THE JOURNAL OF PHYSICAL CHEMISTRY Letters

Letter

Cite This: J. Phys. Chem. Lett. 2018, 9, 4495−4501

pubs.acs.org/JPCL

Discovering a Transferable Charge Assignment Model Using Machine Learning

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

[†]Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90089, United States

[‡]Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States

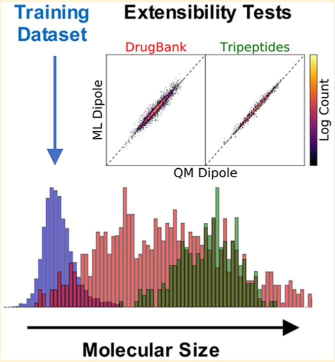
[¶]Center for Integrated Nanotechnologies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States

[§]Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States

[∥]UNC Eshelman School of Pharmacy, University of North Carolina Chapel Hill, Chapel Hill, North Carolina 27599, United States

S Supporting Information

ABSTRACT: Partial atomic charge assignment is of immense practical value to force field parametrization, molecular docking, and cheminformatics. Machine learning has emerged as a powerful tool for modeling chemistry at unprecedented computational speeds given accurate reference data. However, certain tasks, such as charge assignment, do not have a unique solution. Herein, we use a machine learning algorithm to discover a new charge assignment model by learning to replicate molecular dipole moments across a large, diverse set of nonequilibrium conformations of molecules containing C, H, N, and O atoms. The new model, called Affordable Charge Assignment (ACA), is computationally inexpensive and predicts dipoles of out-of-sample molecules accurately. Furthermore, dipole-inferred ACA charges are transferable to dipole and even quadrupole moments of much larger molecules than those used for training. We apply ACA to dynamical trajectories of biomolecules and produce their infrared spectra. Additionally, we find that ACA assigns similar charges to Charge Model 5 but with greatly reduced computational cost.



Electrostatic interactions contribute strongly to the forces within and between molecules. These interactions depend on the charge density field $\rho(r)$, which is computationally demanding to compute. Simplified models of the charge density, such as atom-centered monopoles, are commonly employed. These partial atomic charges result in faster computation as well as provide a qualitative understanding of the underlying chemistry.[1−4] However, the decomposition of charge density into atomic charges is, by itself, an ambiguous task. Additional principles are necessary to make the charge assignment task well-defined. Here we show that a machine learning model, trained *only* on the dipole moments of small molecules, discovers a charge model that is *transferable* to quadrupole predictions and *extensible* to much larger molecules.

Existing popular charge models have also been designed to reproduce observables of the electrostatic potential. The Merz−Singh−Kollman (MSK)[5,6] charge model exactly replicates the dipole moment and approximates the electrostatic potential on many points surrounding the molecule, resulting in high-quality electrostatic properties exterior to the molecule. However, MSK suffers from basis set sensitivity, particularly for "buried atoms" located inside of large molecules.[7−9] Charge Model 5 (CM5)[8] is an extension of Hirshfeld analysis,[10] with additional parametrization in order to approximately reproduce a combination of ab initio and experimental dipoles of 614 gas-phase dipoles. Unlike MSK, Hirshfeld and CM5 are nearly independent of basis set.[9] This insensitivity allows CM5 to use a single set of model parameters. The corresponding trade-off is that its charges do not reproduce electrostatic fields as well as MSK.

A limitation of these conventional charge models is that they require expensive ab initio calculation, which can be computationally impractical, especially for large molecules, long time scales, or systems exhibiting great chemical diversity. Recent advances in machine learning (ML) have demonstrated great potential to build quantum chemistry models with ab initio-level accuracy while bypassing ab initio costs.[11] Trained to reference data sets, ML models can predict energies, forces, and other molecular properties.[12−27] They have been used to discover materials[28−37] and study dynamical processes such as charge and exciton transfer.[38−41] Most related to this work are ML models of existing charge models,[9,42−44] which are orders of magnitude faster than ab initio calculation. Here we show that ML is able to go beyond emulation and *discover* a charge

Received: June 20, 2018
Accepted: July 24, 2018
Published: July 24, 2018

ACS Publications

DOI: 10.1021/acs.jpclett.8b01939

model that closely reproduces electrostatic properties by training directly to the dipole moment.

In this Letter, we use HIP-NN (Hierarchically Interacting Particle Neural Network),[45] a deep neural network for chemical property prediction, to train our charge model, called Affordable Charge Assignment (ACA). ACA is effective at predicting quadrupoles despite being trained only to dipoles, demonstrating the remarkable ability of ML to infer quantities not given in the training data set. Furthermore, its predictions are extensible to molecules much larger than those used for training. We validate ACA by comparing it to other popular charge models and find that it is similar to CM5. We then apply ACA to long-time dynamical trajectories of biomolecules and produce infrared spectra that agree very well with ab initio calculations.

We briefly review HIP-NN's structure. A more complete description is reported elsewhere in ref 45. HIP-NN takes a molecular conformation as input. The input representation consists of the atomic numbers of all atoms and the pairwise distances between atoms. This representation is simple and ensures that the network predictions satisfy translational, rotational, and reflection invariances. Figure 1 illustrates how HIP-NN processes molecules using a sequence of on-site and interaction layers. On-site layers generate information specific to each local atomic environment, and interaction layers allow sharing of information between nearby atomic environments.

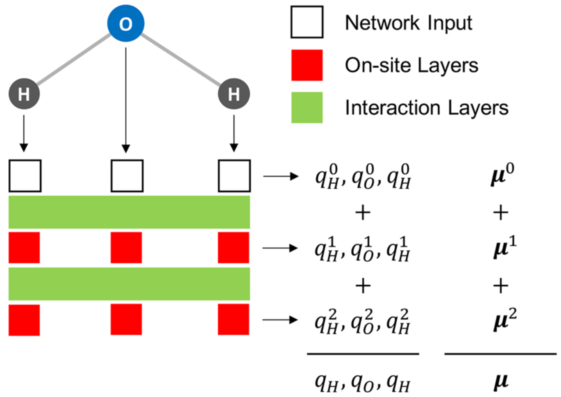


Figure 1. Abstract schematic of HIP-NN in the context of dipole prediction, illustrated for a water molecule.

HIP-NN has previously been successful in modeling energy[45] and pre-existing charge models.[9] In this work, we extend the model for dipole prediction using

$$\boldsymbol{\mu} = \sum_{i=1}^{N_{\text{atoms}}} q_i \mathbf{r}_i \tag{1}$$

where $\mathbf{r}_i$ and q_i are the position and charge of atom i. HIP-NN's learned charge assignment q_i (the ACA charge) is decomposed as a sum over hierarchical corrections

$$q_i = \sum_{l=0}^{N_{\text{interactions}}} q_i^l \tag{2}$$

As depicted in Figure 1, each q_i^l is calculated from the activations (i.e., outputs) of the lth set of HIP-NN on-site layers. An equivalent decomposition is $\boldsymbol{\mu} = \sum_l \boldsymbol{\mu}^l$ where $\boldsymbol{\mu}^l = \sum_i q_i^l \mathbf{r}_i^l$ is the lth hierarchical dipole correction. HIP-NN is designed such that higher-order corrections (i.e., $\boldsymbol{\mu}^l$ for larger l) tend to decay rapidly.

Training of HIP-NN proceeds by iterative optimization of the neural network model parameters using stochastic gradient descent. The goal of training is to maximize the accuracy of HIP-NN's dipole predictions (as quantified by the root-mean-square error (RMSE)) subject to regularization. The full ACA model of this Letter was generated by an ensemble of four networks. More details about HIP-NN and its training process are provided in ref 45 and the Supporting Information.

The HIP-NN training and testing data are drawn from the ANI-1x data set, which includes nonequilibrium conformations of molecules with C, H, N, and O atoms.[46] The ANI-1x data set was constructed through an active learning procedure[47−49] that aims to sample chemical space with maximum diversity. Although ANI-1x was originally designed for potential energy modeling, its chemical diversity also enhances the transferability of ML predictions for other properties, such as the dipole moment. We restrict molecule sizes to 30 atoms or less and randomly select 396k for training and 44k for testing. Data set calculations were performed with Gaussian 09 using the ωB97x density functional and 6-31G* basis set.[50] This level of theory will be referred to as the quantum-mechanical (QM) standard throughout this Letter.

We benchmark the ACA model according to the accuracy of its dipole and quadrupole predictions. To demonstrate extensibility, we test on the DrugBank (∼13k structures) and Tripeptides (2k structures) subsets of the COMP6 benchmark,[46] which contain nonequilibrium conformations of drug molecules and tripeptides. Figure 2 shows the molecular size distribution of these data sets; the molecules in the extensibility sets are roughly four times larger on average than those of ANI-1x, which we used to train ACA.

Figure 3 shows 2D histograms comparing ACA predicted dipoles and quadrupoles to the QM reference, for all three data

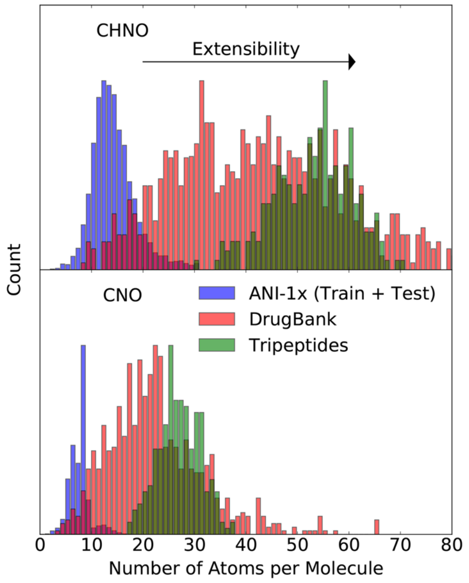


Figure 2. Size distributions of molecules in three data sets. The top panel counts the number of all atoms (C, H, N, O), and the bottom panel counts the number of heavy atoms (C, N, O) per molecule. Each histogram is normalized by its maximum bin count. Although ACA is only trained to ANI-1x, its predictions are extensible to the much larger molecules in the DrugBank and Tripeptides data sets.

DOI: 10.1021/acs.jpclett.8b01939

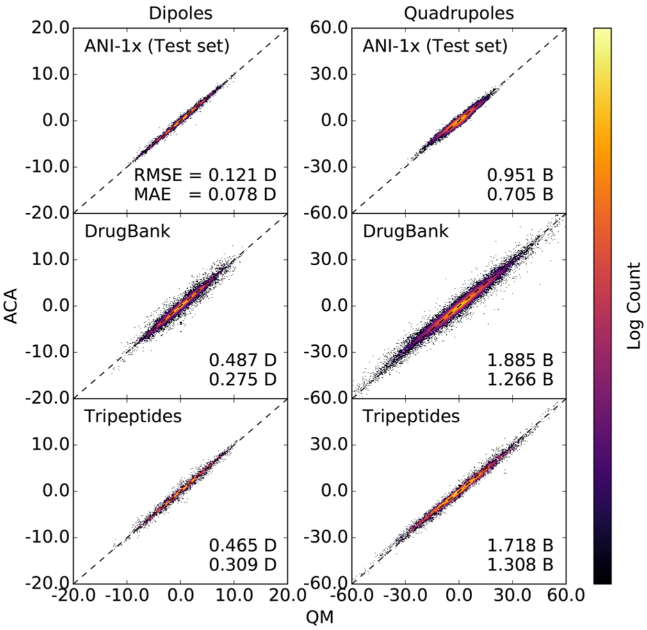


Figure 3. 2D histograms showing the correlation between predicted (ACA) and reference (QM) electrostatic moments using three test data sets: ANI-1x, DrugBank, and Tripeptides. Left and right panels show dipole and quadrupole correlations, respectively. The values for the RMSE and MAE are provided in the lower right corner of each subpanel. The color scheme for each histogram is normalized by its maximum bin count. ACA is surprisingly effective in predicting quadrupoles, given that it was only trained to ANI-1x dipoles.

sets. We measure the RMSE and mean absolute error (MAE). Left panels of Figure 3 compare Cartesian dipole components in units of Debye (D). The MAE of 0.078 D for predicting ANI-1x dipoles is comparable to the error between the QM level of theory and experimental dipole measurements.[51] The MAE of ~0.3 D for predicting DrugBank and Tripeptides dipoles demonstrates the strong extensibility of ACA. Right panels of Figure 3 compare quadrupole Cartesian components in units of Buckingham (B). The agreement with QM is remarkable (MAE = 0.705 B for the ANI-1x tests) in light of the fact that ACA was trained only to dipoles. Furthermore, ACA continues to make good quadrupole predictions for the much larger COMP6 molecules. We conclude that the ACA charges are physically useful for reproducing electrostatic quantities. Additional material quantifying the distributions depicted in Figures 2 and 3, including the error as a function of molecular size, are available in the Supporting Information.

Next, we compare the dipole-inferred ACA model to some conventional charge models. This analysis uses a subset of GDB-11, denoted here as GDB-5, which contains up to five heavy atoms of types C, N, and O.[52] The data set contains a total of 517 133 structures, including nonequilibrium conformations. Four charge models were included in the reference data set: Hirshfeld,[10] MSK,[5,6] CM5,[8] and population analysis from natural bond orbitals[53] (NBOs). Hirshfeld assigns atomic contributions to the electron density based on their relative weighting to the protodensity. MSK charges are constrained to reproduce the dipole moment while attempting to match the electrostatic potential at many points surrounding the molecule. CM5 is an extension of Hirshfeld, empirically parametrized to reproduce ab initio and experimental dipoles. NBO charges are computed as a sum of occupancies from all natural atomic orbitals on each atom. The NBO model is more popular for capturing features such as bond character.

Figure 4 shows the correlation between each pair of charge models and demonstrates the inconsistency between different approaches for charge partitioning. The strongest correspondence is between CM5 and ACA, with a mean absolute

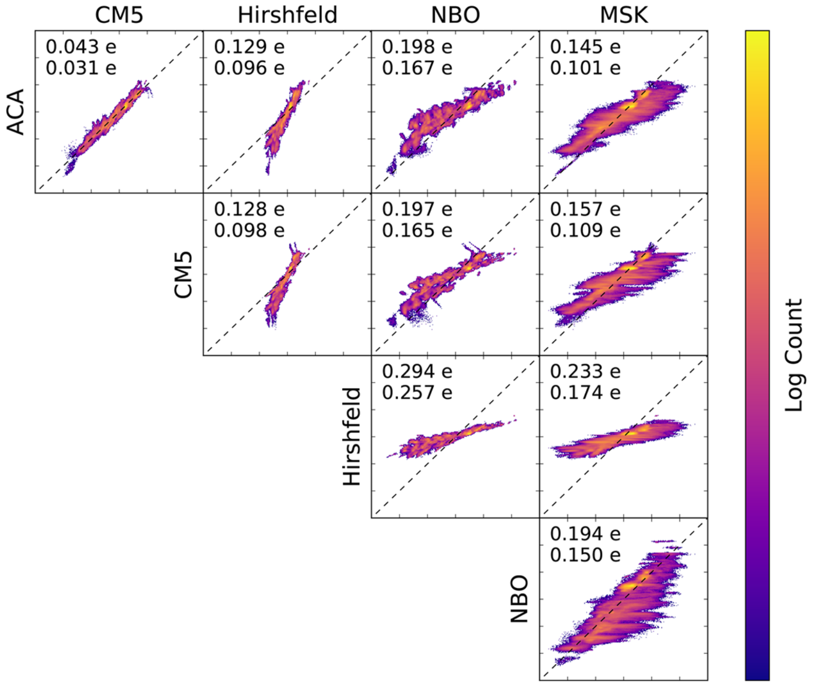


Figure 4. 2D histograms showing correlations between all pairs of charge models. The upper and lower values in each subpanel are the root-mean-square deviation and mean absolute deviation, respectively. The color scheme for each histogram is normalized by its maximum bin count. The strong agreement between ACA and CM5 charge assignments was unexpected.

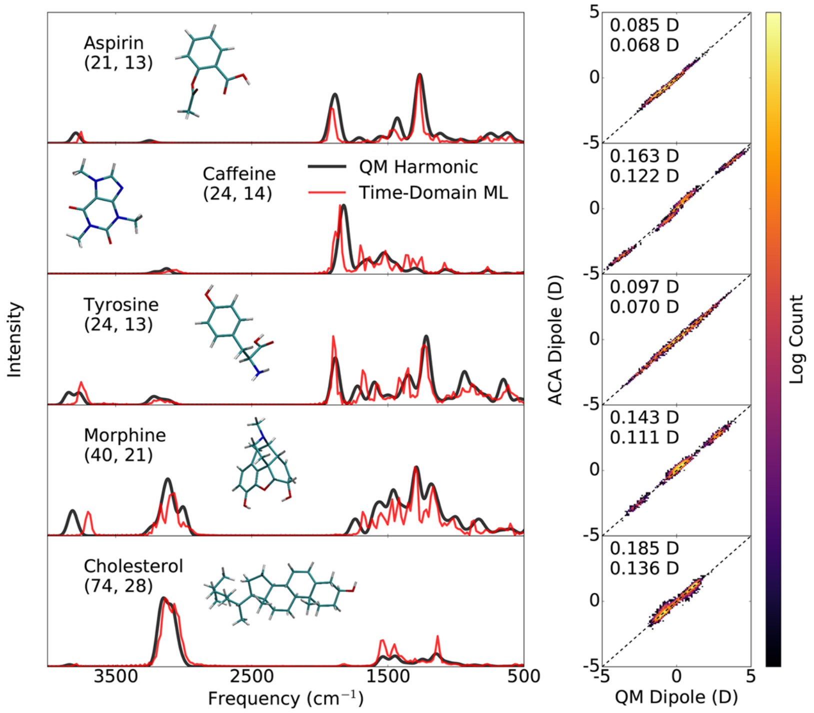


Figure 5. (Left) Infrared spectra of select molecules, computed without polarization effects due to solvation. The values in parentheses are the total number of all atoms (C, H, N, O) and of heavy atoms (C, N, O). The agreement between QM- and ACA-derived spectra is reasonable, given that the harmonic approximation is not exact. (Right) 2D histograms of predicted (ACA) versus true (QM) dipoles at 10^3 subsampled time steps throughout the 100 ps trajectories. The upper and lower values in each subpanel are the RMSE and MAE, respectively. The color scheme for each histogram is normalized by its maximum bin count.

deviation of 0.031 e. Other model pairs have mean absolute deviations that range from 3 to 8 times larger, a consequence of differing principles used to design these models.

Conceptually, MSK, CM5, and ACA are similar in that they attempt to partition charge such that the molecular dipole moment is preserved in the point charge representation. We note, however, that MSK differs significantly from CM5 and ACA (Figure 4). MSK is constrained to match the QM dipole *exactly* for each given input molecular configuration. This constraint alone is underdetermined, and MSK therefore invokes additional principles for its charge assignment, attempting to fit the far-field electrostatic potential. However, the far-field potential is relatively insensitive to the partial charge assignments of internal atoms.[7−9] Because MSK performs its charge assignments according to global (rather than local) criteria, the assigned charges can deviate significantly from the local charge density field. Another related difficulty of MSK is that it exhibits a noticeable basis set dependence.[7,9]

CM5 was designed to address such drawbacks.[8] Like CM5, our ACA charge model is local-by-design, thus averting the problem of artificial long-range effects. Specifically, ACA seeks a *local* charge assignment model that best reproduces the QM dipoles over the whole training data set. We remark that the ACA dipole predictions do not perfectly reproduce the QM dipoles. Allowing for this imperfection may actually be important; collapsing a charge density field into a relatively small number of monopoles while simultaneously forcing the molecular dipole to be exact may be incompatible with locality of the charge model.

As we show in Figure 4, the CM5 and ACA charges are remarkably consistent, a result that we did not anticipate. CM5 reproduces the molecular dipole well but not as accurately as ACA (see the Supporting Information). The reduced accuracy of CM5 dipoles may be due to the fact that it is fit to a hybrid of ab initio and experimental data. In contrast, ACA trains to a homogeneous database of QM dipoles. The ML approach has a conceptual advantage: it is fully automated and requires few design decisions (primarily, the specification of an error metric for training). As a consequence, the extension of ACA to new atomic species and to new classes of molecules should be straightforward.

A strong practical advantage of ACA is that assignment does not require any new QM calculations. We highlight this advantage of efficiency by applying ACA to calculate an experimentally relevant quantity. Inspired by the work of ref 26, we use ACA to calculate dynamic dipoles and subsequently infrared spectra for select molecules. Ground-state trajectories were generated from the ANI-1x potential[46] and were 100 ps in length with a 0.1 fs time step, amounting to a total of 10^6 time steps. Dipoles were predicted along these trajectories using ACA. Both the molecular dynamics and dipole prediction were performed using only ML, that is, without any QM calculation. Spectra were made by Fourier transforming the dipole moment autocorrelation function. Harmonic spectra were calculated with the Gaussian 09 software. A comparison of time domain ML spectra to QM harmonic spectra is shown in Figure 5, left panels. Although time domain and harmonic spectra are not one-to-one, the comparison is reasonable because spectral features are harmonic to first order. ACA recovers the harmonic features across all molecules.

To further validate the ACA dipole predictions, QM calculations were performed at 10^3 subsampled time steps throughout the trajectories. Figure 5, right panels, shows that the ACA dipole predictions are in excellent agreement with QM, another validation of ACA's extensibility. The dipole errors are consistent with those observed in the data sets of Figure 3. Note that cholesterol and morphine have 74 and 40 atoms, respectively, whereas our training data set has no molecules with more than 30 atoms. The quality of the ML-predicted spectra for cholesterol and morphine is similar to those of smaller molecules, such as aspirin.

We carried out an additional test with smaller molecules of sizes 6−15 atoms, making it feasible to calculate QM dipoles at all 10^6 time steps. The resulting infrared spectra are shown in the Supporting Information and are in excellent agreement with our ML-based approach. For these smaller molecules, ACA yields a factor of greater than 10^4 computational speed-up. The relative speed-up is even more dramatic for large molecules.

In summary, the key contribution of this Letter is the formulation of an electrostatically consistent charge model called ACA. We construct the ACA model using a deep neural network that outputs charges. The network is trained to DFT-computed molecular dipole moments over a diverse set of chemical structures. The fast and accurate predictive power of the model was evidenced with extensibility tests (Figure 3) and infrared spectra (Figure 5). Although ACA is only trained directly to the molecular dipole, we show that it also captures quadrupole moments, demonstrating transferability.

ACA is compared with four conventional charge models on a data set containing over 500k molecules (Figure 4). The rather poor correlation between most model pairs confirms the ambiguity in charge partitioning. The ACA model correlates well to CM5. CM5 was designed to combine advantages of the Hirshfeld and MSK models. It is parametrized to reproduce a combination of ab initio and experimental dipoles. ACA, like CM5, is a local model that is designed to reproduce dipoles but, unlike CM5, is built entirely from ab initio data. In addition to fast charge assignments, a potential advantage of ACA is its applicability to a wide range of chemically diverse systems, assuming that appropriate training data is available. This work is also a testament to how physics-informed ML can be used to discover properties (here, charge assignment) not employed as an explicit target in the training process. We would also like to note an independent and concurrent study (i.e., ref 54) that took a similar approach in constructing dipole-driven partial charges. The authors of ref 54 confirm that inferred charges produce interpretable insight into chemical structure.

Future work will focus on improving and utilizing ACA for quantum chemical prediction. Improvements to extensible dipole prediction may be made by engaging in dipole-driven active learning. Furthermore, ACA could be trained to higher-order multipole moments such as quadrupoles; this could be important for systems where the dipole does not provide enough of a constraint for charge assignments. Currently, ACA is limited to C, H, N, and O atoms, but this could be overcome when more diverse data sets are available. Another important drawback of the current model is that charged systems, such as anionic and cationic species, cannot yet be treated. An application underway is to predict dynamic charges in neutral biomolecular systems to parametrize force fields for molecular dynamics. We hope that this study not only sheds light on charge models that best reproduce dynamical data but also helps explain the interesting correlations among charge models (visible in Figure 4).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpclett.8b01939.

> Additional details on ACA training and charge assignment, correlation plots of ACA charge predictions between different neural networks, table summarizing test and extensibility data sets along with statistical measures of dipole and quadrupole prediction, error in dipole prediction as a function of the number of atoms in ANI-1x, DrugBank, and Tripeptides data sets, correlation plots between predicted and reference electrostatic moments (i.e., dipoles and quadrupoles) using several popular charge models, and infrared spectra and dipole correlations of small molecules ranging from 6 to 15 total atoms (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: kbarros@lanl.gov (K.B.).

*E-mail: serg@lanl.gov (S.T.).

ORCID

Andrew E. Sifain: 0000-0002-2964-1923

Adrian E. Roitberg: 0000-0003-3963-8784

Sergei Tretiak: 0000-0001-5547-3647

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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. 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 the Advanced Computing Laboratory (ACL) for providing computational resources. A.E.S. acknowledges support of the U.S. Department of Energy, Grant No. DE-SC0014429. A.E.S., J.S.S., and O.I. thank CNLS for their support and hospitality.

■ REFERENCES

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

(2) 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.

(3) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Jr 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.

(4) Provorse, M. R.; Peev, T.; Xiong, C.; Isborn, C. M. Convergence of Excitation Energies in Mixed Quantum and Classical Solvent: Comparison of Continuum and Point Charge Models. *J. Phys. Chem. B* **2016**, *120*, 12148−12159.

(5) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.

(6) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(7) Sigfridsson, E.; Ryde, U. Comparison of Methods for Deriving Atomic Charges from the Electrostatic Potential and Moments. *J. Comput. Chem.* **1998**, *19*, 377−395.

(8) 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.

(9) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A.; Lokhov, A.; Isayev, O.; Roitberg, A.; Barros, K.; Tretiak, S. *Transferable Molecular Charge Assignment Using Deep Neural Networks*. https://arxiv.org/abs/1803.04395 (2018).

(10) Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chem. Acc.* **1977**, *44*, 129−138.

(11) Behler, J. Constructing High-Dimensional Neural Network Potentials: A Tutorial Review. *Int. J. Quantum Chem.* **2015**, *115*, 1032−1050.

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

(13) 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.

(14) 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.

(15) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.

(16) 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.

(17) 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.

(18) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309−3313.

(19) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-The-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.

(20) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, e1603015.

(21) Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8*, 2689−2694.

(22) 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.

(23) 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.

(24) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 1668−1673.

(25) 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.

(26) 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.

(27) 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.

(28) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.

(29) 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.

(30) Morawietz, T.; Behler, J. A Density-Functional Theory-Based Neural Network Potential for Water Clusters Including van der Waals Corrections. *J. Phys. Chem. A* **2013**, *117*, 7356−7366.

(31) Artrith, N.; Hiller, B.; Behler, J. Neural Network Potentials for Metals and Oxides-First Applications to Copper Clusters at Zinc Oxide. *Phys. Status Solidi B* **2013**, *250*, 1191−1203.

(32) 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.

(33) Kondati Natarajan, S.; Morawietz, T.; Behler, J. Representing the Potential-Energy Surface of Protonated Water Clusters by High-Dimensional Neural Network Potentials. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8356−8371.

(34) Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8*, 5137−5152.

(35) 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.

(36) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064−1071.

(37) Tong, Q.; Xue, L.; Lv, J.; Wang, Y.; Ma, Y. Accelerating CALYPSO Structure Prediction by Data-Driven Learning of Potential Energy Surface. *Faraday Discuss.* **2018**, DOI: 10.1039/C8FD00055G.

(38) Häse, F.; Valleau, S.; Pyzer-Knapp, E.; Aspuru-Guzik, A. Machine Learning Exciton Dynamics. *Chem. Sci.* **2016**, *7*, 5139−5147.

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

(40) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.

(41) 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.

(42) Geidl, S.; Bouchal, T.; Raček, T.; Svobodová Vařeková, R.; Hejret, V.; Křenek, A.; Abagyan, R.; Koča, J. High-Quality and Universal Empirical Atomic Charges for Chemoinformatics Applications. *J. Cheminform.* **2015**, *7*, 59.

(43) Bereau, T.; DiStasio, R. A., Jr; 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.

(44) 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.

(45) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.

(46) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(47) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discovery Today* **2015**, *20*, 458−465.

(48) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.

(49) Podryabinkin, E. V.; Shapeev, A. V. Active Learning of Linearly Parametrized Interatomic Potentials. *Comput. Mater. Sci.* **2017**, *140*, 171−180.

(50) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; et al. *Gaussian 09*, revision D. 01; 2009.

(51) Hickey, A. L.; Rowley, C. N. Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *J. Phys. Chem. A* **2014**, *118*, 3678−3687.

(52) 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.

(53) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural Population Analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.

(54) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R. *Quantum-Chemical Insights from Interpretable Atomistic Neural Networks*. https://arxiv.org/abs/1806.10349 (2018).