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 Califronia, Los Angeles, CA 90089*

‡*Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545*

¶*Center for Integrated Nanotechnologies, Los Alamos National Laboratory, Los Alamos, NM 87545*

§*Department of Chemistry, University of Florida, Gainesville, FL 32611*

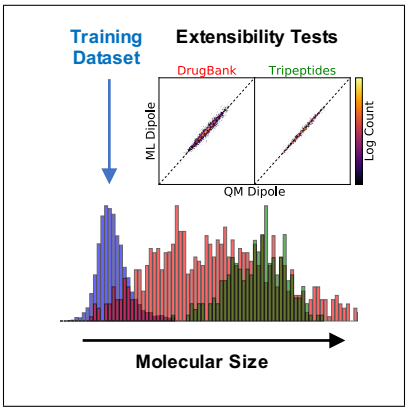
‖*UNC Eshelman School of Pharmacy, University of North Carolina Chapel Hill, Chapel Hill, North Carolina 27599*

E-mail: kbarros@lanl.gov; serg@lanl.gov

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 ground-truth values, but for the task of charge assignment, the choice of ground-truth may not be obvious. In this letter, we use machine learning to *discover* a charge model by training a neural network to molecular dipole moments using a large, diverse set of CHNO molecular conformations. 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 long dynamical trajectories of biomolecules and successfully produce their infrared spectra. Additionally, we compare ACA with existing charge models and find that ACA assigns similar charges to Charge Model 5, but with a greatly reduced computational cost.

Graphical TOC Entry



Keywords

machine learning, neural networks, quantum chemisty

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 large molecules.[7–9] Charge model 5 (CM5)[8] is an extension of Hirshfeld analysis,[10] with additional parametrization in order to approximately reproduce *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 tradeoff 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 datasets, 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 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 Assignments (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 dataset. 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.

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

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

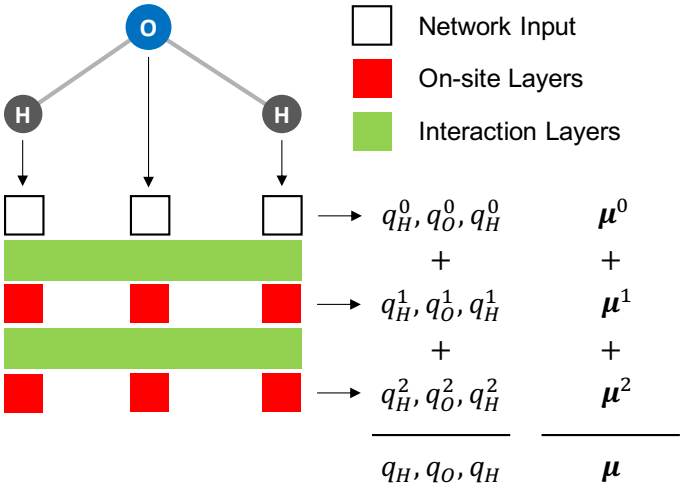


Figure 1: Abstract schematic of HIP-NN in the context of dipole prediction, illustrated for a water molecule.

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,

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

As depicted in Fig. 1, each q_i^ℓ is calculated from the activations (i.e. outputs) of the ℓ-th set of HIP-NN on-site layers. An equivalent decomposition is $\boldsymbol{\mu} = \sum_\ell \boldsymbol{\mu}^\ell$ where $\boldsymbol{\mu}^\ell = \sum_i q_i^\ell \mathbf{r}_i^\ell$ is the ℓ-th hierarchical dipole correction. HIP-NN is designed such that higher-order corrections (i.e. $\boldsymbol{\mu}^\ell$ 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) subject to regularization. The full ACA model of this paper was generated by an ensemble of four networks. More details about HIP-NN and its training process are provided in Ref. [45] and Supporting Information.

The HIP-NN training and testing data are drawn from the ANI-1x dataset, which includes non-equilibrium conformations of molecules with C, H, N, and O.[46] The ANI-1x dataset 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. Dataset 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 paper.

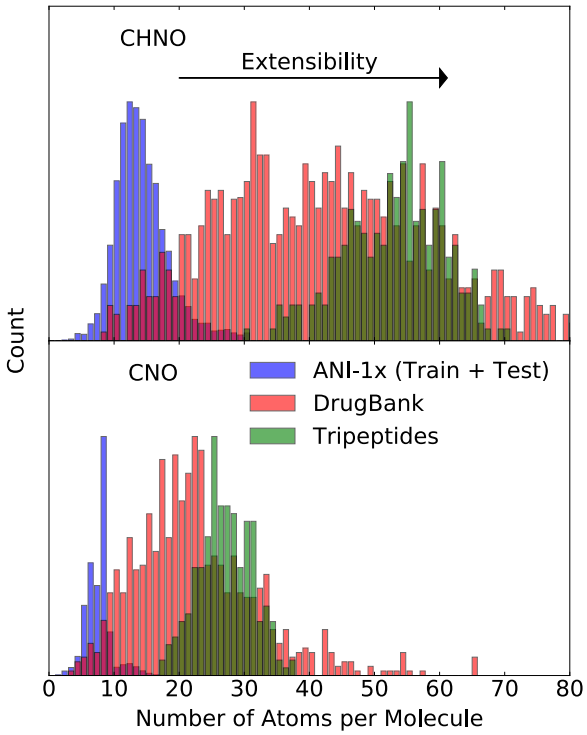


Figure 2: Size distributions of molecules in three datasets. Top panel counts the number of all atoms (C, H, N, O) and bottom panel counts the number of heavy atoms (C, N, O), per molecule. Each histogram is normalized by its maximum bin count. Although ACA is only trained to ANI-1x, its predictions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

We benchmark the ACA model according to the accuracy of its dipole and quadrupole predictions. To demonstrate extensibility, we test on the DrugBank ($\sim$ 13k structures) and Tripeptides (2k structures) subsets of the COMP6 benchmark,[46] which contain non-

equilibrium conformations of drug molecules and tripeptides. Figure 2 shows the molecular size distribution of these datasets; 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 datasets. We measure the root-mean-square-error (RMSE) and mean-absolute-error (MAE). Left panels of Fig. 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 Fig. 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 Figs. 2 and 3, including error as a function of molecular size, are available in 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 5 heavy atoms of types C, N, and O.[52] The dataset contains a total of 517,133 structures, including non-equilibrium conformations. Four charge models were included in the reference dataset: Hirshfeld,[10] MSK,[5,6] CM5,[8] and population analysis from natural bond orbitals[53] (NBO). Hirshfeld assigns atomic contributions to the electron density based on their relative weighting to the proto-density. 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

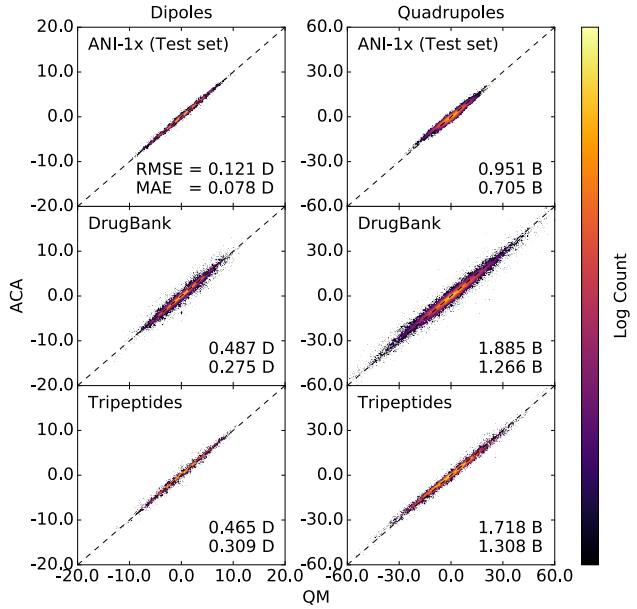


Figure 3: 2D histograms showing the correlation between predicted (ACA) and reference (QM) electrostatic moments using three test datasets: ANI-1x, DrugBank, and Tripeptides. Left and right panels show dipole and quadrupole correlations, respectively. The upper and lower values in each subpanel are RMSE and MAE, respectively. 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.

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 to charge partitioning. The strongest correspondence is between CM5 and ACA, with a mean-absolute-deviation of 0.031 e. Other model pairs have mean-absolute-deviations that range from three to eight 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 (Fig. 4). MSK is constrained to match the QM dipole *exactly* for each given input molecular configuration. This constraint alone is under-determined, and so MSK therefore invokes additional principles for its charge assignment, attempting to fit the far-field electrostatic potential. However,

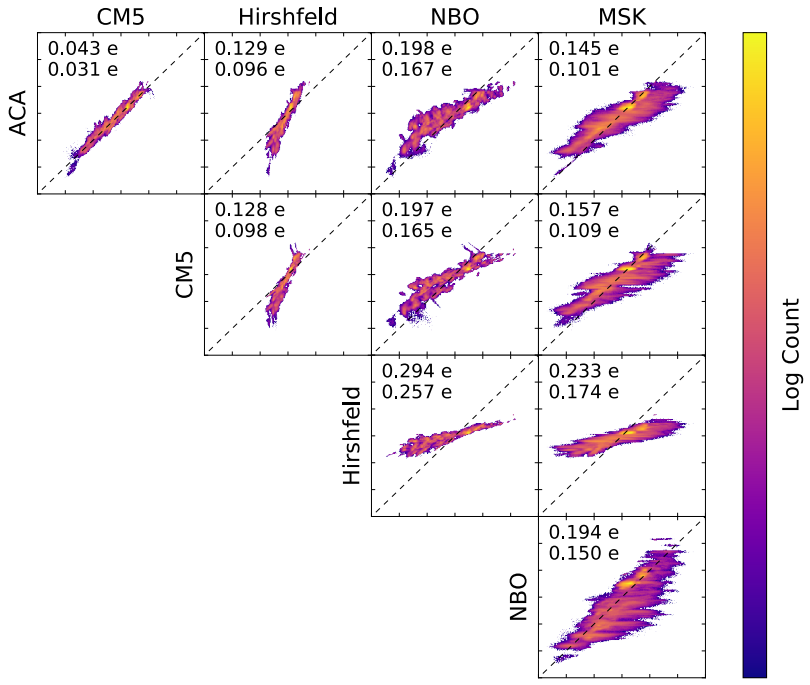


Figure 4: 2D histograms showing correlations between all pairs of charge models. The upper and lower values in each subpanel are root-mean-square-deviation and mean-absolute-deviation, respectively. The strong agreement between ACA and CM5 charge assignments was unexpected.

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 dataset. 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.

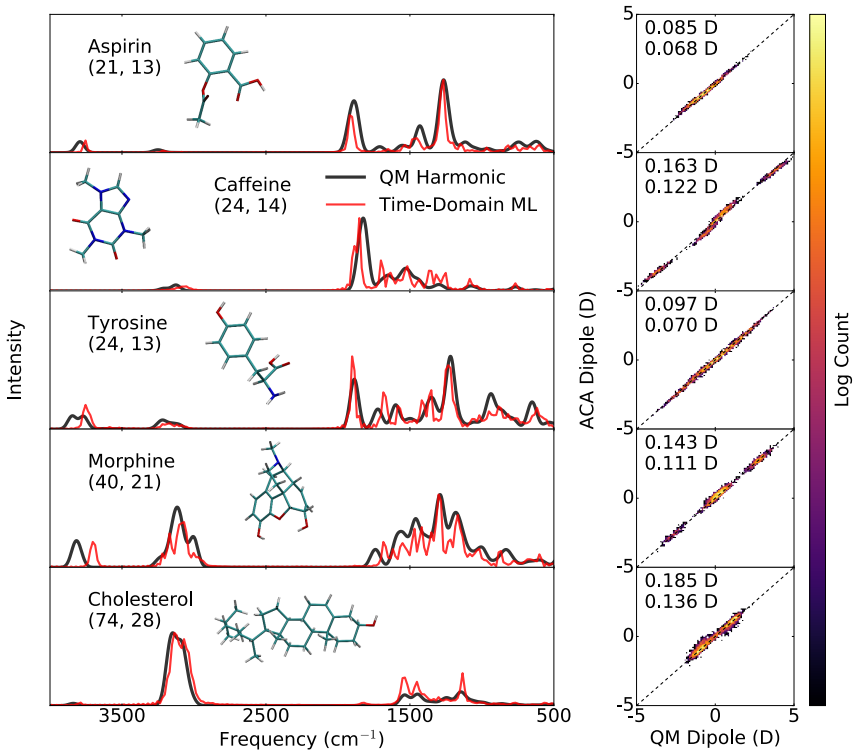


Figure 5: (Left) Infrared spectra of select molecules, computed without polarization effects due to solvation. The values in parentheses are the total number of all atoms (C, H, N, O) and of heavy atoms (C, N, O), respectively. The agreement between QM and ACA-derived spectra is reasonable, given that 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 RMSE and MAE, respectively.

As we showed in Fig. 4, the CM5 and ACA charges are remarkably consistent, a result we did not anticipate. CM5 reproduces the molecular dipole well, but not as accurately as ACA (See Supporting Information). The reduced accuracy of CM5 dipoles may be due to the fact that it is a 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 efficiency advantage 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, i.e., 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 Figure 5, left panels. Although time-domain and harmonic spectra are not one-to-one, the comparison is reasonable since 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. Fig. 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 datasets of Fig. 3. Note that cholesterol and morphine have 74 and 40 atoms, respectively, whereas our training dataset 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 to 15 atoms, using QM to calculate dipoles at all 10^6 time-steps. The resulting infrared spectra are shown in 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 results are even more dramatic for large molecules.

In summary, the key contribution of this paper is the formulation of an electrostatically consistent charge model called Affordable Charge Assignments (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 (Fig. 3) and infrared spectra (Fig. 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 dataset containing over 500k molecules (Fig. 4). The rather poor correlation between most model pairs confirms the ambiguity in charge partitioning. The ACA model correlates well to Charge Model 5 (CM5). CM5 was designed to combine advantages of the Hirshfeld and MSK models. It is parameterized 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.

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 CHNO atoms, but this could be overcome when more diverse datasets 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 using ACA is underway to predict dynamic charges in neutral biomolecular systems to parametrize force fields for molecular dynamics.

Supporting Information Available

More details on ACA training and charge assignment. Correlation plots of ACA charge predictions between different neural networks. Table summarizing test and extensibility datasets along with statistical measures of dipole and quadrupole prediction. Error in dipole prediction as a function of number of atoms in the following test datasets: test set of ANI-1x, DrugBank, and Tripeptides. Correlation plots between predicted and reference electrostatic moments (i.e. dipoles and quadrupoles) using several popular charge models: ACA, Hirshfeld, MSK, CM5, and NBO. Infrared spectra and dipole correlations of small molecules ranging from 6 to 15 total atoms.

Acknowledgement

The authors acknowledge support of the US 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 US 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. AES acknowledges support of the US Department of Energy, Grant No. DE-SC0014429. AES, JSS, and OI 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 and Comput.* **2011**, *7*, 4026–4037.

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

(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 and 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. *arXiv: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.; Von Lilienfeld, O. A. 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 and 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* **2015**, *92*, 014106.

(33) Natarajan, S. K.; 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**,

(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.; Vařeková, R. S.; 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 Jr, 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.

(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 Discov. 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.

Supporting Information: 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 Califronia, Los Angeles, CA 90089*

‡*Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545*

¶*Center for Integrated Nanotechnologies, Los Alamos National Laboratory, Los Alamos, NM 87545*

§*Department of Chemistry, University of Florida, Gainesville, FL 32611*

∥*UNC Eshelman School of Pharmacy, University of North Carolina Chapel Hill, Chapel Hill, North Carolina 27599*

E-mail: kbarros@lanl.gov; serg@lanl.gov

HIP-NN Architecture and Training Details

HIP-NN Architecture

The HIP-NN model closely follows the methodology given in Ref. [1]. A key difference is that linear layers are used to construct partial atomic charge, rather than a molecular energy, and so no sum over atoms is employed. The network has 2 interaction blocks, each consisting of 1 interaction layer, followed by 3 on-site layers, and a linear layer to form hierarchical contribution to charge. Each layer was given a width of 40 neurons. The network architecture contains approximately 60k parameters.

Training

Training also closely follows Ref. [1]. The main difference is the cost function, adapted for dipole regression. The cost function used here consists of dipole RMSE, total charge RMSE, and L2 regularization (as described in Ref. [1]):

$$\mathcal{L} = \sqrt{\frac{1}{3}\langle(\boldsymbol{\mu}' - \boldsymbol{\mu})^2\rangle} + \sqrt{\langle Q'^2\rangle} + \mathcal{L}_{L2} \tag{1}$$

where the angle brackets $\langle ... \rangle$ denote a quantity averaged over each training batch of 30 molecules, $\boldsymbol{\mu}'$ and $\boldsymbol{\mu}$ represent the predicted and QM dipole, respectively, and Q' represents the predicted total charge for the molecule (i.e. the total QM charge is set to zero). The factor of $\frac{1}{3}$ is a normalization reflecting the three cartesian degrees of freedom in the dipole.

Training is then given by the gradient-based optimization and annealing/early-stopping algorithm in Ref. [1]. A validation set of 1% of the training dataset (approximately 4385 molecules) was used for the annealing procedure, and the dipole RMSE was used as the validation criterion for annealing. For training to the ANI-1x dataset used in this work, the algorithm terminates after roughly 1000 epochs.

Details of Charge Assignment

The full charge assignments are given by an ensemble prediction using four different random initializations of HIP-NN, each separately trained to the same data. Figure S1 shows the correlation between charge predictions by the members of the ensemble; networks agree to approximately 0.01 e (Fig. S1).The charges produced by the ensemble are not exactly neutral, and so when predicting the charge on a molecule, excess total charge is redistributed evenly across atoms. This redistribution constitutes a very small change, typically 0.001 e or less per atom.

Dipoles for each datapoint $\boldsymbol{\mu}$ are constructed as

$$\boldsymbol{\mu} = \sum_{i=1}^{N_{\text{atoms}}} \mathbf{r}_i \, q_i \tag{2}$$

and traceless quadrupoles are constructed as

$$\boldsymbol{Q} = \sum_{i=1}^{N_{\text{atoms}}} \left(\mathbf{r}_i \otimes \mathbf{r}_i - \frac{1}{3} \mathbf{I} \, r_i^2 \right) q_i \tag{3}$$

where $\otimes$ is the outer product, and $\mathbf{I}$ is the unit dyadic (or Kronecker delta).

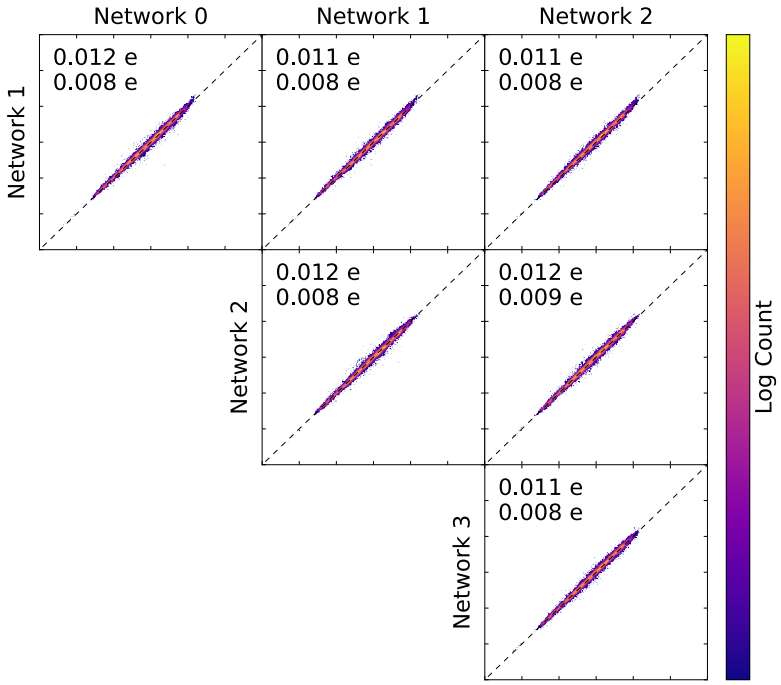


Figure S1: Pair correlation plots of charge predictions from the four neural networks constituting our ensemble, labeled as $0, 1, 2, 3$. Upper and lower values in each subpanel are RMSD and MAD, respectively.

Additional Data

This section contains additional data quantifying the performance of ACA. We include a a table summarizing extensibility results (Table S1), bar charts showing error as a function of molecule size (Figures S2, S3, and S4), dipole and quadrupole correlations plots comparing ACA to other existing popular charge models (Figures S5 and S6), and infrared spectra computed with ML dynamics + ACA, ML dynamics + QM dipoles, and harmonic QM (Figure S7).

Table S1: Summary of test and extensibility datasets along with statistical measures for dipole and quadrupole prediction.

	ANI-1x	Drug Bank	Tripeptides
Total # molecules	438481[a]	13379	2000
Total atoms (CHNO) per molecule, min / mean / max	2 / 14 / 30	8 / 44 / 140	30 / 53 / 70
Heavy atoms (CNO) per molecule, min / mean / max	1 / 7 / 17	3 / 22 / 65	17 / 27/ 37
Dipole MAE, RMSE (D)	0.08[b], 0.12[b]	0.28, 0.49	0.31, 0.47
Quadrupole MAE, RMSE (B)	0.71[b], 0.95[b]	1.27, 1.89	1.31, 1.72
Mean $\left\|\boldsymbol{\mu}_{\mathrm{ACA}} - \boldsymbol{\mu}_{\mathrm{QM}}\right\|$ (D)	0.16[b]	0.59	0.66

[a] This is 10% of the full ANI-1x dataset, which consists of more than 4M molecules.

[b] Error metrics computed on held-out test set of 43849 molecules.

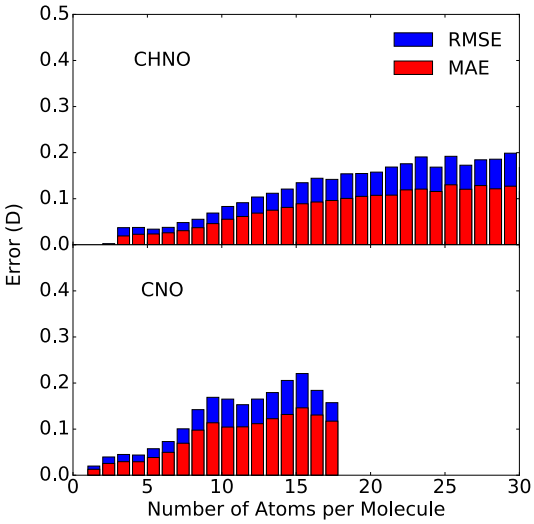


Figure S2: Bar charts showing RMSE and MAE for each molecule size in the 43849 test datapoints selected from ANI-1x. Top and bottom panels correspond to total atoms (CHNO) and heavy atoms (CNO), respectively.

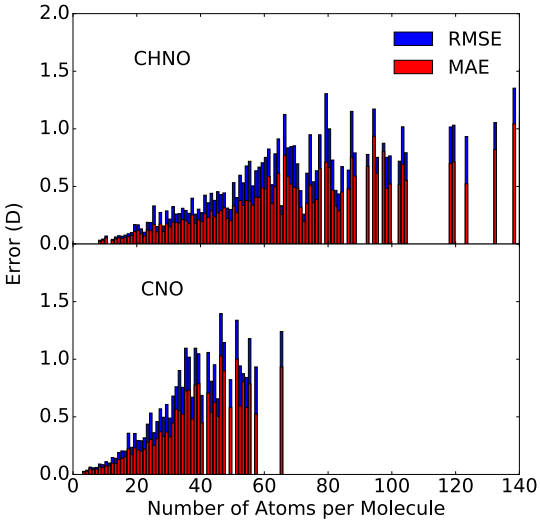


Figure S3: Bar charts showing RMSE and MAE for each molecule size for the DrugBank extensibility set. Top and bottom panels correspond to total atoms (CHNO) and heavy atoms (CNO), respectively.

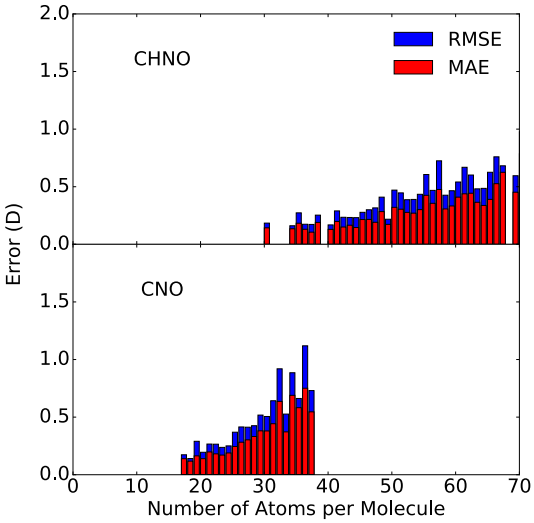


Figure S4: Bar charts showing RMSE and MAE for each molecule size for the Tripeptide extensibility set. Top and bottom panels correspond to total atoms (CHNO) and heavy atoms (CNO), respectively.

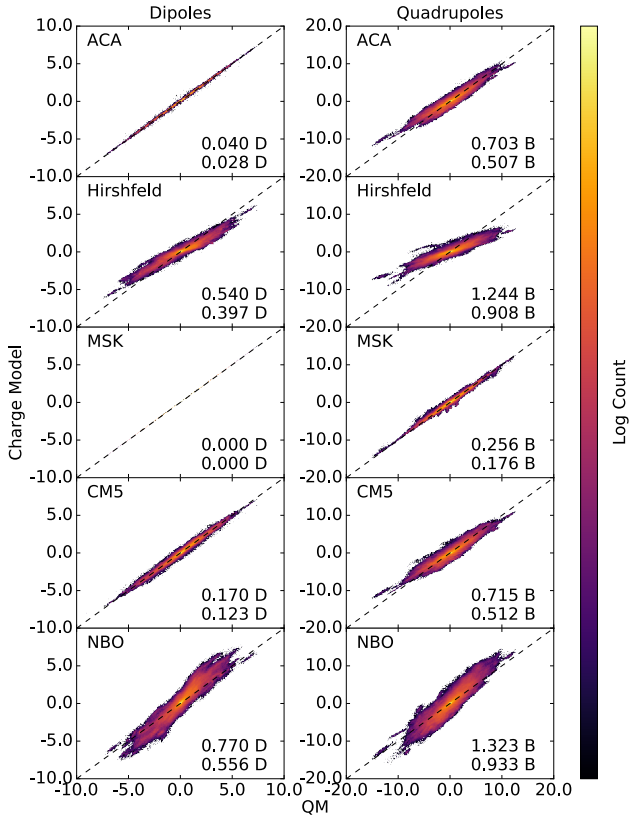


Figure S5: 2D histograms showing correlations between predicted (Charge Model) and reference (QM) electrostatic moments using five different charge models: ACA, Hirshfeld, MSK, CM5, and NBO. The test dataset is GDB-5, which contains a total of 517,133 molecules. The upper and lower values in each subpanel are RMSE and MAE, respectively. Each histogram is normalized by its maximum bin count. ACA recovers both the dipole and quadupole moments of the test dataset at better accuracy than all other models except for MSK. ACA recovers quadrupoles, despite being only trained to dipoles.

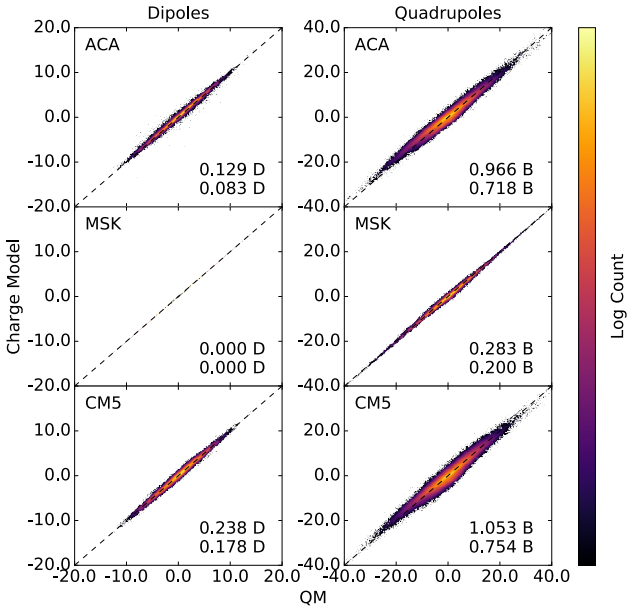


Figure S6: 2D histograms showing correlations between predicted (Charge Model) and reference (QM) electrostatic moments using three different charge models: ACA, CM5, and MSK. The test dataset is a random subset of ANI-1x, which contains a total of BLAH molecules. The upper and lower values in each subpanel are RMSE and MAE, respectively. Each histogram is normalized by its maximum bin count. ACA recovers both the dipole and quadupole moments of the test dataset at better accuracy than CM5. ACA recovers quadrupoles, despite being only trained to dipoles.

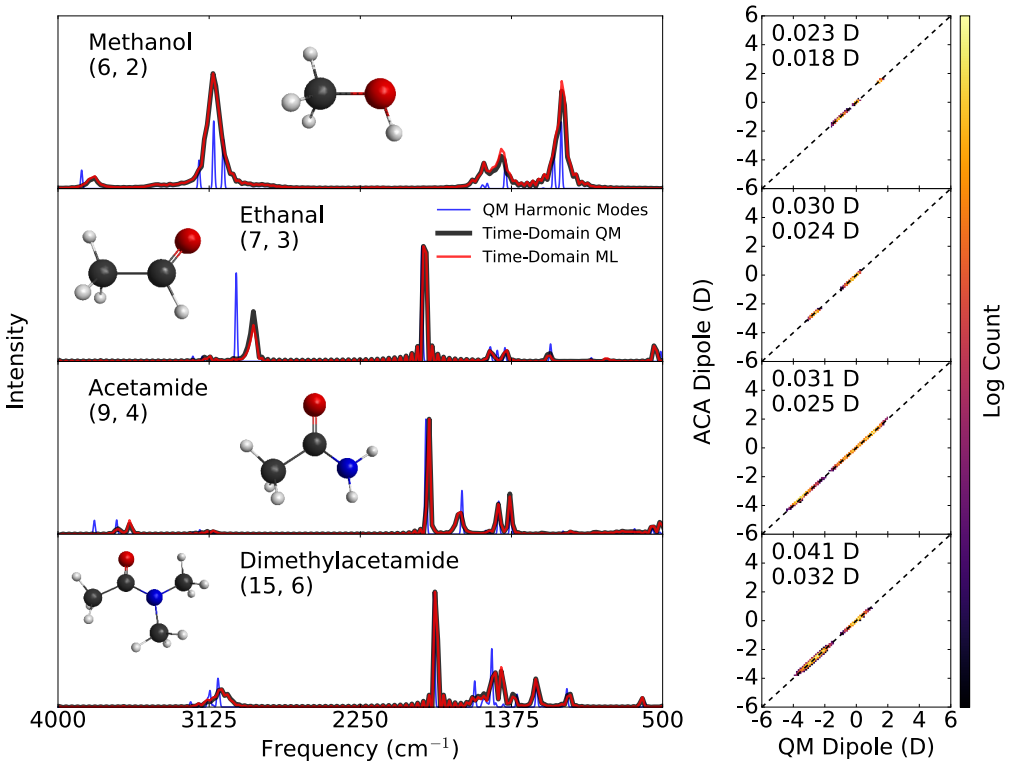


Figure S7: Infrared spectra of small molecules calculated using ACA (red) and QM (black). The values in parentheses are the total number of all atoms (C, H, N, O) and heavy atoms (C, N, O), respectively. For each molecule, both ACA and QM dipoles were predicted at 10^6 time-steps of the same ANI-1x molecular dynamics trajectory. Frequencies determined from DFT harmonic mode analysis are also shown (blue). Right panels are dipole correlation plots of ACA versus QM. Upper and lower values in each subpanel are RMSE and MAE, respectively.

References

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