Transferable Neural Network Potential Energy Surfaces for Closed-Shell Organic Molecules: Extension to Ions

Leif Jacobson,* James Stevenson, Farhad Ramezanghorbani, Delaram Ghoreishi, Karl Leswing, Ed Harder, and Robert Abel

Schrödinger Inc. 120 West 45th Street New York NY

E-mail: leif.jacobson@schrodinger.com

Abstract

Transferable high dimensional neural network potentials (HDNNP) have shown great promise as an avenue to increase the accuracy and domain of applicability of existing atomistic force fields for organic systems relevant to life science. We have previously reported such a potential (Schrödinger-ANI) that has broad coverage of druglike molecules. We extend that work here to cover ionic and zwitterionic druglike molecules expected to be relevant to drug discovery research activities. We report a novel HDNNP architechture, which we call QRNN, that predicts atomic charges and uses these charges as descriptors in an energy model which delivers conformational energies within chemical accuracy when measured against the reference theory it is trained to. Further, we find that delta learning based on a semi-empirical level of theory approximately halves the errors. We test the models on torsion energy profiles, relative conformational energies, geometric parameters and relative tautomer errors.

1 Introduction

Over the last decade, techniques borrowed from the field of machine learning (ML) have greatly impacted many fields within computational chemistry. No field seems to be safe from the intrusion, even the historically empiricism-averse field of *ab initio* quantum chemistry. Techniques have emerged at nearly every level, between solving the electronic Schrödinger equation using a Neural Network (NN) based *ansatz* of the many-electron wave function,[1,2] machine learned density functionals,[3,4] empirical corrections of semi-empirical, density functional, or Hartree-Fock energies to higher level theories,[5–7] machine learned force fields,[8–11] and property prediction as complex as chemical reactivity.[12] Of particular interest to this work is what we might refer to as a NN potential energy surface (NN-PES). The goal of a NN-PES is to compute the potential energy of a chemical system, given the atomic positions. These models are typically trained to reproduce a particular model chemistry and are expected to reproduce the total electronic energy to chemical accuracy.[13] A NN-PES can be considered a type of atomistic force field: it maps atomic coordinates to energies, it is comparatively efficient to compute, and its parameters are empirically determined. However, a NN-PES differs from traditional biomolecular force fields[14] in that the total electronic energy is reproduced, as opposed to the energy relative to an arbitrary reference conformation. Additionally, the energy is purely a function of the chemical elements, coordinates, and net charge (as in an *ab initio* method), not relying on additional discontinuous input data such as atom types or assigned bonds. As such, NN-PES models promise both to increase the accuracy of empirical force fields and to expand their application domain to include important processes such as chemical reactions.

The most common approach to construct a NN-PES is to first transform the atomic coordinates into local atomic descriptors, or "features", which describe the local environment of each atom.[15–17] Alternatively some models allow the features to be "learnable" parameters, typically referred to as an embedding.[9,10] These features are transformed and reduced in dimensionality by a machine learning method to produce an energy. The parameters of the

model are determined by minimizing the error of predicted energies relative to a selected reference level of theory, typically density functional theory (DFT). In one of the simplest forms, and the form we focus on here, the feature vector describing each atom's environments is independently transformed by a neural network to output an atomic energy; these atomic energies are then summed to produce a molecular energy. This approach is often termed a high dimensional neural network potential (HDNNP).[15,18] Many of the works that have applied this algorithm have focused on applications in which the HDNNP is trained to the same chemical system to which it is applied. In this application, a model is trained very accurately within a small and well defined chemical space and then used to perform energy sampling within this chemical space, on larger simulation cells or time scales than would otherwise be possible. There is no expectation that the model could be applied to systems for which it is not trained. Conversely, it is also possible to develop a *transferable* HDNNP in which it *is* expected that the model will be applied to chemical systems for which it is not trained. Smith *et al.* demonstrated that an HDNNP, trained only to a representative, but still large, subset of the vast diversity of organic molecules, could produce accurate energies for molecules not in the training set.[8] A transferable HDNNP rests on two pillars: A model with the capacity to reproduce energies to within chemical accuracy, and a dataset which contains a representative sample of the space of relevant atomic environments.

Previously, some of the current authors have extended the work of the Roitberg lab to increase element coverage[19] (which was initially only four elements but has recently increased to seven[20]) and increased the precision when tested on rotamer scans of a diverse set of druglike organic molecules. The motivation for that work was to develop an HDNNP that could be confidently used to generate training data for parameterization of intramolecular terms in a biomolecular forcefield such as OPLS.[14] We termed this model Schrödinger-ANI (here abbreviated SANI) in homage to the ANI model from which it descends. There are several shortcomings of that model, the most significant for parameterization of force fields is that molecular ions were not in the domain of applicability. SANI and other models like it

have no way to distinguish charged from uncharged species, nor did we have a dataset with sufficient coverage for such training. Here we report our solutions to both of these problems.

In what follows, we will describe a model called QRNN (charge recursive NN), trained directly to DFT, and QRNN-TB, trained to the difference between DFT and GFN2-xTB. We will also compare to a version (QeqNN) with a charge equilibration method like that of Ko *et al*,[21] which has similar accuracy but less favorable computational scaling. We will demonstrate the effectiveness of these models by testing conformational energies of a broad set of neutral and ionic systems, as well as the Hutchison conformer test set[22] and relative energies of tautomers in the Tautobase dataset.[23]

2 Models

The main issue to be resolved is that the electronic energy is not uniquely defined by the nuclear positions alone: one must also define the system's net charge and spin multiplicity. In a model such as SANI these two electronic inputs are implicitly defined to be zero and one (neutral closed-shell), simply by having only neutral closed-shell examples in the training set. It is tempting to add ionic examples without modifying the model and to hope that the model can interpret which of these system are ionic and which not. Most chemists could guess on sight that a deprotonated carboxylic acid will be negatively charged, or a protonated amine will be positively charged. It is possible that one could proactively curate a dataset of ions such that the charge state can be inferred from coordinates alone, but we believe this approach is flawed and will eventually fail as the training set achieves broad coverage of chemical space. A simple example of the problem is any tertiary carbocation, which (for a given set of coordinates) would be indistinguishable from a tertiary carbanion. Any model that has features only depending on nuclear positions would unavoidably fail to distinguish these systems. Thus, one must specify the net charge to a model if the training set has broad coverage of geometries of closed shell ions. Here we do not consider systems

that are not closed-shell, and thus the spin multiplicity continues to be defined as 1 for all inputs.

The crucial idea of the models we study in this report is to use a simple physical model[24] to predict atomic charges, and then to use these charges as inputs to an energy model along with the usual geometric features. Essentially we transform the global net charge into local charges that can be used as part of description of the local atomic environment in a natural way. The parameters entering a charge model can be predicted by a neural network, a strategy first used by Ghasemi *et al.*[25]

Before describing the details of the models studied here we would like to highlight three recent highly relevant works. The first is the work of Zubatyuk *et al*[26] which extends AIMNet to charged systems. These authors have resolved the dilemma of adding a net charge in a novel and interesting way: a network which predicts energies of multiple charge states simultaneously. By definition, this scheme requires either the ionic state or the neutral state to be a radical. This is a disadvantage for our desired use case (closed-shell systems), but it could be useful in other contexts. Since this algorithm requires at least three energy labels for each training point (charge states +1, 0, -1), it seems that extending the scheme to any other charge states (+2, -2, etc) would require large increases in the size of the training set. A second work from Ko *et al*[18,21] reports a "fourth generation" HDNNP (4G-HDNNP) which is quite similar to one of the two models we present below. While our model was independently developed, it was motivated by earlier work by some of the same authors,[25] and so it is not unexpected that both would develop in the same direction. One can therefore interpret our work as an extension of Ko *et al* in which we demonstrate the ability to construct a *transferable* charge-aware HDNNP with broad coverage of organic molecules, and with improved computational scaling. Third is the work of Qiao *et al* who reports OrbNet,[5,6] a method which makes extensive use of features from the tight binding quantum mechanics method GFN2-xTB.[27] While the authors of that work have not explicitly demonstrated that OrbNet accurately reproduces energies of ions, our work suggests that their model likely has

the capacity to work for such systems. In fact, one could interpret the charge features we use as coarse-grained approximations to the quantum mechanical charge density which is used as a feature in OrbNet. Further, OrbNet relies on delta learning and is trained to the difference between GFN2-xTB and their reference level of theory. We demonstrate here that one can use this technique to boost the accuracy of our HDNNP models as well, at the cost of having to perform a tight-binding calculation.

The geometric features for the neural networks are modified Behler-Parrinello symmetry functions,[15] as described by Smith *et al.*[8] The Cartesian coordinates are transformed into a set of element resolved radial,

$$g^{(R)}_{i,Z,\eta,s} = \sum_{j\neq i}^{N_{atoms}} exp(-\eta(R_{ij} - R_s)^2)\ \delta(Z - Z_j)\ f_C(R_{ij}), \tag{1}$$

and angular symmetry functions,

$$g^{(A)}_{i,Z,Z',\eta,s} = 2^{1-\zeta} \sum_{j,k\neq i}^{N_{atoms}} (1 + cos(\theta_{ikl} - \theta_s))^{\zeta}\, exp(-\eta(\frac{R_{ij} + R_{ik}}{2} - R_s)^2)\ \delta(Z - Z_j)\delta(Z' - Z_k)\ f_C(R_{ij}) f_C(R_{ik}). \tag{2}$$

Here R_{ij} is the distance between atoms i and j, $f_C(R)$ is a switching function[8] which decays to zero at a radial cutoff, Z is the atomic number defining an element type and θ_{ijk} is the angle formed by atoms i, j and k, centered on atom i. R_s, θ_s, η and ζ are hyper-parameters which direct these symmetry functions to probe different regions in distance and angle space. Together, the radial and angular symmetry functions for atom i form a fixed length vector (G_i^{AEV}), referred to as the atomic environment vector (AEV)[8] which describe the local environment of that atom. By construction of the BP symmetry functions, the AEV is invariant to overall translation, rotation, and permutation of atom indices.

Next let us define an atomic neural network (ANN) which transforms the atomic AEV

of dimension f into an output dimension o, i.e.

$$\mathrm{NN}_Z : \mathbb{R}^f \rightarrow \mathbb{R}^o. \tag{3}$$

A commonly used algorithm to calculate atomic charges (q_i) is the QEq method.[24] This algorithm defines the charges as minimizing a simple energy expression,

$$\mathbf{q} = \mathrm{argmin} \left[\sum_i^{N_{atoms}} \chi_i q_i + \frac{1}{2} \sum_{ij}^{N_{atoms}} q_i J_{ij} q_j - \lambda (\sum_i q_i - Q_{tot}) \right] . \tag{4}$$

Here λ is a Lagrange multiplier, Q_{tot} is the total charge of the system, χ is the electronegativity of atom i and J_{ij} is the Coulomb interaction matrix. Following others[25,28] we parameterize the Coulomb interaction by assuming atom centered, spherical Gaussian charge distributions with a standard deviation σ_i, yielding

$$J_{ij} = \begin{cases} \frac{1}{\sqrt{\pi}\sigma_i} & i = j \\ \frac{\mathrm{erf}\left(\frac{R_{ij}}{\sqrt{2(\sigma_i^2 + \sigma_j^2)}} \right)}{R_{ij}} & i \neq j \ . \end{cases} \tag{5}$$

To determine the atomic charges from eq. 4 one can set the derivative with respect to q_i and λ to zero and solve the resulting linear equations.[21,24,25,28] The atomic electronegativity, χ_i, and width parameters, σ_i, are both allowed to be environmentally dependent and are predicted from an ANN. More precisely, we compute

$$X_i, S_i = \mathrm{NN}_{Z_i}^{Qeq}(G_i^{\mathrm{AEV}}) \tag{6}$$

and then calculate the Qeq parameters as $\chi_i = X_i^2$ and $\sigma_i = \sigma^0 + S_i^2$. This is done to ensure $\chi_i > 0$ (for physicality) and $\sigma_i > \sigma^0$ (for numerical stability). We have found σ^0 as small as 0.05 Angstrom to be sufficient to avoid numerical issues. Once the charges have been determined they can be used as features to a second ANN which defines the energy.

For atomic features, in addition to atomic charge and the standard AEV, we also use a charge-weighted radial AEV which describes the local charge environment:

$$G^{qR}_{i,\eta,s} = \sum_{j\neq i}^{N_{atoms}} q_j exp(-\eta(R_{ij} - R_s)^2)\ f_C(R_{ij}). \tag{7}$$

Finally, we compute the total energy as

$$E = \sum_{i}^{N_{atoms}} NN^{QeqNN}_{Z_i}(q_i, G^{qR}_i, G^{AEV}_i) + E_{disp} + E_{coul}(q)\ , \tag{8}$$

where E_{disp} is the empirical dispersion correction of the ωB97X-D functional[29,30] and E_{coul} is a truncated Coulomb energy which decays smoothly to zero at short range.[31] Eq. 8 along with Eq. 4 can be seen to be a fairly straightforward extension of the recently reported 4G-HDNNP[21] with the main difference being the charge-weighted radial AEV shown in Eq. 7 which we have found increases the capacity of the model by providing information about the local distribution of charge around an atom. The truncated Coulomb expression we use is given by

$$E_{coul} = \frac{1}{2}\sum_{i\neq j}^{N_{atoms}} \frac{q_i q_j}{R_{ij}} \frac{1}{1 + exp(b(a - R_{ij}))}\ . \tag{9}$$

In order to solve the Qeq equations and determine the charges one must first compute the Coulomb matrix in eq. 5 and then solve a set of linear equations that has the dimension of the number of atoms. This yields a method with the same asymptotic scaling as the linear solver, approximately $\mathcal{O}(N^3_{atoms})$, a distinct disadvantage compared to standard force fields which can be computed in quasi-linear time $\mathcal{O}(N_{atoms}\log(N_{atoms}))$.

The model we focus on in this report is an approximate form of the Qeq method with reduced computational scaling. This model removes the need to solve a system of linear equations, by shifting the burden onto the neural networks to predict a more difficult parameter $\tilde{\chi}_i$. We begin by separating the diagonal and off diagonal contributions to the Coulomb sum

in Eq. 4, thus defining an "effective" electronegativity:

$$\tilde{\chi}_i = \chi_i + \frac{1}{2}\sum_{j\neq i} J_{ij} q_j \tag{10}$$

to yield

$$\mathbf{q} = \text{argmin}\left[\sum_i \tilde{\chi}_i q_i + \frac{1}{2}\sum_i J_{ii} q_i^2 - \lambda(\sum_i q_i - Q_{tot})\right] \tag{11}$$

Without the explicit off-diagonal terms in J_{ij}, this expression now has a simple analytic solution in terms of the effective electronegativities $\tilde{\chi}_i$,

$$q_i = -\frac{1}{J_{ii}}\left(\tilde{\chi}_i - \lambda\right) \tag{12a}$$

$$\lambda = \frac{Q_{tot} + \sum_i \frac{\tilde{\chi}_i}{J_{ii}}}{\sum_i \frac{1}{J_{ii}}}. \tag{12b}$$

We interpret the effective electronegativities, $\tilde{\chi}_i$, as environmentally-dependent learnable atomic properties. From eq. 10 it is seen that these parameters have an implicit dependence on all other atomic charges. To approximate this effect without using non-local interactions, the charges are predicted iteratively, with each iteration using only local information for each atom. On iteration I we predict effective electronegativities

$$\tilde{\chi}_i^{(I)} = \text{NN}_{Z_i}^{QR}(q_i^{(I-1)}, G_i^{qR}, G_i^{AEV}). \tag{13}$$

which then enter Eq. 12 to predict the atomic charges. Note that it is not necessary to compute the full coulomb matrix for this method. Further, we fix the diagonal hardness parameters and use those taken from Caldeweyher *et al.*[28] The charge-weighted AEV, G_i^{qR}, of each atom is computed at each iteration, allowing charge information to propagate locally in a way reminiscent of message passing NNs,[9,10,26] though with a pre-defined type of message (atomic charges). Surprisingly, we have found that two iterations are sufficient for converging results to accuracy equal to that of the Qeq method. The charges are then used in an energy

model identical in form to Eq. 8. The resulting model, which calculates accurate system properties but preserves the quasi-linear scaling of standard force fields, we call QRNN for charge recursive neural network.

3 Dataset Construction

We have constructed a large dataset of ions of druglike molecules and their tautomers using active learning[32] which consists of approximately 18 million examples. We initialize the process with a dataset of neutral molecules which has been previously described.[19] A QeqNN model trained on the previous data at each active learning step is used to perform geometry optimization and normal model sampling (NMS).[8] As such, we must begin by appending to the neutral dataset an initial set of ionic data that roughly represents ionic species in general.

Active learning cycles are initiated with a dataset of very small ionic fragments optimized by DFT. This is done by fragmenting molecules appearing in ChEMBL[33] and ZINC[34] molecular datasets with the BRICS[35] implementation in the RDKit.[36] The fragmentation points are methyl-capped, unless they are carbon atoms already in which case they are hydrogen-capped. We retain all fragments with five heavy atoms or fewer, and then generate tautomers of charge states with a relative charge difference of -1, 0, and +1 electrons using Jaguar and EPIK tautomer enumerators.[37,38] All unique fragments (according to canonicalized SMILES strings) are retained and for each fragment we generate a single starting conformer with our in-house methodology, Fast3D. Each resulting conformer is then optimized with our reference level of DFT. For any optimization not resulting in a chemical reaction (see below), we generate 30 examples with NMS based on the DFT Hessian. This procedure resulted in 289,572 conformations which were added to our neutral dataset and used to provide an initial model.

From this point forward we use the following workflow to generate new training exam-

ples: select a new SMILES string from a molecular dataset, optimize the molecule with the HDNNP, generate charge states and tautomers which differ in charge from the initial molecule by -1, 0 and +1 electrons, optimize all tautomers with the HDNNP, perform sampling of torsions following our previously described methodology,[19] apply either NMS or an empirical sampling (ES) scheme as previously described,[19] and filter new examples by measuring the uncertainty of the model, excluding points with low uncertainty. The uncertainty is given by $\rho = \frac{\sigma}{N_{atoms}}$, where σ is the standard deviation of energies from an ensemble: A small set of trained models whose only difference is the initial values of the weights at the start of training. This uncertainty measure is required to be greater than 0.25 kcal/mol. Further, we skip NMS or ES on examples which have uncertainty below this threshold. We sample two SMILES strings at a time and pool all examples found with the above protocol, then randomly select a maximum of 1000 examples in each cycle. We use Chembl, ZINC, GDB and an internal proprietary dataset (CACDB) of druglike molecules to generate molecular examples in the form of SMILES strings. NMS and ES are performed in equal proportions. In the initial stages tautomers were not filtered and we retained all found tautomers. As the complexity of molecules grows the number of tautomers also grows combinatorially; to avoid an explosion of the number of high energy tautomers we filter them using a probability distribution given by

$$P_t(\epsilon_t) = \begin{cases} 1 & \epsilon_t < \mu_t \\ exp\left(-\left(\frac{\epsilon_t - \mu_t}{\sigma_t}\right)^2\right) & \epsilon_t \geq \mu_t \end{cases} \tag{14}$$

where ϵ_t is the tautomer energy relative to the lowest known tautomer and the mean and standard deviation are 10.0 and 20.0 kcal/mol respectively. In the final rounds of active learning we narrow our focus further by only using ionic examples directly from the molecular datasets, we expect that these examples are low lying tautomers as judged by some expert or software program that constructed the dataset or SMILES string. The rounds of active learning are summarized in Table 1. After each round DFT labels are generated and an ensemble of five members is trained.

Table 1: Summary of active learning cycles. Round 0 refers to the dataset built at the start of active learning by fragmentation, as described in the text.

round	max heavy atoms	new training points (thousands)	tautomer filtration	conformation search
0	5	289	None	No
1	6	158	None	No
2	6	328	None	No
3	8	357	None	No
4	8	302	None	No
5	10	391	No	No
6	10	1500	Eq. 14	No
7	12	2191	Eq. 14	No
8	12	1088	dataset tautomer only	No
9	14	1000	dataset tautomer only	Yes
10	14	846	dataset tautomer only	Yes
Total	14	8450	mixed	mixed

When generating tautomers by our enumeration protocol, potentially with very high energies, it is not uncommon that a geometry optimization will produce a chemical reaction. This reactivity may be enhanced because we do not use a solvent model, which could stabilize some tautomeric forms. As a general rule, we have found that the approximate model must be provided with some data about unfavorable, high-energy configurations, or else it may predict them to be favorable. If spurious chemical reactions take place, sampling these reactive pathways is a good way to correct the errors of the model; by labeling the erroneous points and re-training. Here we are focused on the description of conformational and tautomer energies, not the description of arbitrary chemical reactivity, and thus we do not necessarily wish to concentrate sampling on all such processes. There are two main types of reactions we have observed and we handle the two types differently: fragmentation and proton transfer. If an intramolecular proton transfer reaction occurs we retain examples along the entire optimization path, conversely, if fragmentation occurs we retain samples only up to the point of fragmentation. The latter allows us to potentially correct spurious fragmentation events while avoiding concentrating samples of molecular complexes which we

will focus on in future work. The fragmentation point or chemical reaction is detected by re-assigning bonds using distance thresholds based on covalent radii.

4 Details

We train models directly to DFT energies (see below) as well as to the difference between GFN2-xTB and DFT energies. We refer to these as *direct* and *delta* learning. For direct learning we train models of the form of eq. 8 to the atomization energy, that is, we subtract per-element atomic energy offsets from the DFT energies and train to the (much smaller and more tractable) residuals. For delta learning, we first generate delta energy labels

$$\delta E = E_{DFT} - E_{GFN2-xTB} \ , \tag{15}$$

then fit per-element atomic energy offsets to this difference (as for DFT energies) and again train to the residual of the labels and the per-element atomic energy offsets. When performing delta learning we omit the long-range terms E_{disp} and $E_{coul}(q)$ from Eq. 8 since we expect GFN2-xTB to reproduce the reference DFT reasonably well at long range.

All models are trained using a multitask loss function,[31,39] where the two tasks are charge prediction and energy prediction,

$$\mathcal{L}_{mtl} = \frac{\mathcal{L}_E}{2\sigma_E^2} + \frac{\mathcal{L}_q}{2\sigma_q^2} + log(\sigma_E \sigma_q) \ . \tag{16}$$

$\mathcal{L}_E$ and $\mathcal{L}_q$ are the loss functions for the energy task and charge task, respectively. The inverse weights, σ_E and σ_q are trainable parameters. This approach obviates the need to hand tune the weights of the tasks in the overall loss function and we have found that models trained with this method outperform models trained sequentially. The energy loss function is taken to be the squared error between the predicted energies and energy labels. For delta learning we use the squared error between the predicted charges and GFN2-xTB

charges. For direct learning we train to the squared error of predicted *dipole moments* to those predicted by DFT. The dipole moment is a physical observable which avoids the well-known arbitrary nature of atomic partial charge schemes. More importantly, training to dipole moments ensures correct long range electrostatic interactions, which is not true for charge decomposition schemes in general. This is less important for the delta learned models where we rely on GFN2-xTB to provide a good description of the long range interactions.

To minimize the loss function we use the AdamaxW optimizer (Adamax with decoupled weight decay) with a weight decay of 1.0e-4. We utilize early stopping with a patience of zero and a maximum of 100 epochs. Each ensemble member is trained to a 90/10 random split of the training data which is performed independently at run time (as such there is some degree of overlap between the training sets of the ensemble members). We use exponential moving averaged weights and biases to evaluate validation and test errors, the model parameters are updated every 10 batches with a mixing fraction of 0.999. As described in previous work,[19] we weight training examples with a Boltzmann-inspired weighting function which focuses the training on low-energy examples. The parameters of the weighting function are dependent on the net charge - for example, the distribution of atomization energies of cations is shifted relative to that of neutrals by the mean ionization energy. Each charge state is weighted separately so as to remove this energy shift. This shift is less important for delta learning where the distributions have much greater overlap. All network dimensions and Behler-Parinello type symmetry function hyperparameters are taken from the work of Smith[8] the only difference being that we use GELU activation function[40] for all neural networks, which have a continuous second derivative. All models are implemented and trained in a locally modified copy of the torchani open source software package,[41] which is an implementation of HDNNP type models utilizing PyTorch. Each neural network is trained on a single GPU with single-precision floating point, and all results are computed with double-precision inference.

All DFT calculations were computed with the Jaguar molecular electronic structure package[42] and utilize the pseudo-spectral approximation to accelerate computation of J and K

matrices. The reference functional is ωB97X-D[29] and we utilize the 6-31G* basis set. All calculations are run with default accuracy settings. All minima optimized with DFT are verified as such by checking that the number of imaginary vibrational frequencies is zero. Optimizations using the HDNNP models and GFN2-xTB were performed inside Jaguar with a local modification that allows specification of an external program which returns energy and gradient data. All PM7 calculations are performed with MOPAC2016[43] and all GFN2-xTB energy evaluations are computed with the xtb python API.[44]

5 Results and Discussion

We have trained both QRNN and QeqNN to a large dataset of organic molecules containing the elements H, C, N, O, P, S, Cl and F. This training set consists of roughly 18 million examples of conformations of neutral, ionic and zwitterionic organic species, see Sec. 3 for details. The molecules in the training set are representative of druglike molecules, ions and their tautomers. We have trained both models directly to DFT energies at the ωB97X-D/6-31G* level which we will refer to as *direct* learning as well as the difference between DFT and GFN2-xTB, which we will refer to as *delta* learning and denote by appending -TB to the name of the model, as in QRNN-TB. We evaluate the performance of these four models on test datasets which probe the accuracy of predicted torsional energy profiles, relative conformational energies, optimized geometries and relative tautomer energies. We focus on testing the models against our chosen reference level of theory, not against the highest possible level of theory. We expect that our chosen reference will have (potentially significant) basis set incompleteness errors for some systems, and this will be addressed in future work. Nonetheless we expect that if our model accurately reproduces the chosen reference it will still be generally useful for many applications and that we can have success in recapitulating higher levels of theory after some subset of the data is recomputed at that higher level.[45] All test sets that were generated in this work are reported in the supporting

information with DFT, GFN2-xTB and model energies labeled to facilitate reproduction of this work and comparison to other works. We hope the addition of test sets of ionic molecules contributes to the growing body of test sets available for machine-learned force fields.

5.1 Delta learning on the QM9 dataset

Table 2: Test errors for delta learning to the QM9 dataset, energies are given in meV. OrbNet results are taken from Qiao *et al.*[5]

	Direct Learning		Delta Learning		
size of training set	HDNNP	SchNet	HDNNP	SchNet	OrbNet
25,000	43	32	30	15	12
50,000	24	21	15	11	8
110,000	14	13	9	7	5

We start by evaluating the effectiveness of delta learning on a standard dataset, QM9.[46] We were inspired by recently reported results of OrbNet,[5,6] a model which uses quantum mechanically derived features. It was reported that OrbNet can achieve higher accuracy results with fewer training data points as compared to other methods. Besides its unique featurization, OrbNet also differs from other methods in its use of delta learning between GFN2-xTB and DFT, rather than learning DFT directly. We were interested in how much of the reported impressive performance could be replicated in a much more simple HDNNP, simply by using delta learning. To this end we trained an HDNNP to the ground state energy labels of QM9 using train/test splits reported by Qiao *et al.*[5] Details of the training procedure for this section are available in the supporting information. QM9 is a popular dataset used to test the "capacity" of a model, or it's ability to accurately train to complex data. In this section we also train a message passing model, SchNet,[47] as its design more closely resembles OrbNet and can be thought of as a more modern design.

Table 2 shows that the capacity of a message passing model such as SchNet is generally higher than that of an HDNNP as evidenced by the lower test errors. Also shown is that the use of delta learning reduces the test errors by nearly a factor of two for both HDNNP and

SchNet and improves test errors on the smallest training set to be competitive with direct learning on the largest training set. Still, the errors reported for OrbNet are lower than those we see for the other models. We cannot speculate that delta learning is the only advantage of OrbNet (which has other favorable attributes), but our results show that delta learning can increase the effective model capacity of other model types, not only OrbNet. In general, we can infer that learning the difference between a semi-empirical quantum method and DFT is more accurate than learning DFT energies directly. While a message passing model is expected to have higher capacity, training to such a model is much more expensive than an HDNNP and likely would be prohibitive for our full dataset. For that reason, the remainder of the article will focus on training charge-aware HDNNP models that were described in Sec. 2.

5.2 Torsion energies

We next turn to testing models trained on the full training set described in Sec. 3. Here our intent is to validate the models for describing geometries and energetics of conformers of organic molecules and their ions. We are primarily concerned with transferability and therefore test on molecules outside of the training set. One of our target applications is the use of HDNNP energies as reference data for fitting torsion parameters of traditional force fields.[14] As such, we test the accuracy of torsion scans of druglike molecules. We have generated relaxed torsion scans (optimized at the reference level of DFT) by fragmenting 1000 ionic or zwitterionic druglike molecules randomly selected from an internal dataset of such species (CACDB). This fragmentation results in 388 unique torsion scans of species containing at least one non-zero formal charge, which we will refer to as "charged" species, and 112 unique "neutral" molecules which contain no formal charges.

Using each method, we compute single-point energies for the DFT-optimized geometries in this test set. The errors are listed in Table 3 and shown graphically in Fig. 3 which can be found in the supporting information. The errors reported here are computed as root

mean squared error (RMSE) in relative energy (relative to the DFT minimum geometry) over each torsion scan, which consists of 12 equally spaced points. From these data it is clear that QRNN and QeqNN behave quite similarly and that delta learning improves the errors by nearly a factor of two. The direct learning models perform much worse on charged systems ($\sim$1.0 kcal/mol) than does a model trained only to neutrals and evaluated on neutrals (0.56 kcal/mol). This situation recovers when applying delta learning and we are able to achieve a mean error of only $\sim$0.5 kcal/mol. All of the charge-aware models dramatically outperform the tested semi-empirical theories, which for charged systems perform similarly to our previous HDNNP trained solely to neutral systems.[19] These results are displayed as a box and whisker plot in the supporting information which shows that remarkably, the largest outliers with the delta learned models are only slightly higher than the upper quartile of errors for the semi-empirical theories. As a general rule we find that QRNN and QeqNN models perform very similarly, with the QRNN version slightly outperforming regardless of whether or not we use direct or delta learning. For this reason and the fact that QRNN has superior formal computational scaling we will focus on this model for the remainder of this article. Tables which include results for the QeqNN models can be found in the supporting information.

Figure 1 shows the largest outlier for each of three methods: QRNN, QRNN-TB and GFN2-xTB. It is clear from these figures that the delta learned model is not a simple linear combination of GFN-xTB and QRNN; In panel (a) the curves for GFN2-xTB and QRNN-TB are quite similar whereas in panel (c) the QRNN and QRNN-TB curves nearly overlap. Finally, it is interesting to speculate that panel (a) represents a type of non-local resonance that the QRNN has difficulty recognizing (a resonance form for this molecule exists which indicates the bond being rotated has double bond character), whereas GFN-xTB and QRNN-TB seem to perform quite well. It is unclear if this is a general observation or simply a statistical outlier and warrants further investigation.

Table 3: Comparison of ML and semi-empirical methods against ωB97X-D/6-31G* for torsion scans. Energies are given in kcal/mol. There are 388 ionic or zwitterionic torsion scans and 112 torsion scans with no formal charges, each with 12 samples. The systems range from 10 to 46 heavy atoms, cover net charge states from -2 to +3 and have up to three atoms bearing a formal charge.

method	neutral subset	ionic subset	full set
QeqNN	0.58	1.04	0.94
QRNN	0.56	0.92	0.84
QeqNN-TB	0.35	0.54	0.50
QRNN-TB	0.33	0.51	0.47
SANI	0.56	2.05	1.71
PM7	1.72	2.18	2.07
GFN2-xTB	1.33	1.84	1.72

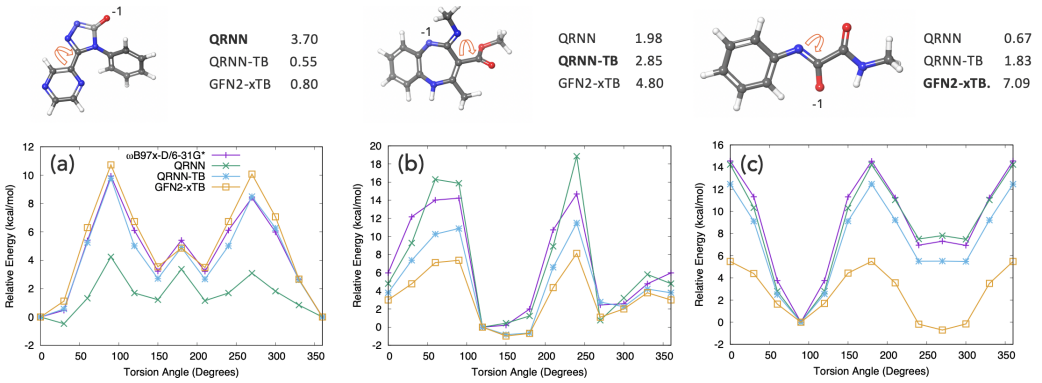

Figure 1: The worst performing torsion scan for the (a) QRNN, (b) QRNN-TB and (c) GFN2-xTB methods. The relative energy RMSE for each method is given in kcal/mol.

5.3 Relative Conformational Energies

We now turn our attention from torsional profiles to relative conformational energies of flexible molecules. Torsion scans probe energies which include torsion barrier heights, whereas relative conformational energies probe minima on the potential energy surface. To evaluate the conformational performance of our models we will use a test set reported by Folmsbee and Hutchison.[22] This test set is constructed from 622 neutral molecules and 86 charged systems, all of which are expected to show some degree of conformational freedom. Hutchison and co-workers performed a conformational search on each of these molecules and reported up to ten low-lying conformers of each species (see Ref[22] for details). The test molecules

contain up to 50 heavy atoms and 23 rotatable bonds, representing a strong test of transferability for our models, since our training molecules are substantially smaller. We have re-computed DFT energy labels at our reference level of theory and filtered any geometries that had unconverged self-consistent-field equations or contained chemical elements not supported by our model. This filtration left us with 576 neutral and 81 charged systems to study.

We follow the analysis performed in the original work on this dataset and compute the Mean Absolute Relative Error (MARE) and the square of the Pearson correlation coefficient (R^2) for each set of conformers. The median of these two metrics over all conformer sets are used to assess the quality of reproducing the reference energies and rank ordering of conformations. Hutchison showed that when comparing against a DLPNO-CCSD(T)/cc-pVTZ reference, DFT methods typically have median R^2 values of greater than 0.8 and MARE of less than 0.3 kcal/mol whereas the best empirical methods (GFN2-xTB, ANI) have R^2 less than 0.65 and MARE greater than 0.4 kcal/mol. We would consider results for our models (versus our reference level of theory) that are similar to those reported for DFT versus DLPNO-CCSD(T) to be a very encouraging sign. Tables 5 and 4 show our results for the neutral and charged subsets, respectively. Indeed, comparing the delta learned models to our reference level of theory, we see that we can achieve MARE less than 0.2 kcal/mol and R^2 values greater than 0.9 for both subsets. Again, we see that delta learning provides an impressive gain in accuracy over direct learning and that both methods show improved error statistics relative to the semi-empirical QM methods PM7 and GFN2-xTB alone. When comparing to the DLPNO-CCSD(T) references the delta learned models perform only slightly worse than when comparing to our reference level of theory, ωB97X-D/6-31G*. We are also encouraged by the observation that the difference between ωB97X-D/6-31G* and DLPNO-CCSD(T)/cc-pVTZ is similar to that between our delta learned models and ωB97X-D/6-31G*. Further, when comparing directly to DLPNO-CCSD(T), the delta learned models perform nearly as well as ωB97X-D/6-31G*.

Table 4: Results of relative conformational energies on the charged subset of the Hutchison test set.

reference	ωB97X-D/6-31G*		DLPNO-CCSD(T)	
method	Median MARE (kcal/mol)	Median R^2	Median MARE (kcal/mol)	Median R^2
QRNN	0.25	0.87	0.28	0.67
QRNN-TB	0.14	0.94	0.18	0.84
SANI	0.50	0.72	0.39	0.73
PM7	0.48	0.28	0.40	0.39
GFN2-xTB	0.36	0.62	0.44	0.55
ωB97X-D/6-31G*	–	–	0.17	0.88

Table 5: Results on relative conformational energies on the neutral subset of the Hutchison test set.

reference	ωB97X-D/6-31G*		DLPNO-CCSD(T)	
method	Median MARE (kcal/mol)	Median R^2	Median MARE (kcal/mol)	Median R^2
QRNN	0.27	0.87	0.34	0.74
QRNN-TB	0.16	0.94	0.27	0.84
SANI	0.28	0.84	0.35	0.75
PM7	0.54	0.36	0.61	0.31
GFN2-xTB	0.35	0.74	0.41	0.65
ωB97X-D/6-31G*	–	–	0.19	0.91

5.4 Accuracy of optimized geometries

In order to assess the accuracy of optimized geometries we have constructed a dataset of ionic conformers of druglike molecules. Two hundred molecules were drawn at random from the ZINC dataset.[34] The first hundred of these molecules were required to have sixteen heavy atoms or fewer, and the second hundred examples were required to have thirty-two heavy atoms or fewer. For each molecule we perform a mixed-mode conformational search with MacroModel utilizing the OPLS3e force field.[14] For each molecule, a maximum of 200 conformations is returned, with a maximum energy range of 12.0 kcal/mol.

In an effort to increase the diversity of our test set, we choose not to select only low-energy conformers for each molecule, since these are often similar to one another. Instead, we take the minimum-energy conformer and then up to nine other conformers drawn uniformly at random from the remainder. Each of the conformers is then geometry-optimized at the ωB97X-D/6-31G* level and then re-optimized, starting with the DFT geometry, with each model tested. After removing saddle point geometries we were left with 190 conformer sets that contained more than one conformer. We do not filter duplicate conformers after the DFT optimization and thus some similar minima will remain. However, conformational diversity of our conformer set is much greater than that of the Hutchison test set, with a mean relative energy range per molecule of 10.9 kcal/mol for our test versus only 2.9 kcal/mol for the Hutchison set.

To assess the accuracy of the optimized geometries, relative to ωB97X-D/6-31G* we compute errors in bond distances, angles, torsions and RMSD over all cartesian coordinates. We also report errors in relative energies and the median R^2 value, as in Sec. 5.3. The Cartesian RMSD is computed by first maximally aligning the geometries by minimizing the RMSD over translations and rotations of one of the pairs of molecules. Further, in order to test the geometries near charged groups, we compute errors over bond lengths, angles, and torsions for which at least one atom bears a formal charge differing from zero. Finally, we exclude from Table 6, any test geometry that undergoes a reaction during optimization

Table 6: Comparison of model optimized geometries to ωB97X-D/6-31G* optimized geometries. The errors in bond lengths, angles, and torsion are only computed for groups of atoms for which at least one atom bears a formal charge. In addition, geometries for which a reaction occurs during geometry optimization are excluded from the analysis, see text for details.

method	Mean Energy RMSD (kcal/mol)	Median R^2	Bond Distance RMSD (Å)	Angle RMSD (degrees)	Torsion RMSD (degrees)	Cartesian RMSD (Å)
QRNN	0.51	0.96	0.006	0.72	6.99	0.26
QRNN-TB	0.42	0.98	0.004	0.52	6.62	0.24
SANI	1.27	0.77	0.020	1.88	10.18	0.34
PM7	1.27	0.72	0.019	2.08	11.74	0.33
GFN2-xTB	1.23	0.79	0.022	1.49	11.81	0.42

with the model PES. These reactions exclusively involve proton transfer and ring closure. We have found that GFN2-xTB exhibits this behavior to a much greater extent than any other method, reacting on 155 inputs, compared to 19, 1, 5 and 3 inputs for PM7, SANI, QRNN and QRNN-TB respectively. This may indicate that GFN2-xTB tends to underestimate proton transfer barriers, a subject which warrants further investigation.

The results of the geometry optimizations are shown in Table 6. We see that errors in bond lengths, angles, torsion angles as well as overall Cartesian RMSDs improve from semi-empirical to QRNN and improve further for the delta learned model, QRNN-TB. The difference between direct and delta learned models for geometries is small. Our best models achieve errors of 0.004 Å, 0.52°, 6.62° and 0.24 Å for bond distances, angles, torsion angles and cartesian RMSDs respectively We also compute relative energy errors and R^2 values for this dataset and again find very good correlation with the reference. We believe the linear correlations and energy errors are higher than for the Hutchison dataset due to the larger range of energies in our test set, as discussed above. Overall the geometries produced by the delta learned models are in excellent agreement with DFT references and warrant future work related to energy ranking of conformers with these methods.

5.5 Relative Tautomer Energies

Table 7: Mean absolute relative errors for the tautomer pairs in Tautobase in kcal/mol

SANI	QRNN	QRNN-TB	GFN2-xTB	PM7
1.62	1.14	0.57	5.29	4.89

Finally, we evaluate our models' ability to reproduce relative tautomer energies. Our primary interest in this task is in the impact it may have upon workflows which compute the pKa of organic molecules, as in drug design. In order to compute pKa it is often necessary to rank-order a large number of tautomers of a certain charge state.[38] It is very difficult to rank these tautomers using a purely rules-based scheme, and computationally expensive to do so with DFT or other *ab initio* methods. Generally, this type of bond-changing energy difference is outside of the range of applicability of classical force fields, so this is an example of an area where an HDNNP could have a large impact in computational life science. Currently in common workflows, tautomers are ranked with semi-empirical methods, but the low accuracy of these methods for this task (see below) means that a wide energy window must be used for selecting samples for re-ranking with DFT, increasing the DFT workload.

Tautomerization free energies have been recently studied with an ML/MM model based on ANI.[48] In addition Vazquez-Salazar *et al.*[49] have recently explored the impact of the diversity of training set on the ability to compute relative tautomer energies in a public dataset, Tautobase,[23] and we use the same dataset here. Tautobase consists of 1673 tautomer pairs stored as SMIRKS strings. Here we neglect solvent effects and focus on reproduction of relative tautomer energies in the gas phase. (Solvent effects would need to be accounted for in order to make contact with the experimentally observed populations, which are also available in Tautobase.) We convert each tautomer to a single (arbitrary) three-dimensional starting conformation and optimize with ωB97X-D/6-31G*. Relative energies are then computed using each of the tested models. After filtering unsupported elements, failed SCF or geometry optimization jobs, and optimizations that landed on saddle-points, we are left with 1552

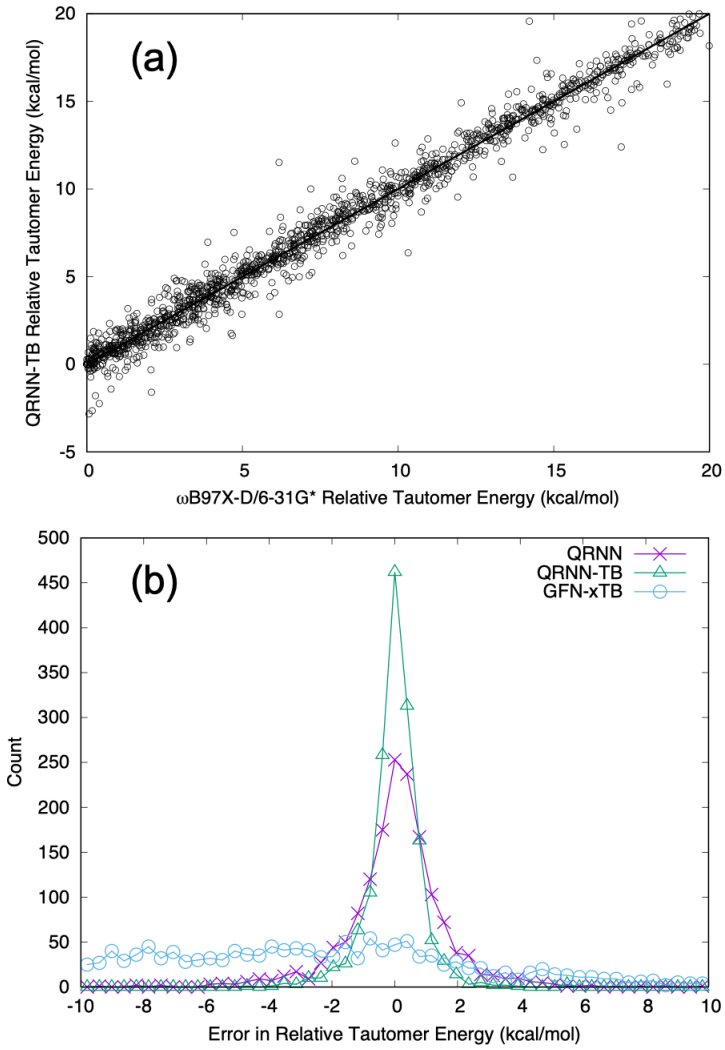

Figure 2: Correlation of ωB97X-D/6-31G* to QRNN-TB relative tautomer energies in Tautobase (a), the DFT minimum is chosen as the reference energy geometry. (b) displays the distribution of errors in the predicted relative tautomer energies for QRNN, QRNN-TB and GFN2-xTB.

tautomer pairs to analyze.

The mean error over the tautomer pairs are listed in Table 7; again we see a dramatic improvement in going from GFN2-xTB to direct learned, charge-aware HDNNP and finally to a delta learned model. For the best model the overall mean error is only 0.57 kcal/mol and only a very few tautomers are mis-ranked as shown in Fig. 2, where qualitative misrankings appear as negative values on the vertical axis. These results suggest that it is possible to replace semi-empirical ranking of relative tautomer energies with a charge-aware HDNNP, and with this replacement utilize a significantly smaller energy window to re-rank tautomers with an *ab initio* method, greatly reducing costs. This motivates further work to incorporate solvation effects into such a model.

6 Conclusions

In this work we have reported on the construction of a *transferable*, charge-aware, HDNNP with broad applicability to organic molecules, including their conformational energies, ions, and tautomers. We have presented the results of two models, QRNN and QeqNN, when these models are trained directly to DFT based energy labels and to the difference between GFN2-xTB and DFT. While a model almost identical to QeqNN has recently been reported[21] this is, to our knowledge, the first report of a transferable HDNNP applied to a broad range of closed shell ionic systems and their tautomers. We also report a novel charge model, QRNN, which performs at least as well as Qeq methods like the previously reported 4th generation HDNNP[21] and has superior scaling properties. We find that all models are able to achieve errors below 1 kcal/mol on relative energies of conformations on a broad range of ionic systems outside of the training set. Further, we find that delta-learning based on GFN2-xTB can nearly halve this error. We believe this technique could make a large impact in workflows involving conformational energy analysis and ranking, by providing geometries and energies that are highly accurate relative to the level of theory they are trained to.

Finally, we show that our models are able to rank relative tautomers to less than 1 kcal/mol accuracy, a large advance for efficient energy ranking of tautomers. These results neglect solvent effects, which are important in both conformation and tautomer energy rankings, and this will be addressed in future work. In addition, we do not expect inter-molecular interactions to be well described by any model trained to the current data set due to the fact that the training examples are all single molecules. This will be extended in future work.

Acknowledgement

We would like to thank Art Bochevarov and Adrian Roitberg for many useful discussions.

Supporting Information Available

6.1 QM9 training details

The results shown in Sec. 5.1 utilize the QM9 dataset [46,50] and use the train/test splits given in the supporting information of Ref. [5] We train to the electronic energy as this is consistent with the training tasks we are most interested in. Because the dataset is so small, rather than using early stopping the HDNNP results are trained using a fixed number (200) of epochs. The SchNet results utilized the exact number of examples listed in Table 2 whereas the HDNNP models were trained to 24,999, 49,995 and 110,000 examples because we specify the percentage of the dataset to use as training instead of the exact number of examples; we expect this discrepancy to have minimal impact on the results. The SchNet results were trained using the SchNetPack open source package with hyperparameters recommended in SchNetPack tutorials and given in the supporting information. [47] The models reported in Sec. 5.1 do not use dispersion corrections, long range coulomb energies or charge features, they are simply short range models.

6.2 Addition figures and tables with all models results

Tables 8, 9, 10 and 11 show the error statistics on the full set of models. Figure 3 shows graphically the error distributions and outliers for the rotamer energy tests presented in 5.2 in the main text.

Table 8: Results on relative conformational energies on the charged subset of the Hutchison test set.

reference	ωB97X-D/6-31G*		DLPNO-CCSD(T)	
method	Median MARE (kcal/mol)	Median R^2	Median MARE (kcal/mol)	Median R^2
QeqNN	0.29	0.78	0.27	0.67
QRNN	0.25	0.87	0.28	0.67
QeqNN-TB	0.16	0.94	0.18	0.87
QRNN-TB	0.14	0.94	0.18	0.84
SANI	0.50	0.72	0.39	0.73
PM7	0.48	0.28	0.40	0.39
GFN2-xTB	0.36	0.62	0.44	0.55
ωB97X-D/6-31G*	–	–	0.17	0.88

Table 9: Results on relative conformational energies on the neutral subset of the Hutchison test set

reference	ωB97X-D/6-31G*		DLPNO-CCSD(T)	
method	Median MARE (kcal/mol)	Median R^2	Median MARE (kcal/mol)	Median R^2
QeqNN	0.28	0.84	0.35	0.76
QRNN	0.27	0.87	0.34	0.74
QeqNN-TB	0.16	0.94	0.27	0.83
QRNN-TB	0.16	0.94	0.27	0.84
SANI	0.28	0.84	0.35	0.75
PM7	0.54	0.36	0.61	0.31
GFN2-xTB	0.35	0.74	0.41	0.65
ωB97X-D/6-31G*	–	–	0.19	0.91

6.3 Supplementary data files

All geometries used to analyze errors statistics in the main text with energy labels for QeqNN, QeqNN-TB, QRNN, QRNN-TB, GFN2-xTB, PM7 and ωB97X-D/6-31G* are available for

Table 10: Comparison of optimized geometries to ωB97X-D/6-31G* optimized geometries. The errors in bond lengths, angles, and torsion are only computed for groups of atoms for which at least one atom bears a formal charge. In addition, geometries for which a reaction occurs during geometry optimization are excluded from the analysis, see main text for details.

method	Mean Energy RMSD (kcal/mol)	Median R^2	Bond Distance RMSD (Å)	Angle RMSD (degrees)	Torsion RMSD (degrees)	Cartesian RMSD (Å)
QeqNN	0.61	0.96	0.007	0.75	7.27	0.26
QRNN	0.51	0.96	0.006	0.72	6.99	0.26
QeqNN-TB	0.44	0.98	0.004	0.53	6.83	0.23
QRNN-TB	0.42	0.98	0.004	0.52	6.62	0.24
SANI	1.27	0.77	0.020	1.88	10.18	0.34
PM7	1.27	0.72	0.019	2.08	11.74	0.33
GFN2-xTB	1.23	0.79	0.022	1.49	11.81	0.42

Table 11: Mean absolute relative energies for the tautomer pairs in Tautobase in kcal/mol

SANI	QeqNN	QRNN	QeqNN-TB	QRNN-TB	GFN2-xTB	PM7
1.62	1.28	1.14	0.59	0.57	5.29	4.89

download in the file supplementary_data_files.tar.gz. Each test set is stored in a separate directory and each set of conformers, rotamers or tautomers is stored in a separate json file. The format for the files is given in a README.format file in each directory.

References

(1) Pfau, D.; Spencer, J. S.; Matthews, A. G. D.; Foulkes, W. M. C. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2020**, *2*.

(2) Hermann, J.; Schätzle, Z.; Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **2020**, *12*, 891–897.

(3) Li, L.; Hoyer, S.; Pederson, R.; Sun, R.; Cubuk, E. D.; Riley, P.; Burke, K. Kohn-Sham

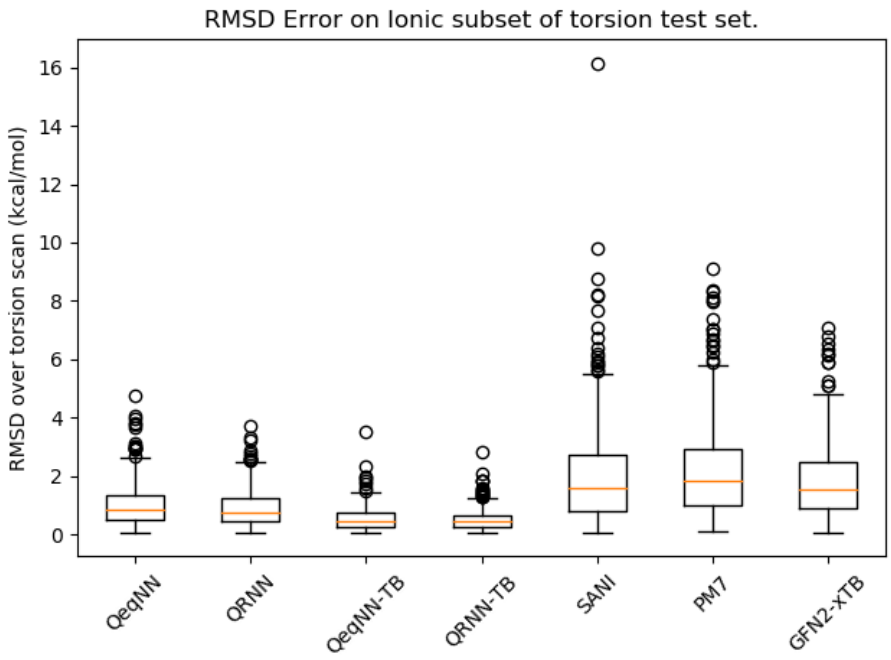

Figure 3: Box and whisker plot for the relative energy RMSE over torsion scans of charged systems. The yellow line inside the box shows the median RMSE value and the edges of the box display the lower and upper quartile. The upper whisker is drawn at 1.5 times the interquartile range above the upper quartile while the lower is at zero. Points outside of this range are shown explicitly.

Equations as Regularizer: Building Prior Knowledge into Machine-Learned Physics. *Phys. Rev. Lett.* **2021**, *126*, 036401.

(4) Nagai, R.; Akashi, R.; Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Materials* **2020**, *6*.

(5) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., 3rd OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153*, 124111.

(6) Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G. A.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R.; Miller, T. F., III OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *ArXiv* **2021**,

(7) Sinitskiy, A. V.; Pande, V. S. Physical machine learning outperforms "human learning" in Quantum Chemistry. *ArXiv* **2019**,

(8) 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.

(9) 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.

(10) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(11) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci Adv* **2019**, *5*, eaav6490.

(12) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

(13) Herr, J. E.; Yao, K.; McIntyre, R.; Toth, D. W.; Parkhill, J. Metadynamics for training neural network model chemistries: A competitive assessment. *J. Chem. Phys.* **2018**, *148*, 241710.

(14) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863–1874.

(15) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(16) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*.

(17) 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.

(18) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **2021**, *54*, 808–817.

(19) Stevenson, J.; Jacobson, L. D.; Zhao, Y.; Wu, C.; Maple, J.; Leswing, K.; Harder, E.; Abel, R. Schrodinger-ANI: An Eight-Element Neural Network Interaction Potential with Greatly Expanded Coverage of Druglike Chemical Space. *ChemRxiv* **2019**,

(20) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(21) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **2021**, *12*, 398.

(22) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies using Electronic Structure and Machine Learning Methods. *Int. J. Quantum Chem.* **2020**, *121*, e26381.

(23) Wahl, O.; Sander, T. Tautobase: An Open Tautomer Database. *J. Chem. Inf. Model.* **2020**, *60*, 1085–1089.

(24) Rappe, A. K.; Goddard, W. A. Charge equilibration for molecular dynamics simulations. *The Journal of Physical Chemistry* **1991**, *95*, 3358–3363.

(25) Ghasemi, S. A.; Alireza Ghasemi, S.; 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* **2015**, *92*.

(26) Zubatiuk, T.; Isayev, O. Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence. *Acc. Chem. Res.* **2021**, *54*, 1575–1585.

(27) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - an Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Compute.* **2019**, *15*, 1652–1671.

(28) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122.

(29) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(30) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(31) Ramezanghorbani, F. Developing Machine Learning Models to Enhance Applicability of Neural Network Potentials in Drug Discovery. Ph.D. thesis, 2020; Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-06-07.

(32) 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.

(33) Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

(34) Stirling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.

(35) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *Chem. Med. Chem.* **2008**, *3*, 1503–1507.

(36) Landrum, G. RDKit. `https://www.rdkit.org/`.

(37) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol.* **2007**, *21*, 681–691.

(38) Watson, M. A.; Yu, H. S.; Bochevarov, A. D. Generation of Tautomers Using Micro-pKa's. *J. Chem. Inf. Model.* **2019**, *59*, 2672–2689.

(39) Cipolla, R.; Gal, Y.; Kendall, A. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2018**,

(40) Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arxiv* **2016**,

(41) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60*, 3408–3415.

(42) Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A. Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *Int. J. Quantum Chem.* **2013**, *113*, 2110–2142.

(43) Stewart, J. MOPAC2016. `http://openmopac.net/MOPAC2016.html`, Accessed: 2021-6-10.

(44) grimme-lab, xtb-python. `https://github.com/grimme-lab/xtb-python`, Accessed: 2021-6-10.

(45) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.

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

(47) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.-R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448–455.

(48) Wieder, M.; Fass, J.; Chodera, J. D. Fitting quantum machine learning potentials to experimental free energy data: Predicting tautomer ratios in solution. *bioRxiv* **2021**, 2020.10.24.353318.

(49) Vazquez-Salazar, L. I.; Boittier, E.; Unke, O. T.; Meuwly, M. Impact of the characteristics of quantum chemical databases on machine learning predictions of tautomerization energies. *ArXiv* **2021**,

(50) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. and Model.* **2012**, *52*, 2864–2875.