begin decreasing, and the hard cutoff is when all interactions fall to zero. Red squares in Figure 1 illustrate on-site layers, which perform processing of activations on each atom independently. Additionally, HIP-NN is organized into interaction blocks, each of which contains a single interaction layer, followed by a series of on-site layers. At the end of each series of on-site layers, an output layer gathers the current information on a given atom and produces a contribution to the atomic charge. In this way, the total charge on an atom is given by a sum over layer-wise contributions:

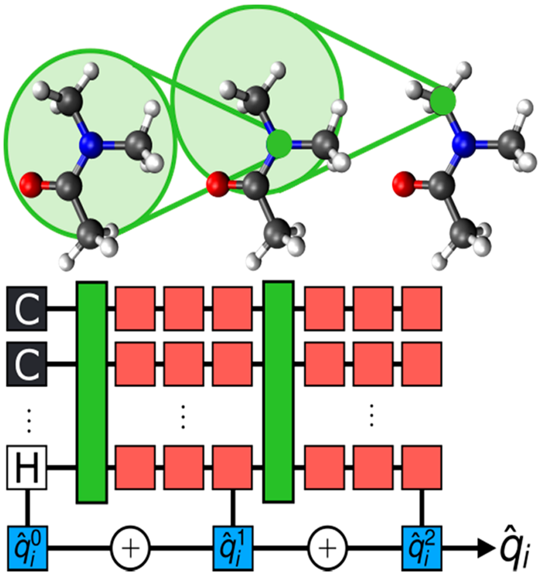


Figure 1. A diagrammatic representation of HIP-NN. Green bars represent interaction layers and red squares represent on-site layers for atoms. The blue squares are the output layers which return a series of corrections to the atomic charge. The molecule on top illustrates how information can be passed from one atom to another. This includes information being indirectly passed from distance atoms through an intermediate interaction layer.

$$\hat{q}_i = \sum_{j=0}^{N_{\text{interaction\_layers}}} \hat{q}_i^{\,j}$$

We use ADAM,[4,59] a variant of stochastic gradient descent (SGD) to learn the parameters for our networks. Briefly, the source data set is randomly partitioned into sets: training (60%), validation (20%), and test (20%). The training data is repeatedly fed through the network, and the parameters of the network are adjusted so that the outputs of the network tend toward the true reference charges (i.e., those from the given charge assignment). After each pass of the training data set, the performance on the validation set is recorded in order to estimate how the current parameters generalize to new data. The training process is terminated once the performance on the validation set stops improving. This early stopping procedure helps to prevent overfitting (i.e., helps the network to learn generalizable patterns, rather than irrelevant details specific to the training set). The model that scored the best on the validation set is kept and is used to measure the out-of-sample performance on the test set. Since the test set does not play any role in the fitting process, the error on the test set constitutes a fair measure of the network performance on data

DOI: 10.1021/acs.jctc.8b00524

Table 2. Parameterization for Various Networks Applied to Charge Training[a]

	layer 1				layer 2			
network name	sensitivity functions	soft cut-off (Å)	hard cut-off (Å)	number of units	sensitivity functions	soft cut-off (Å)	hard cut-off (Å)	number of units
R = 4.0	8	2.5	4.0	20	8	2.5	4.0	20
R = 6.0	10	3.0	6.0	20	8	2.5	4.0	20
R = 8.0	20	5.5	8.0	40	20	5.5	8.0	40

[a] Sensitivity functions shows the number of unique functions that facilitate the interactions between atoms at each interaction layer. Cut-off is the largest possible distance to the maximum of the last spatial sensitivity function, and hard cut-off is when all spatial sensitivity functions go to zero. Each network has three atomic layers following each interaction layer. Networks are named for the hard cut-off of the first interaction layer. As each network has two interaction layers, the maximum possible range (receptive field) of the network prediction is the sum of the cut-offs for the interaction layers.

that is similar in scope to the training data set. Section 2 of the Supporting Information contains a more detailed description of the training process.

Training Databases and Tested Networks. Here our choices were motivated by previous experience of learning energies within HIP-NN and ANI-1[4] approaches. To help understand the transferability of ML algorithms, we use two different databases to train the HIP-NN charge model. The first we denote GDB-5: a subsample of GDB-11[60,61] data set used to train the ANI-1 ML force field.[4,5] The configurations include up to 5 heavy atoms (of types C, N, and O) all in charge neutral molecules. The conformations are based on the ANI-1 normal mode sampling scheme. This data set contains 517133 individual molecular structures. The subsampling of the GDB-11 was done for two reasons: first, the smaller data set allowed us to produce charges for many different partitioning schemes without expending excessive computational resources on DFT calculations. Second, the data set of only small structures is used to determine how networks trained to small molecules can predict charges on larger systems. The Gaussian09[62] computational suite was used to run the DFT calculations with the wB97*x*/6-31G*[63,64] functional and basis set. For each molecule in the data set, Hirshfeld, CM5, NBO, and MSK charges were generated for every atom. This allowed us to construct networks trained to reproduce each type of charge assignment.

The second database examined, specifically the data set used to train the ANI-1x potential, was developed by Smith et al.[9] using an active learning method, whereby the training set is progressively built by selectively performing new DFT simulations to cover the regions of greatest model uncertainty.[4,40,65−68] ANI-1x data set samples vast and very diverse chemical space and include aromatic and heteroaromatic systems, most common substituents and functional groups containing only C, H, N, and O atoms. While a summary of methods can be found in Section 3 of the Supporting Information, one key component of this procedure is "Query-by-Committee".[69] QBC is performed by training an ensemble of ANI models to related data and using them to predict values on the same unknown system. The variation between the predictions of each model becomes a proxy for the confidence of prediction on each data point. In other words, if the chemical system is well represented in the training data, all the networks will give similar predictions, while if the system is not well-represented by the training data, the predictions of the various networks are likely to be widely disparate. The complete technical details are described in ref 9. ANI-1x was produced by sampling chemical space and only performing reference calculations on those systems for which an ensemble of neural networks produced no consensus energy value. The procedure produced a data set of 5.5 million structures, with an average of 15 atoms per structure and a maximum size under 60. For our charge modeling, we use 6% of ANI-1x for the training set and an additional 2% each for test and validation sets. All molecules in both training data sets are charge neutral and closed shell, thus our networks are only applicable to neutral, closed shell molecules.

The active learning process to construct the ANI-1x data set is based on learning the energy of thermally accessible molecules and conformations, rather than charge learning. While it would be possible to construct an active learning database using charge rather than energy for QBC, we believe our results show this is not necessary. Due to the active learning procedure, the ANI-1x data set contains a vast array of diverse configurations and conformations which covers the space of interest to bio and organic chemists. While ANI-1x does a very good job of sampling organic molecules, it has two limitations of note: first, the AL sampling only searched for molecules that can be constructed with hydrogen, carbon, nitrogen, and oxygen atoms. Second, while dimer interactions were sampled, extensive solvent interactions arising from many body interactions were not. This limits the applicability of this training data set when examining solvated systems.

Training a single HIP-NN model to the GDB-5 energy data set takes roughly 12 h with GPU acceleration using a NVIDIA 1080Ti, while training to ANI-1x takes roughly 120 h. Roughly a factor of 3 can be saved by initially training on GDB-5 and then "transferring" the model to train on ANI-1x database. We observe a similar cost savings for HIP-NN charge models trained using the same transfer learning protocol. No significant loss of accuracy was observed between the traditionally trained network and the transfer network, as seen in Table S1.

Table 2 describes the various forms of HIP-NN that are examined in this work. The R = 6.0 network is trained on both the small GDB-5 as well as the active learned ANI-1x training data. The goal of this network is to allow all distance functions to be trained even when using data sets of very small molecules. The R = 4.0 network is a network designed to focus only on nearest neighbor interactions. The R = 8.0 network closely resembles the design for predicting energy in the original HIP-NN paper.[11]

RESULTS

Benchmark Databases. To illustrate the generality of HIP-NN, we computed Mulliken, Hirshfeld, CM5, MSK, and NBO reference charges for the entire GDB-5 database and trained the same HIP-NN network on each of the charge

DOI: 10.1021/acs.jctc.8b00524

schemes. Figure 2 reports the mean absolute error (MAE) and root mean squared error (RMSE) for each network trained to its respective charge scheme. An RMSE much greater than MAE indicates that distribution of errors has a long tail (i.e., it contains outliers). HIP-NN learns almost all of the charge schemes with a MAE of about 0.005 e$^-$ and a slightly larger RMSE. This is significantly better than our target precision of 0.01 e$^-$, set by the consistency of the charge models. Additionally, the prediction of these charges takes approximately 0.234 ms per conformation. Thus, using these neural networks to predict charges on molecules similar to those found in GDB-5 is significantly computationally cheaper, but no less precise, than using a quantum calculation. The one exception to this is MSK charges, which are replicated to approximately a precision of 0.02 e$^-$. We believe the explanation is as follows: MSK charges are constrained posthoc to exactly replicate the dipole moment of the molecule. This makes the charge scheme nonlocal; charges may differ between similar conformations in order to produce the correct dipole moment. This nonlocality is potentially difficult to represent within the HIP-NN model, which describes only local interactions.

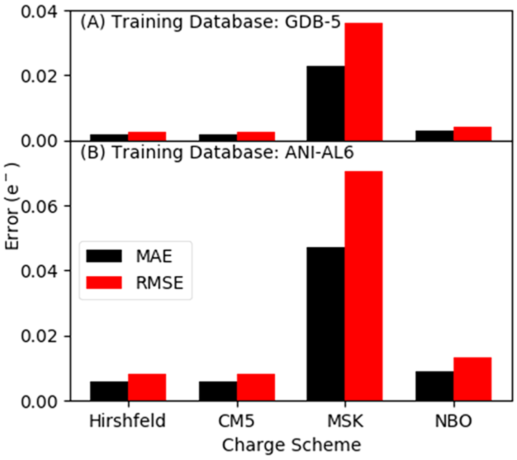


Figure 2. Test set mean absolute and root-mean-square errors for the R = 6.0 (see Table 2) network trained to the GDB-5 database (A) and ANI-1x database (B) using different charge schemes. HIP-NN is able to learn almost all charge schemes to equal precision, with the exception of the MSK scheme. While the test set error for the more diverse ANI-1x data set is larger than the test set error for the GDB-5 data set, the predictive accuracy of networks trained to ANI-1x is significantly better (see Figure 3).

To test the transferability and extensibility of networks trained to both the GDB-5 and ANI-1x databases, we applied various versions of HIP-NN (Table 2) to the COMP6 benchmark suite[9] of organic molecules (Figure 3). COMP6 samples a diverse selection of molecular configurations and conformations aimed at validating the accuracy of an ML potential. This suite supplies energies, forces, and Hirshfeld and CM5 charges for the validation of ML methods, all computed using wB97x/6-31G*. COMP6 contains six benchmarks, each containing different types of molecules: GDB-7to9 (built from GDB-11[60,61]), GDB-10to13 (built from GDB-11[60,61] and GDB-13[70]), Tripeptides, DrugBank,[71] ANI-MD, and S66 × 8 noncovalent interaction[72] benchmarks. These benchmarks are described below.

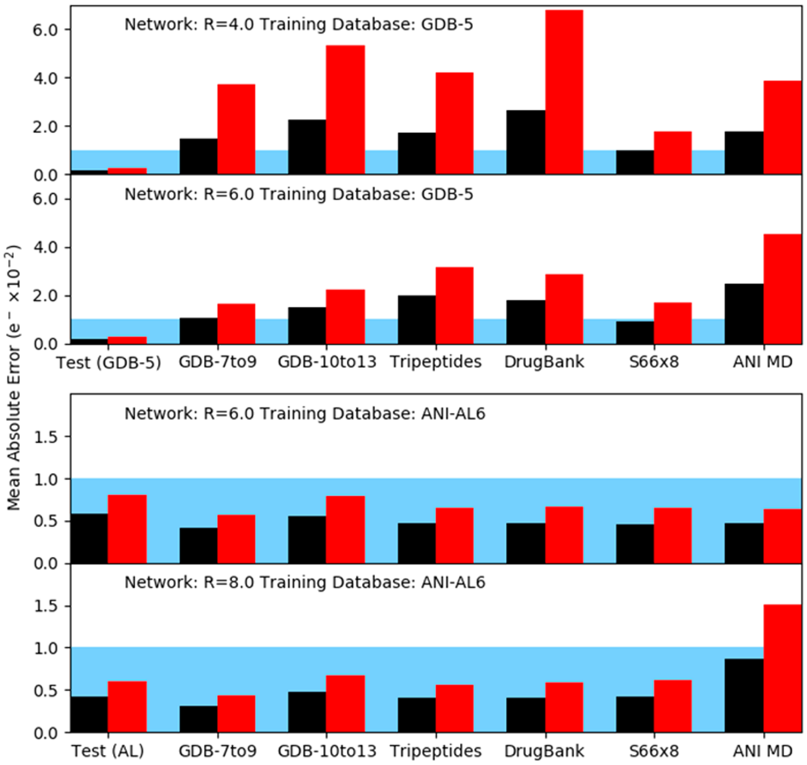


Figure 3. Predictions of various versions of HIP-NN on the COMP6 benchmark with all charges originating from the Hirshfeld partitioning scheme. Mean absolute error (MAE) and root mean squared error (RMSE) given for all data sets. The blue shading indicates 0.01e$^-$, the target error for this method. The R = 6.0 network trained to ANI-1x has an error significantly less than the target for all pieces of the COMP6 benchmark. A full table of this data can be found in Section 4 of the Supporting Information.

The GDB-7to9 and GDB-10to13 benchmarks aim to validate the universality of the predictor on a comprehensive list of artificially generated small molecules with many nonequilibrium conformations per molecule. These sets contain 1500 and 2996 configurations with 36000 and 47670 total conformers, respectively. The tripeptide benchmark provides 1984 random conformations for 248 generated tripeptides. The DrugBank benchmark provides 13379 random configurations of 837 drug molecules from the DrugBank[71] database. The S66 × 8[72] is a benchmark of 66 different interacting dimers for validating a methods accuracy in reproducing noncovalent interactions. Interactions included in this data set include hydrogen bonding, London interactions, and $\pi-\pi$ stacking. ANI-MD contains random structures from MD trajectories generated with the ANI-1x active learned ML potential[9] for 13 common drug molecules and two small proteins. The largest of these structures contains 312 atoms, which is more than 6 times larger than the largest system used to train the network. More details on these data sets are provided in Section 4 of the Supporting Information.

Figure 3 illustrates the accuracy of various combinations of training set and network shape applied to the COMP6 benchmark. Critically, we consider the R = 6.0 network trained to ANI-1x to be the best performing network with a MAE and RMSE not exceeding 0.004 e^- and 0.007 e^-, respectively, on any of the sub-benchmarks. Further, when examining only the largest structures in this database (128 conformations of 1L2Y contained in ANI-MD), the error rises negligibly to 0.00606 e^- and 0.00795 e^- MAE and RMSE, respectively. This protein is 6 times larger than the largest system in ANI-1x and clearly shows extensibility of the charge prediction scheme. Additionally, the MAE of these predictions is only slightly larger than the dependence on the basis set size for even Hirshfeld charges (0.0048 e^- for carbon). This is to say that the accuracy of the predictions on extremely large systems is roughly the same as the certainty of the charge assignment itself.

Figure 3, column 1, shows accuracies of the networks given in Table 2 on their respective charge schemes and test sets. This shows that GDB-5 charges are much easier to learn than those in ANI-1x. This is due to the smaller size of GDB-5, as well as the lower variety in the data. For example, GDB-5 lacks nonbonding interactions. The rest of the columns in Figure 3 demonstrate the network performance on each benchmark in COMP6. The networks trained to GDB-5 predict charges on the benchmark data with significantly lower accuracy; training to GDB-5 does not generalize well.

Interestingly, when training to the ANI-1x data set, the R = 6.0 network outperforms the R = 8.0 network (Figure 3, bottom half). This illustrates that local models need to ensure proper special coverage (a cutoff radius the size of the molecules in the training set) of the given training set, or else generalization to large structures can be hindered. Thus, we would like to emphasize the importance of limiting longer range interactions when considering the extensibility of NNs. Section 5 of the Supporting Information contains the exact training accuracies across all benchmarks.

Figure 4 displays the correlation between charge and predicted charge on the COMP6 benchmark for an R = 6.0 network trained to ANI-1x. A similar plot for a network trained to the GDB-5 database is reported as Figure S1 in the Supporting Information. The ANI-1x network (Figure 4) performs significantly better and exhibits far fewer outlying predictions than the network trained to GDB-5. We emphasize that the HIP-NN model form is completely identical for each network, the difference in prediction quality is dictated entirely by the parameter values learned from the training data set. . In both cases, the network contains far fewer parameters ($\sim 10^4$) than training conformations ($\sim 3 \times 10^5$ for each data set). This illustrates the need to train to larger and more diverse chemical structures than present in GDB-5 in order to capture all of the subtle chemical interactions pertinent to charge assignment.

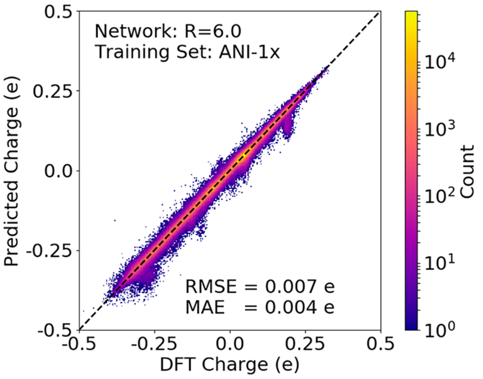


Figure 4. Correlation plot of the R = 6.0 network (see Table 2) when applied to the Hirshfeld charges for the entire COMP6 benchmark. The networked trained to ANI-1x is pictured here while the correlation plot for the network trained to GDB-5 can be found in Section 5 of the Supporting Information.

In general, the previous results show that accurate and extensible charge prediction can be obtained by learning to a data set that contains larger molecules and more data diversity, even though such data sets are more difficult to learn, producing a larger MAE test set error. However, we concluded that properly constructed network trained on a diverse data set is capable of making accurate predictions on systems significantly larger than those seen in the training set with only a minimal loss in accuracy.

Simulation of IR Spectra. One of the applications for a rapid charge prediction scheme is the construction of IR spectra from molecular dynamics data. Here we construct fully ML-based IR spectra by generating gas phase MD trajectories for a few small molecules (methanol, ethanal, acetamide, and dimethylacetamide) using the ANI-1x potential.[9] Trajectories of 100 ps with a 0.1 fs time step were collected on a single molecule at 300 K using an NVT thermostat at 300 K. From the trajectory conformations, neural network charges were generated at every point using the R = 6.0 HIP-NN trained to GDB-5 Hirshfeld charges. GDB-5 serves as a sufficient training database because the molecules in the IR study are of a similar size and composition. Using the position and ML charge data for each frame, a molecular dipole is constructed. The IR spectrum is constructed (red dashed line) from the Fourier transform of the autocorrelation of the dipole moment using the code by Efrem Braun.[73] To produce a fair comparison to DFT, we also generate the IR spectrum with the same algorithm but using true Hirshfeld and CM5 charges computed from the same trajectory. We refer to these as the "ML charge" and the "QM charge" in Figure 5, respectively. Figure 5 shows that both curves virtually coincide, demonstrating the accuracy of ML charge predictions.

This needs to be contrasted to the other sources of error underlining the modeling. First of all, accuracy of ANI-1x force field potential can be tested for selected cases of methanol and acetamide (owing to a large numerical expense) by calculating

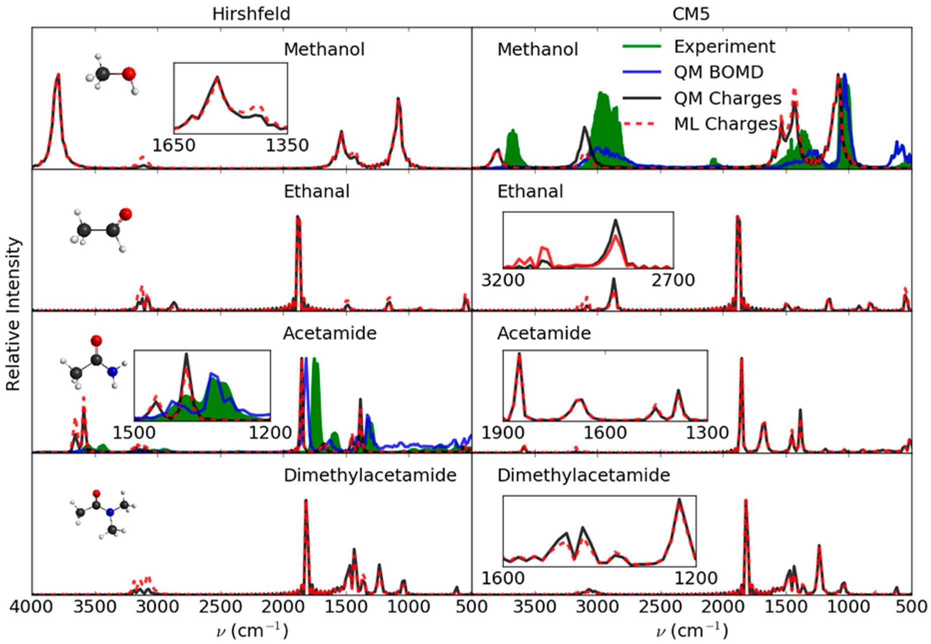


Figure 5. Comparison of IR spectrum predicted by ML (red dashed line) to experiment (green shade), quantum BOMD simulations (blue line), and exact charges along the ML dynamics trajectory (black). The red and black lines utilize Hirshfeld and CM5 charges on the left and right frames, respectively. Comparison between the red and black lines illustrates how precise the ML charge scheme is, as there is almost no difference between these curves. Comparison between the ML spectra and experiment show reasonable agreement for all molecules except ethanal. For methanol and acetamide, full QM trajectories and dipoles were computed, which show that a significant amount of error in the ML spectra arise from the underlying QM methods.

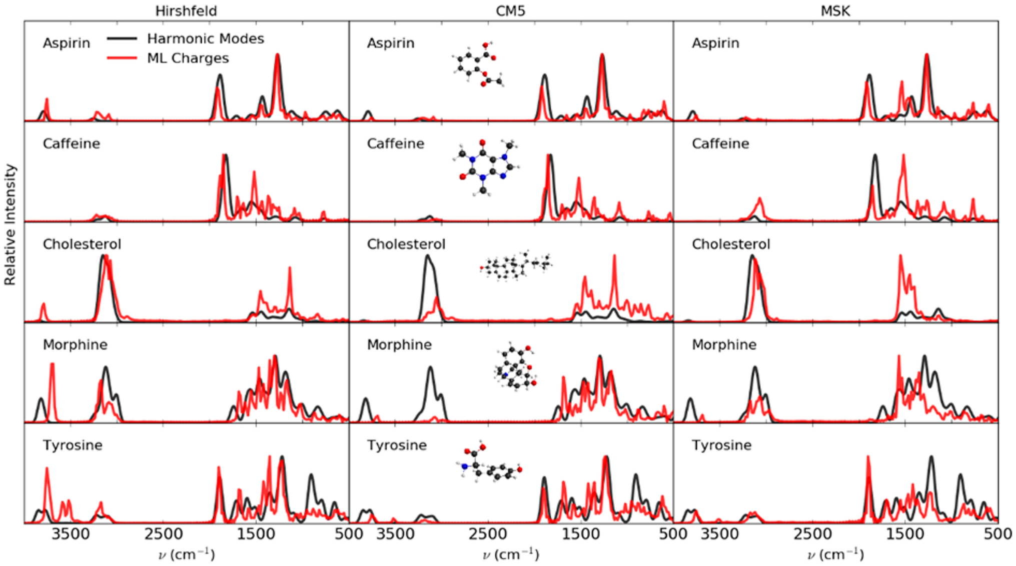


Figure 6. IR spectrum generated completely by machine learning (red line) and from a normal-mode analysis (black line). The ML spectrum was generated with an identical method as in Figure 5. Different charge assignment schemes produce significantly different IR spectra with Hirshfeld charges being the most reliable for larger molecules.

the IR spectrum using the DFT Born–Oppenheimer Molecular Dynamics. Here 20 parallel trajectories were run with a time step of 0.2 fs at 300 K for 5.5 ps using the fully quantum gradient computed at the wB97*x*/6-31G* level. The autocorrelation function of the quantum dipole moment was taken for each of these trajectories separately and averaged together. Then, the IR spectrum was generated in the same way (Figure 5, blue line). The second error is related to inaccuracy of the DFT model compared to gas phase experimental IR spectra.[74] Both of the charge-based IR spectra are significantly different from the full DFT Born–Oppenheimer Molecular Dynamics IR spectra. While the agreement to experiment is generally good (green shade), it is not perfect. Much of this failure can be attributed to the underlying QM methodology, as the DFT Born–Oppenheimer Molecular Dynamics simulations exhibit the same failures. Additionally, the DFT-BOMD spectrum is red-shifted compared to the ML spectrum. This is likely due to the larger time step of the DFT-BOMD calculations, which was used due to computational limitations. See Section 6 of the Supporting

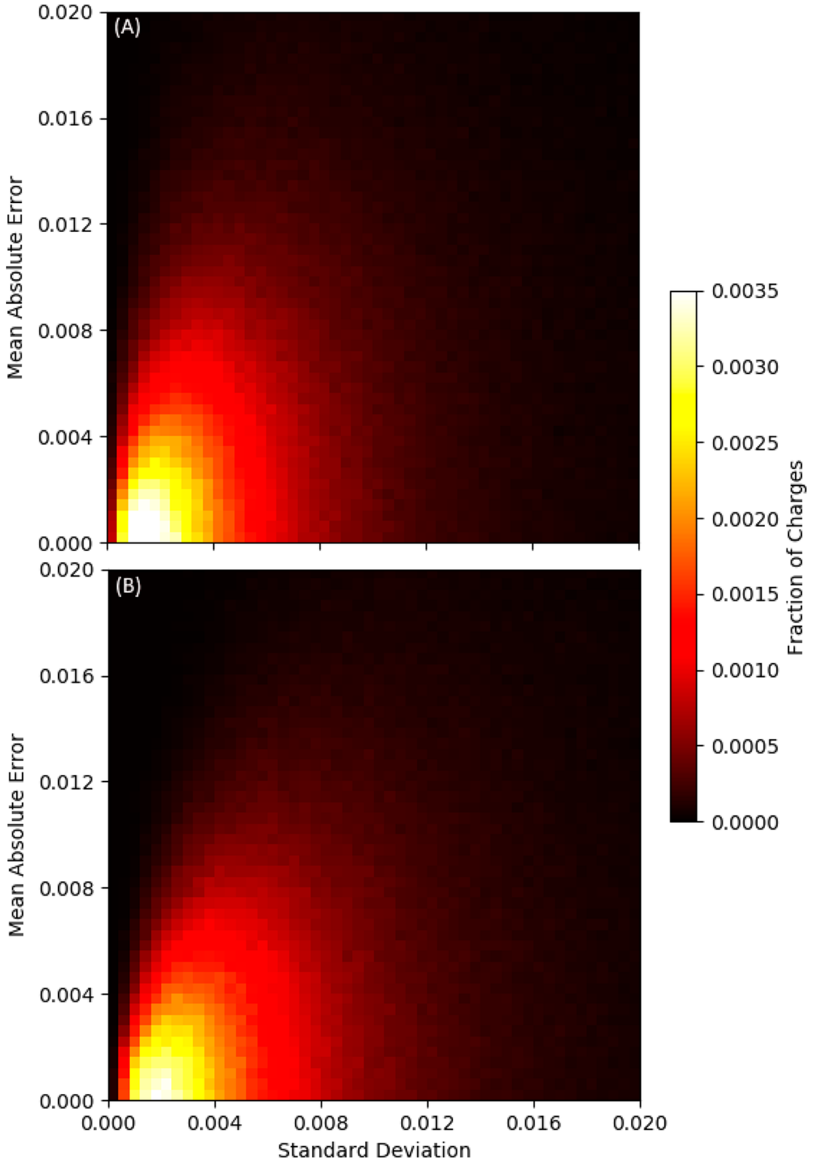


Figure 7. Ensembles of (A) 4 and (B) 8 networks, trained on GDB-5, predicting charges on GDB-7−9 structures. This demonstration shows that standard deviation between an ensemble of neural networks can be used to limit the maximum absolute error of the ensemble. The larger number of neural networks makes this effect more prominent.

Information for a thorough comparison of ML, DFT, harmonic mode, and experimental charge spectra on several small molecules.

Figure 6 is an application of the ML calculation of IR spectra, using the R = 6.0 network trained to ANI-1x, to larger molecules including aspirin, caffeine, cholesterol, morphine, and tyrosine. Caffeine, cholesterol, and morphine do not appear in the training data set, and so this application also tests the extensibility of our methods. Since DFT-BOMD would be too expensive for systems of this size and experimental IR spectra were all performed in the condensed phase, these ML results are compared to spectral calculations using vibrational normal modes. There is noticeable variance between networks trained to different charge schemes, with Hirshfeld charges giving the best agreement with the normal mode spectra. The overall agreement is good regardless of charge scheme, giving further evidence that our methods produce an extensible model.

Computationally, molecular dynamics and charge generation using ANI and HIP-NN is extremely fast, requiring roughly 25 min on a GPU-equipped workstation to generate an IR spectrum. For comparison, the DFT-BOMD calculation for a single molecule took 20000 CPU hours, leading to a speedup of over 4 orders of magnitude. For larger molecules such as cholesterol, the ML compute time increased negligibly to 30 min, while performing this calculation with BOMD would require an intractable 300000 CPU hours.

Charge Prediction on Proteins. Another application of our models is the prediction of charges on systems too large for quantum chemistry (e.g., proteins). Due to the speed of ML charge prediction, it is possible to dynamically assign charges during an MD run for an entire protein, allowing for a dynamic estimate of the coulomb contribution to the force field. As a test of the accuracy of this method on systems of this size, we extracted two proteins from the ANI-MD benchmark: Chignolin (IUAO)[75] and a Trp-Cage mini-protein Construct (1L2Y).[76] These proteins contain 149 and 312 atoms, respectively, and are the largest systems where reference DFT charges were gathered. The geometries originate from molecular dynamics calculations using the ANI-1x potential. As additional Supporting Information, we have included a spreadsheet with the reference and ML predicted CM5 and Hirshfeld charges for lowest energy configuration of these two systems. Importantly, the MAE of our charge predictions on these systems is just over 0.006 e^-, which is only slightly greater than the error on the total ANI-MD benchmark. Thus, our ML charge predictor is extensible: it is accurate when applied to systems 6 times larger than those included in the training set. Further, we analyze our prediction accuracy by atom type in Section 7 of the Supporting Information. Finally, as a test of speed, we predicted charges on the glucoamylase protein (1AYX), which has a mass of over 50 kDa and a length of 492 residues.[77] Since this protein is too large for DFT calculations, the goal was to explore how fast we could perform our analysis. For prediction, crystallographic water was removed and protons were added using the Reduce program.[78] In total, this methodology took less than 2 min to predict charges on the entire system. Such a prediction would be almost impossible with traditional quantum-chemical methods. On very large systems such as this protein, we find that the total predicted charge of the system does not remain precisely zero. However, the total charge defect is around 0.001 e^- per atom. A more in depth discussion can be found in Section 8 of the SI.

Error Estimation. The ability of ML to determine the uncertainty of a prediction through schemes like "Query by Committee" is extremely powerful because it allows for a minimum number of reference calculations to be run when constructing optimized training sets. Although active learning has been studied in the context of molecular energy modeling,[4,9,40,65−68] this concept has yet to be applied to charge prediction. Here we demonstrate how the consensus of an ensemble of neural networks can be used as a proxy for their accuracy in a "Query by Committee" framework (see earlier discussion in Training Databases and Tested Networks section).[69]

To test this for charges, two ensembles (of size 4 and 8) of R = 6.0 HIP-NN were trained to the GDB-5 database. The networks were then applied to predict charges in GDB-7to9. For each atomic charge, the ensemble average prediction error and ensemble prediction standard deviation was computed.

In Figure 7, we show density maps of prediction error versus ensemble standard deviation for the GDB-7to9 data set for each ensemble. The standard deviation provides a typical bound for the ensemble error; the upper left quadrant of the density map is empty, demonstrating that there are no points with low ensemble standard deviation and high error. Points in the upper-left quadrant would be problematic for QBC; they correspond to points which are predicted confidently but inaccurately by the ensemble.

CONCLUSION

In this study, we have demonstrated the ability of HIP-NN to learn and predict several charge partitioning schemes for neutral organic molecules made of carbon, hydrogen, nitrogen, and oxygen. HIP-NN was first applied to predicting molecular energies. We find, however, that by minimally modifying the output layer structure, it predicts atomic charges with an accuracy comparable to or better than the basis set error of roughly 0.01 e^-. Our best performing R = 6.0 network can predict charge with a MAE and RMSE, not exceeding 0.004 e^- and 0.007 e^-, respectively, across all of COMP6 benchmark set. Since COMP6 contains structures 6 times larger than those seen in ANI-1x, such high accuracy on the COMP6 benchmark show the HIP-NN charge model is extensible to large molecules. This combined achievement of accuracy and extensibility sets the HIP-NN charge model apart from any previously published work in ML charge prediction. Additionally, the computational cost of ML charge prediction is roughly 4 orders of magnitude faster than reference DFT calculations for small molecules.

However, there are a few limitations to this methodology. Most importantly, the training data sets are only designed to predict charges on organic molecules containing carbon, hydrogen, nitrogen, and oxygen. The network can be made more versatile at the cost of running addition quantum calculations for inclusion in the training set and the respective retraining of the neural nets. Another fundamental limitation is the lack of anionic or cationic (i.e., charged) species in both the training and testing set. Charge localization (i.e., the spatial location of an electron or a hole when considering large system) is a problem that is yet to be solved with ML algorithms. Additionally, we find MSK charges, which are constrained to replicate the quantum molecular dipole, pose a more difficult learning task for local ML models than local charge schemes owing to nonlocal nature of the scheme construction.

We demonstrate two applications of the HIP-NN charge model presented in this work. First, we use a trained model to calculate the IR spectrum for both small and large organic molecules. These spectra are in excellent agreement with other theoretical methods, and no degradation in performance is observed for molecules larger than those in the training set. This demonstrates transferability and extensibility of charge prediction over a variety of systems without specifically tuning the network. This makes the presented HIP-NN model a concrete step toward a universal ML charge predictor. However, quantitative agreement with experiment is not achieved, showing that the underlying DFT methods are insufficient. This leaves us with a path toward systematic improvement: performing higher accuracy quantum-chemical calculations on our data set should lead to a substantial improvement in accuracy while retaining the same computational cost at run-time. Second, we generate charges for a large protein, where the application of quantum mechanics is extremely difficult. The speed of this prediction is evidence that ML methods can generate dynamic charges for fast and accurate electrostatic calculations in MD. In this work, we limited ourselves to a simple protein; however, the same methodology could be extended to metalloenzymes or proteins that contain metal ions. This would require substantial additional calculations of metal compounds in different charge and spin states.

Additionally, we used the concept of query-by-committee to show a correlation between the accuracy and standard deviation of predictions made by an ensemble of ML models; when the ensemble agrees on a charge prediction, the prediction is more likely to be accurate. This is critical to the advancement of ML as applied to chemistry, as it facilitates active learning for the development of strong, diverse training sets. This is also useful when making predictions for unexplored systems, as the confidence for any ML prediction may be obtained without reference QM calculations.

The strong performance on COMP6 benchmarks and other applications demonstrates transferability of our ML approach to systems larger and more diverse than in the training set. Future models need not scale their data to sizes of target molecular systems. Data sets can be built using fragmentation and/or sampling of small molecules, and uncertainty can be estimated with query-by-committee. This provides a path forward to estimating chemical properties of systems intractable by QM methods.

Despite our success in training HIP-NN to various charge schemes, the underlying QM charge partitioning itself has drawbacks. The full charge density cannot be exactly represented within any charge assignment scheme. Two consequences are that different schemes disagree with each other, and that charge-derived quantities (such as the molecular dipole moment) may not be faithfully represented by atomic charge assignment. Future work may focus on overcoming the limitations of charge schemes by focusing instead on predictions of experimentally accessible quantities. Additionally, a more widely applicable model will be obtained by adding more element types (such as chlorine and sulfur) to the training set as well as specific sampling of many-body solvent interactions. The many possible applications of ML to quantum chemical properties (such as bond orders, vibrational and electronic excitations, electron density, and dipole moment) indicate tremendous potential present at the intersection of quantum chemistry and machine learning.

In summary, our results provide a complete analysis of which charge partitioning schemes are easier for ML models to learn and further evidence that machine learning techniques trained to data sets of small molecules can replicate charge partitioning schemes from high-level quantum mechanical calculations of larger molecules with orders of magnitude speedup.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00524.

Description of the charge models used, an overview of the NN training procedure, an explanation of active learning, details on the COMP6 benchmark, quantitative NN accuracy data, experimental IR spectra, prediction accuracy by atom type, and a discussion of charge conservation for the NN models (PDF)

Spreadsheet with the reference and ML predicted CM5 and Hirshfeld charges (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: serg@lanl.gov.

ORCID

Andrew E. Sifain: 0000-0002-2964-1923
Olexandr Isayev: 0000-0001-7581-8497
Adrian E. Roitberg: 0000-0003-3963-8784
Sergei Tretiak: 0000-0001-5547-3647

Funding

The authors acknowledge support of the U.S. Department of Energy through the Los Alamos National Laboratory (LANL) LDRD Program. LANL is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. O.I. also acknowledges support from DOD-ONR (N00014-16-1-2311) and Eshelman Institute for the Innovation award.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was done in part at the Center for Nonlinear Studies (CNLS) and the Center for Integrated Nanotechnologies (CINT) at LANL. We also acknowledge the LANL Institutional Computing (IC) program and ACL data team for providing computational resources. O.I. thanks CNLS for their support and hospitality.

■ REFERENCES

(1) Becke, A. D. Perspective: Fifty years of density-functional theory in chemical physics. *J. Chem. Phys.* **2014**, *140*, 18A301.

(2) Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* **2015**, *87*, 897−923.

(3) Yao, K.; Herr, J. E.; Toth, D. W.; McKintyre, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261−2269.

(4) 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.

(5) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193.

(6) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(7) 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*, DOI: 10.1103/PhysRevLett.108.058301.

(8) 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.

(9) Smith, J. S.; Isayev, O.; Nebgen, B.; Roitberg, A. E.; Lubbers, N. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(10) Huang, B.; von Lilienfeld, O. A. The "DNA" of chemistry: Scalable quantum machine learning with "amons". *arXiv:1707.04146***2017**.

(11) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **2018**, *148*, 241715.

(12) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(13) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212* **2017**.

(14) Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **2015**, *115*, 1032−1050.

(15) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, 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.

(16) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(17) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(18) Han, J.; Zhang, L.; Car, r.; E, W. Deep Potential: a general representation of a many-body potential energy surface. *arXiv: 1707.01478***2017**, 01478.

(19) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156−163.

(20) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.

(21) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure−Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939−8954.

(22) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 014106.

(23) Morgan, D.; Ceder, G.; Curtarolo, S. High-throughput and data mining with ab initio methods. *Meas. Sci. Technol.* **2005**, *16*, 296−301.

(24) Sun, B.; Fernandez, M.; Barnard, A. S. Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 2413−2423.

(25) Häse, F.; Kreisbeck, C.; Aspuru-Guzik, A. Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. *Chem. Sci.* **2017**, *8*, 8419−8426.

(26) Häse, F.; Valleau, S.; Pyzer-Knapp, E.; Aspuru-Guzik, A. Machine learning exciton dynamics. *Chem. Sci.* **2016**, *7*, 5139−5147.

(27) Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. Theory Comput.* **2012**, *8*, 527−541.

(28) Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. The electronegativity equalization method and the split charge equilibration applied to organic systems: Parametrization, validation, and comparison. *J. Chem. Phys.* **2009**, *131*, 044127.

(29) Ionescu, C.-M.; Sehnal, D.; Falginella, F. L.; Pant, P.; Pravda, L.; Bouchal, T.; Svobodová Vařeková, R.; Geidl, S.; Koča, J. AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. *J. Cheminf.* **2015**, *7*, 50.

(30) Verstraelen, T.; Bultinck, P. Can the electronegativity equalization method predict spectroscopic properties? *Spectrochim. Acta, Part A* **2015**, *136*, 76−80.

(31) Nakamura, R.; Machida, K.; Oobatake, M.; Hayashi, S. Molecular dynamics simulation of infrared spectra and average structure of benzoic acid crystal. *Mol. Phys.* **1988**, *64*, 215−227.

(32) Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011**, *7*, 4026−4037.

(33) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760−768.

(34) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155−3168.

(35) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(36) Morawietz, T.; Sharma, V.; Behler, J. A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. *J. Chem. Phys.* **2012**, *136*, 064103.

(37) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 045131.

(38) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.

(39) Artrith, N.; Morawietz, T.; Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2011**, *83*, 153101.

(40) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.

(41) Quaranta, V.; Hellström, M.; Behler, J.; Kullgren, J.; Mitev, P. D.; Hermansson, K. Maximally resolved anharmonic OH vibrational spectrum of the water/ZnO($10\bar{1}0$) interface from a high-dimensional neural network potential. *J. Chem. Phys.* **2018**, *148*, 241720.

(42) Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.

(43) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58*, 579−590.

(44) Mulliken, R. S. Electronic Population Analysis on LCAO−MO Molecular Wave Functions. *J. Chem. Phys.* **1955**, *23*, 1833−1840.

(45) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.

(46) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Landis, C. R.; Weinhold, F. *NBO 6.0*; Theoretical Chemistry Institute, University of Wisconsin: Madison, 2013.

(47) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.

(48) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.

(49) Hands, M. D.; Slipchenko, L. V. Intermolecular Interactions in Complex Liquids: Effective Fragment Potential Investigation of Water−tert-Butanol Mixtures. *J. Phys. Chem. B* **2012**, *116*, 2775−2786.

(50) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P., Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28*, Cortex, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015.

(51) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.

(52) Bartók, A. P. *Gaussian Approximation Potential: An interatomic potential derived from first principles Quantum Mechanics*; Cambridge, 2009.

(53) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.

(54) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084−1093.

(55) Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316−330.

(56) Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153−1173.

(57) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

DOI: 10.1021/acs.jctc.8b00524

(58) 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.

(59) Kingma, D. P.; Ba, J., Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* **2014**.

(60) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−1508.

(61) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.

(62) 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.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian*, D.01; 2016.

(63) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(64) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654−3665.

(65) Podryabinkin, E. V.; Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **2017**, *140*, 171−180.

(66) Browning, N. J.; Ramakrishnan, R.; von Lilienfeld, O. A.; Roethlisberger, U. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. *J. Phys. Chem. Lett.* **2017**, *8*, 1351−1359.

(67) Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **2017**, *146*, 244108.

(68) Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978−10985.

(69) Seung, H. S.; Opper, M.; Sompolinsky, H. Query by committee. *Proceedings of the fifth annual workshop on Computational learning theory* **1992**, 287−294.

(70) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.

(71) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−D1097.

(72) Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. L. The S66 × 8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905−20925.

(73) Braun, E. Calculating an IR Spectra from a LAMMPS Simulation. https://github.com/EfremBraun/calc-ir-spectra-from-lammps (accessed Oct 1, 2017).

(74) Stein, S. E. Infrared Spectra. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, 2017.

(75) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 Residue Folded Peptide Designed by Segment Statistics. *Structure* **2004**, *12*, 1507−1518.

(76) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Biol.* **2002**, *9*, 425−430.

(77) Sevcik, J.; Solovicova, A.; Hostinova, E.; Gasperik, J.; Wilson, K. S.; Dauter, Z. Structure of glucoamylase from Saccharomycopsis fibuligera at 1.7 A resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 854−66.

(78) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.

DOI: 10.1021/acs.jctc.8b00524