Teaching a neural network to attach and detach electrons from molecules

Roman Zubatyuk[a], Justin S. Smith[b], Benjamin T. Nebgen[b], Sergei Tretiak[b,c], Olexandr Isayev[a*]

[a] *Department of Chemistry, Carnegie Mellon University, Pittsburgh PA, 15213, USA*

[b] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

[c] *Center for Integrated Nanotechnologies, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA*

* olexandr@olexandrisayev.com

Physics-inspired Artificial Intelligence (AI) is at the forefront of methods development in molecular modeling and computational chemistry. In particular, interatomic potentials derived with Machine Learning algorithms such as Deep Neural Networks (DNNs), achieve the accuracy of high-fidelity quantum mechanical (QM) methods in areas traditionally dominated by empirical force fields and allow performing massive simulations. The applicability domain of DNN potentials is usually limited by the type of training data. As such, transferable models are aimed to be extensible in the description of chemical and conformational diversity of organic molecules. However, most DNN potentials, such as the AIMNet model we proposed previously, were parametrized for neutral molecules or closed-shell ions due to architectural limitations. In this work, we extend our AIMNet framework toward open-shell anions and cations. This model explores a new dimension of transferability by adding the charge-spin space. The resulting AIMNet model is capable of reproducing reference QM energies for cations, neutrals and anions with errors of 4.1, 2.1, 2.8 kcal/mol, respectively, compared to the reference QM simulations. The spin-charges have errors 0.01-0.06 electrons for small organic molecules containing nine chemical elements {H, C, N, O, F, Si, P, S and Cl}. Thus the proposed AIMNet model allows researchers to fully bypass QM calculations and derive the ionization potential, electron affinity, and conceptual Density Functional Theory quantities like electronegativity, hardness, and condensed Fukui functions. We show that these descriptors, along with learned atomic representations, could be used to model chemical reactivity through an example of regionselectivity in electrophilic aromatic substitution reactions.

Introduction

A large body of research in the field of chemistry is concerned with the flow and behavior of electrons, which gives rise to important phenomena such as making and breaking chemical bonds. Quantum chemistry (QC) provides a mathematical framework for describing the behavior of atomistic systems thorough solution of Schrödinger equation, allowing for a detailed description of charge distribution and molecular energetics. QC provides the tools to accurately construct the potential energy surface (PES) of molecules, i.e., energy as a function of molecular geometry. Density Functional Theory (DFT) framework often underpins the methods of choice for such calculations when working with medium size molecules by providing a good balance between accuracy and computational cost. Unfortunately, standard DFT methods for the treatment of the N-electron system typically require ~$O(N^3)$ numerical cost. This cubic scaling has become a critical challenge that limits the applicability of DFT to a few hundred atom systems. This also limits the accessibility of longer dynamical simulation time scales, which are critical for simulating certain experimental observables. Consequently, a lot of progress has been made in the development of interatomic potentials providing a complex sought out PES functional (geometry -> energy) using machine learning (ML),[1,2] which have been applied to a variety of systems.[3–7] These models tend to provide highly accurate PESs for molecules and materials with a relatively low number of degrees of freedom.[8–11]

Deep neural networks (DNN)[12,13] are a particular class of ML algorithms proven to be universal function approximators.[14] These DNNs are perfectly suitable to learn a representation of the PES for molecules. There are multiple distinct DNN models for ML potentials reported in the literature. They could be divided into two groups. The original Behler-Parrinello (BP)[15] and its modifications ANI[16,17] and TensorMol[18] rely on 2-body (radial) and 3-body (angular) symmetry functions to construct a unique descriptor of atomic environment for a particular atom, then use a DNN to predict atomic properties as a function of that descriptor. Other models, for example, Hip-NN,[19] DTNN,[6] SchNet,[20] and PhysNet[21] use non-invariant radial symmetry functions or interatomic distances and iteratively construct a representation of the atomic environment through message-passing techniques.[22]

The ANAKIN-ME (ANI) method[16,23] is one example of a technique for building transferable DNN-based molecular potentials. The key components of ANI models are the diverse training dataset[24] and BP type descriptors[15] with modified symmetry functions.[16] The ANI-1ccx dataset was built from energies and forces for ~60K small organic molecules containing 5 and 0.5 million non-equilibrium molecular conformations calculated at DFT and high fidelity Coupled Clusters (CCSD(T)) levels, respectively.[24] Notably, these conformations were selected with an active learning technique.[24] Most recent studies agree that ML models obtained with self-adapted training using active learning are more accurate and data-efficient than models with a static, fixed training set.[5,23,25] Test cases showed ANI-1ccx model to be chemically accurate compared to the reference Coupled Cluster calculations and exceeding the accuracy of DFT in multiple applications.[17] Finally, the AIMNet (Atoms-In-Molecules neural Network) architecture, a chemically inspired, modular deep neural network molecular potential improves the performance of ANI models for long-range interactions and continuum solvent effects.[26]

Physical properties of molecular systems are often labeled as *intensive* or *extensive* properties. This nomenclature relates to the dependency of the property upon the size of the system in question.[27] The notation has been introduced by Tolman over one hundred years ago.[28] Only a few recent reports have attempted to use ML for *intensive* properties, independent of the system size, which pose a challenge ML techniques due to spatial non-locality and long-range interactions.

These studies were focused on frontier orbital energies, singlet-singlet, or singlet-triplet transition energies computed with time-dependent DFT (TDDFT).[29–32]

In this work, we examine how DNN models like ANI and AIMNet can be applied to predicting intensive properties like electron attachment (electron affinity) and electron detachment (ionization potential). The conventional wisdom would be to fit different ML potentials for every quantum-mechanical state (neutral, cation, and anion) similar to TDDFT works.[31] QM calculations for ionized states of the molecule are typically more expensive due to the unrestricted Hamiltonian formalism and subsequent spin polarization of orbitals. Therefore, we seek to answer a critical question: Can we fuse information from different molecular charge states to make ML models more accurate, general and data efficient? With the success of deep learning in many applications involving complex multimodal data, this question can be addressed by learning different states of the molecules with one common ML model, and the goal is to use the data in a complementary manner toward learning a single complex problem. We explore two synergistic strategies for joint modeling: multitask learning and data fusion. One of the main advantages of joint learning is that a hierarchical representation can be automatically learned for each state, instead of individually training independent models. In addition to electron attachment and detachment energies, we also choose to learn spin-polarized charges for every state reflecting quantum mechanics of the wavefunctions. This choice of properties is deliberate, as it allowed us to compute reactivity descriptors such as philicity indices and Fukui functions based on conceptual Density Functional Theory (c-DFT) theory.[33,34] c-DFT, or Chemical Reactivity Theory, is a powerful tool for the prediction, analysis, and interpretation of chemical reactions.[35] Here all c-DFT indexes were computed directly from the neural network without additional training that permitted us to bypass quantum mechanical calculations entirely.

Methods

Machine learning models. High-dimensional neural networks (HDNNs)[15] rely on the chemical bonding nearsightedness ('chemistry is local') principle by decomposition of the total energy of a chemical system into atomic contributions. For each atom in the molecule, HDNN models encode the local environment (a set of atoms within a pre-defined cutoff radius) as a fixed-size vector and use it as an input to a feed-forward DNN function to infer individual atomic contribution to the total energy. The ANI model (Figure 1a) transforms coordinates **R** of the atoms in the molecule into atomic environment vectors (**AEV**s): a set of translation, rotation, and permutation invariant two-body radial $g_{ij}^{(r)}$ (gaussian expansion of interatomic distances) and three-body angular $g_{ijk}^{(a)}$ (joint gaussian expansion of average distances to a pair of neighbors and cosine expansion of angle to those atoms) symmetry functions, where index *i* corresponds to a "central" atom and *j* and *k* refer to the atoms from its environment. Using the information of atomic species types **Z**, the **AEV**'s are reduced in a permutation-invariant manner into the **Embedding** vectors ***G,*** which encode both geometrical and type information of the atomic environment. The ANI model uses the concatenation of the sums of $g_{ij}^{(r)}$ and $g_{ijk}^{(a)}$ which correspond to a distinct chemical type of neighbor, or a combination of the types for two neighbors. This is equivalent to multiplication of the matrices $\boldsymbol{g}_{\boldsymbol{i}}^{(\boldsymbol{r})}$ and $\boldsymbol{g}_{\boldsymbol{i}}^{(\boldsymbol{a})}$ with rows composed of **AEV**'s, and corresponding matrices $\boldsymbol{A}^{(r)}$ and $\boldsymbol{A}^{(a)}$ composed with one-hot (categorical) encoded atom or atom-pair types:

$$\boldsymbol{G}_{\boldsymbol{i}} = \left\{ \boldsymbol{g}_{\boldsymbol{i}}^{(\boldsymbol{r})\top} \boldsymbol{A}^{(\boldsymbol{r})}, \boldsymbol{g}^{(\boldsymbol{a})\top} \boldsymbol{A}^{(\boldsymbol{a})} \right\} \tag{1}$$

By definition, the HDNN models suffer from the “curse of dimensionality” problem. Namely, the size of $\boldsymbol{G}$ depends on the number of unique combinations of atomic species included in parametrization (size of vectors in $\boldsymbol{A}^{(a)}$). Also, since the information about the type of the “central” atom is not included in $\boldsymbol{G}$, it uses multiple independent DNNs defined for each atom type ($\mathcal{F}^{(Z_i)}$) to model **Interactions** of the atom with its environment and outputs atomic energy E_i:

$$E_i = \mathcal{F}^{(Z_i)}(\boldsymbol{G_i}) \qquad (2)$$

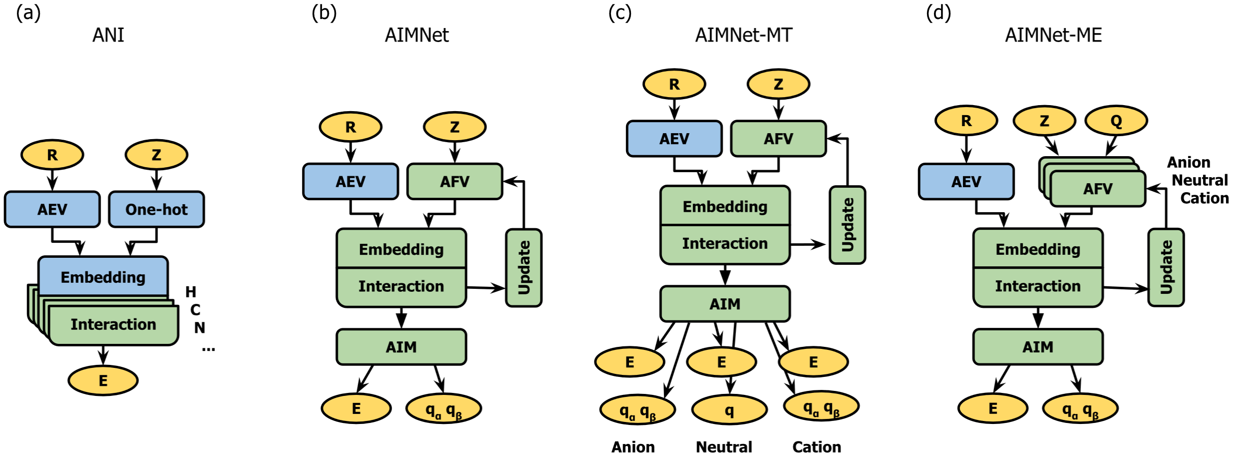

Figure 1. Neural network architectures explored in this work. Models from literature: a) ANI[16], b) AIMNet[26]; Here each model is separately trained for neutral species, cations and ions. Models introduced in this work: c) AIMNet-MT: a multitask model jointly trained on all data which concurrently predicts energies and charges for neutral species such as cations and ions; and d) AIMNet-ME, a multi-embedding model conditioned on a total molecular charge to predict the energy of a particular state. The yellow blocks show input data (coordinates **R**, atomic numbers **Z** and total molecular charge **Q**) and output quantities (energies **E** and spin-polarized charges **q**). The green blocks denote trainable modules and the blue blocks are fixed encodings.

The AIMNet model (Figure 1b) was developed to address the aforementioned issues with the ANI model. Instead of one-hot encoding of atomic species, it uses learnable atomic feature vectors (**AVF**s) $\boldsymbol{A}$ in Eq. 1. The **AFV** vectors encode similarities between chemical elements. This approach eliminates dependence of the size of **Embedding** layer on the number of parametrized chemical species. The AIMNet model utilizes the idea of multimodal learning, making a simultaneous prediction of different atomic properties from several output heads attached to the common layer of multi-layer neural nets. This layer is enforced to capture the relationships across multiple learned modalities and serves as a joint latent representation of atoms in the molecule. Therefore we call this layer an **AIM** vector. Finally, the architecture of AIMNet has a specific implementation of message passing through updating the **AFV** based on neighbor atoms atomic environments. This way, the model operates iteratively, at each iteration *t* predicting atomic properties $\boldsymbol{P}$ and updated features $\boldsymbol{A}$, using the same (shared across iterations) neural network function $\mathcal{F}$:

$$\{P_i^t, \boldsymbol{A}_i^{t+1}\} = \mathcal{F}(\boldsymbol{G}_i^t, \boldsymbol{A}_i^t) \qquad (3)$$

The approach has an analogy with a solution of one-electron Schrodinger equation with self-consistent field (SCF) iterations, where one-electron orbitals (**AFV** in case of AIMNet) adapt to the potential introduced by other orbitals in the molecule (embedding vectors $\boldsymbol{G}$ in case of AIMNet). Though there is no convergence guarantee for AIMNet due to the absence of the variational principle, in practice statistical errors decrease and converge at $t = 3$ being an empirical observation.

We use the ANI and AIMnet models as baselines to compare the results of new AIMNet-MT and AIMNet-ME models developed in this work. AIMNet-MT (Figure 1c) or multitask is a straightforward extension of the AIMNet model by joint training one model that simultaneously predicts energies and spin-polarized charges for neutral species, cations, and anions with multiple output heads from same **AIM** layer. In AIMNet-MT, all three states share the same **AFV** representation, **Interaction,** and **Update** blocks. This setting allows us to evaluate if the common feature representations can capture correlations across different states and, if possible, take advantage of that. In contrast, AIMNet-ME (Figure 1d), the multi-embedding model, shares the same **Interaction**, **Update** blocks, and output heads, but different initial **AFV** for anions, cations, and neutral molecules. An essential feature of AIMNet-ME is its ability to learn different representations inside one model. This feature can be exploited to have a fine-grained control over how learned representations are fused in the **Embedding** layer.

Dataset construction. For the training dataset, we randomly selected about 200k neutral molecules from the UNICHEM database[36] with molecule size up to 16 'heavy' (i.e., non-hydrogen) atoms and set of elements {H, C, N, O, F, Si, P, S and Cl}. We choose molecular dynamics (MD) as a fast and simple method to explore molecular PESs around their minima. We expect that thermally. Notably, all traditional molecular force fields are designed to describe closed-shell molecules only. Therefore, to overcome this limitation, we choose quantum mechanically derived force field (QMDFF[37]) as an efficient method to construct system-specific and charge-specific mechanistic potential for a molecule. We relied on the GFN2-xTB[38] tight-binding model to obtain minimum conformation, force constants, charges, and bond orders that are needed for the QMDFF model.

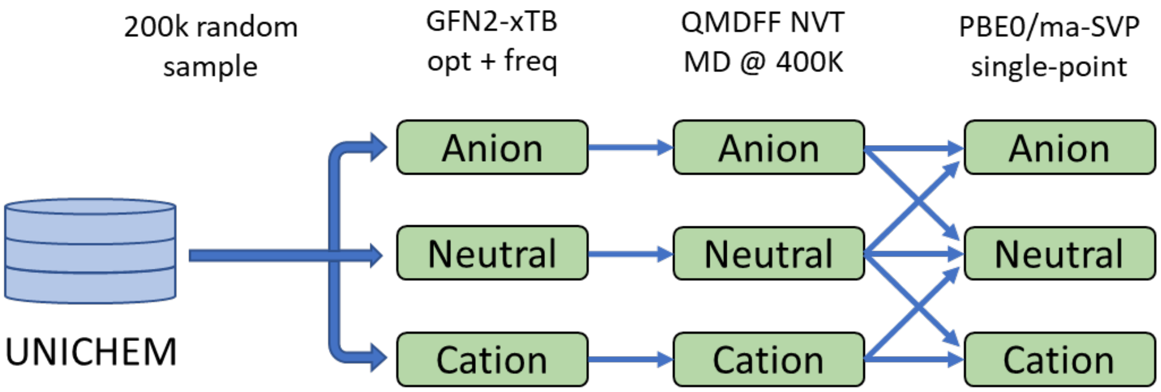

Figure 2. The overall workflow targeting dataset generation for the energetics of neutral and charged molecular species.

The workflow to generate molecular conformations is summarized in Figure 2. Starting from SMILES representations, we generated a single 3D conformation for each molecule using the RDKit[39] library. The molecule in each of three charge states (i.e., neutral, cation and anion) was optimized using the GFN2-xTB method, followed by a calculation of force constants, charges and bonds orders to fit molecule-specific QMDFF parameters. This custom force field was used to perform 500ps NVT MD run, with snapshots collected every 50 ps for the subsequent DFT calculations. For each snapshot, we performed several single-point DFT calculations at PBE0/def2-ma-SVP level, with a charge for the molecule set to the value at which the MD was performed, as well as its neighboring charge state, i.e., -1, 0 for anions, -1, 0, +1 for neutral, and 0, +1 for cations (Figure 2). This results in up to 70 single-point DFT calculations per molecule. The described scheme affords the calculation of both *vertical* and *adiabatic* electronic energy differences. The former corresponds to the vertical transition from the initial (ground) state of the neutral system at its instantaneous non-equilibrium geometry to the lowest-energy state of the cation, and vertical electron attachment energy of a cation at its instantaneous non-equilibrium geometry to the lowest-energy ground state. The same is applicable to anions. All DFT calculations were performed using ORCA 4.0 package.[40] Atomic spin-polarized charges were calculates using minimal basis iterative stockholder (MBIS) scheme[41] as implemented in HORTON package.[42]

We split all data into two subsets: Ions-12 dataset contains 6.44M structures with up to 12 heavy atoms of which 45%, 25% and 30% are neutral, cations and anions, respectively. Ions-16 dataset has 295k structures of 13-16 non-hydrogen atoms size with 48%, 24% and 26% of neutral, anionic and cationic species, respectively. We used Ions-12 dataset for training and validation, whereas Ions-16 was utilized for testing. Ions-16 dataset has larger, more complex structures and thus probes the model transferability.

Training protocol. The ANI model and AIMNet variants were trained using minibatch gradient descent powered by the Adam optimizer.[43] For training performance considerations, all minibatches were composed of molecules with the same number of atoms, to avoid padding. Proper feed data shuffling data was achieved by accumulating gradients on model parameters from 4 random minibatches. The effective batch size was 1000 molecules of different sizes. The training objective was minimization of weighted multi-target mean squared error (MSE) loss function with the general formula:

$$\mathcal{L} = \frac{1}{N}\sum_{t=1}^{3} w_t \sum_{i=1}^{N} \left[\left(E_i - \hat{E}_i\right)^2 + \frac{k^2}{2M}\sum_{j=1}^{M}\sum_{s\in\{\alpha,\beta\}}\left(q_{ij} - \hat{q}_{ij}\right)^2 \right] \quad (4)$$

In this formula, E and $\hat{E}$ are the target and predicted molecular energies, q and $\hat{q}$ – target and predicted atomic charges, respectively, *s* denotes spin component of atomic charge, N and M are corresponding numbers of molecules in minibatch and number of atoms in the molecules, and k is an empirical scaling factor equal to 15 kcal mol^{-1} e^{-1}. The total loss function contains error contributions from all three SCF-like iterations *t* with weights *w* = [0.15, 0.25, 0.60]. Although all final predictions of the AIMNet models were obtained with *t*=3, we found it beneficial to restrain a network to give reasonably accurate results on earlier iterative passes, as it provides regularization to the model.

The baseline ANI and AIMNet models were trained independently for each of the three charge states of the molecules. For AIMNet-ME and AIMNet-MT, joint training for all charge states was performed, and errors for each charge state were averaged in the loss function. The training was done against 5-fold cross-validation data splits. These five independent models were used to build an ensemble for more accurate predictions, denoted as “ens5” later in the text. All AIMNet model variants, as well as the ANI model, were implemented with PyTorch framework[44] and is available in a public code repository at https://github.com/aiqm/aimnet.

Results and Discussions

A summary of the performance for all four models is presented in Table 1. Vertical ionization potentials (IP) and electron affinities (EA) were computed directly from the corresponding differences of energies of neutral and charged states:

$$IP = E_{cation} - E_{neutral}; EA = E_{neutral} - E_{anion} \quad (5)$$

Root mean square errors (RMSE) of the Ions-12 set provide a measure of the performance of the model with respect to the data points similar to those used for training. On the other hand, errors on Ions-16 can be seen as a more appropriate testbed that is probing generalization capabilities of the model across the unknown chemical and conformational degrees of freedom (i.e., unseen molecules).

Table 1. Root mean square errors (RMSEs) in kcal/mol for individual models and ensemble of 5 models (ens5) on Ions-12 test set and Ions-16 external set. The resulting RMSEs for vertical ionization potentials (IP) and electron affinities (EA) are computed from the respective total energies. The smallest errors (within ~0.1 kcal/mol) for a given quantity across the model set are highlighted in bold.

Model	Test Dataset	Total energy RMSE			IP RMSE	EA RMSE
		Cation	Neutral	Anion		
ANI	Ions-12	8.4	5.1	5.0	9.4	6.9
	Ions-16	10.8	4.4	4.9	11.0	5.9
	Ions-16 (ens5)	10.0	4.0	4.6	10.2	5.3
AIMNet	Ions-12	3.8	3.8	3.5	4.7	4.7
	Ions-16	4.6	2.7	3.3	4.8	3.8
	Ions-16 (ens5)	**4.1**	**2.3**	**2.8**	**4.2**	**3.0**
AIMNet-MT	Ions-12	3.6	3.3	2.9	4.1	3.8
	Ions-16	5.0	2.9	3.1	5.2	3.4
	Ions-16 (ens5)	4.6	**2.5**	**2.7**	4.7	**2.9**
AIMNet-ME	Ions-12	3.9	3.7	3.2	4.6	4.2
	Ions-16	5.6	2.8	3.1	5.7	3.4
	Ions-16 (ens5)	5.2	**2.4**	**2.7**	5.2	**2.9**

While ANI models are known to achieve state-of-the-art performance[17,45] on conformational energies and reaction thermochemistry in drug-like molecules, the problem addressed here is challenging due to the presence of charged species. Similarly to our previous results for neutral molecules,[26] all AIMNet flavors substantially improve upon ANI, especially for the total energy of cations and vertical IPs. The original ANI model does not include explicit long-range interactions. All interactions are described implicitly by the neural network; therefore, the interactions described by the model do not extend beyond the AEV cutoff distance (R_{cut} = 5.2 Å in this work). Since the ANI model performs well on neutral molecules and is completely short sighted, we use it as a baseline for comparison. For this data set, because both extra electrons (in case of anions) and holes (in case of cations) are spatially delocalized, the non-local electrostatics extends beyond the cutoff distance and spatially spans over the full molecule.

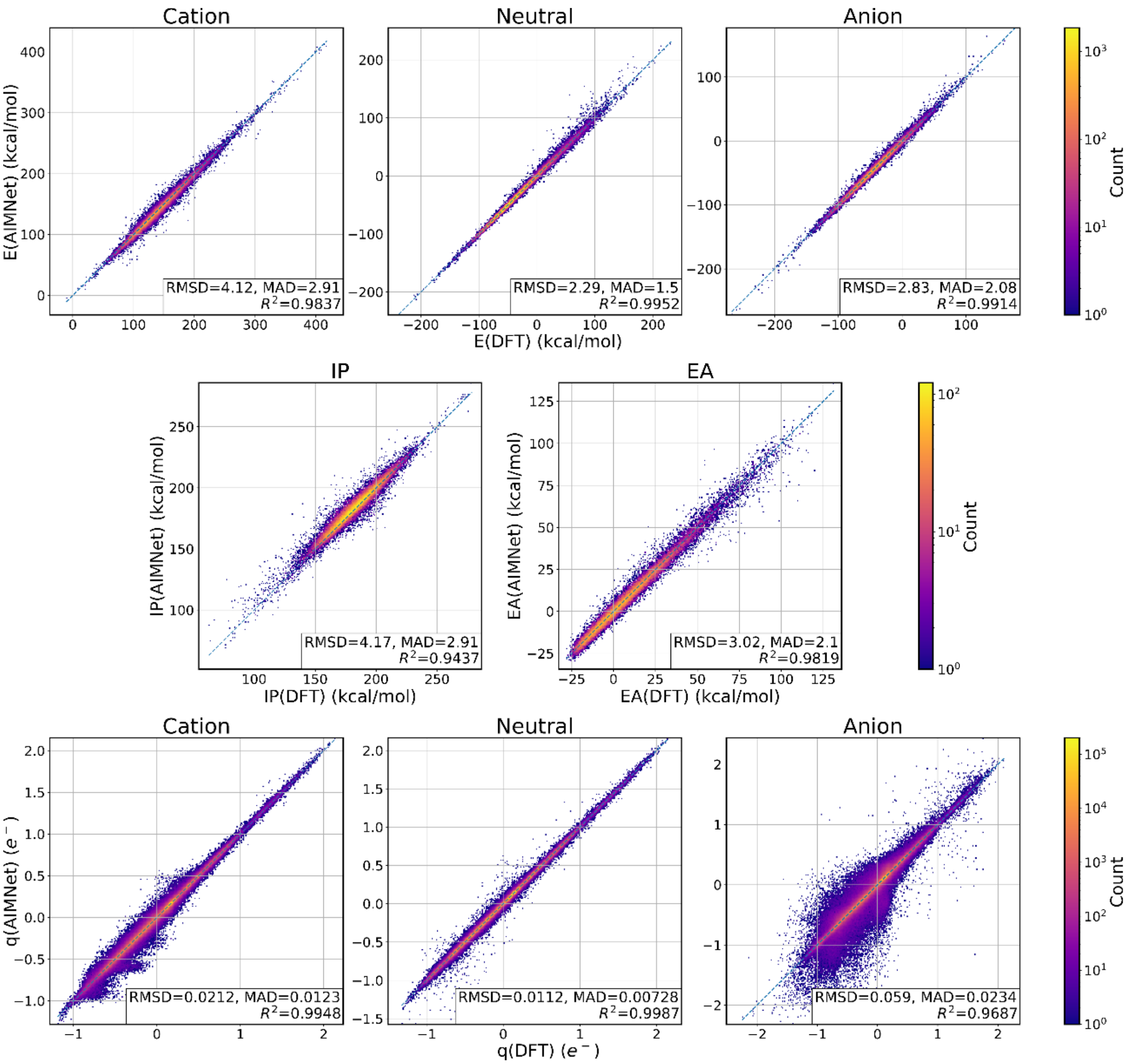

Figure 3. Correlation between DFT PBE0/ma-def2-SVP and AIMNet predictions for total molecular energies (top row), non-equilibrium vertical ionization potentials and electron affinities (middle row) and atomic charges (bottom row) calculated for three charge states for Ions-16 dataset.

The AIMNet model achieves an overall accuracy about 3 kcal/mol for anions and EA and about 4 kcal/mol for cations and IP. Figure 3 provides overall correlation plots for the respective energies and charges. Please see supplementary information for plots for all other models. Note, since regression plots are colored by the density of points on the Log scale, the vast majority of points are on the diagonal line. AIMNet, AIMNet-MT, and AIMNet-ME models consistently provide the same level of performance across the energy range of 400 kcal/mol (~17 eV) without noticeable outliers. All models were able to learn spin-polarized atomic charges up to a "chemical accuracy" of 0.01*e* (electron, elementary charge) as shown in Figure 3 for neutral molecules and cations. The outliers are observed for anions in a realm of negative charges, where the overall RMSD for AIMNet is 0.059 *e.* Table 1 also compares the performance of individual models to the performance of their ensemble prediction (marked as "ens5"). In principle, model ensembling is always desirable and, on average, provide the performance boost by 0.5 kcal/mol for all energy based quantities.

Jointly trained AIMNet-MT and AIMNet-ME data-fusion models only slightly improve the performance for anions and EA by 0.1 kcal/mol. Even though data-fusion did not provide obvious accuracy benefits, such models have significant advantages in practical applications. Namely, only one model needs to be trained and used to predict all properties of interest. AIMNet-ME brings ML and physics-based models one step closer by offering a discrete, physically correct dependence of system energy with respect to a total molecular charge.

To elucidate the importance of iterative "SCF-like" updates, the AIMNet model was evaluated with a different number of passes *t*. AIMNet with $t = 1$ is very similar to the ANI model. The receptive field of the model is roughly equal to the size of the AEV descriptor in ANI; and no updates were made to the AFV vector and atomic embeddings. Figure 4a shows that the aggregated performance of prediction for energies improves with an increasing number of passes *t*. This trend is especially profound for cations. As expected, the accuracy of AIMNet with $t = 1$ is very similar or better compared to the ANI network. The second iteration ($t = 2$) provides the largest improvement in performance for all three states. After $t = 3$, the results are virtually converged.

Therefore, we used $t = 3$ to train all models in this work. These observations for charged molecules are remarkably consistent with results for neutral species.[26]

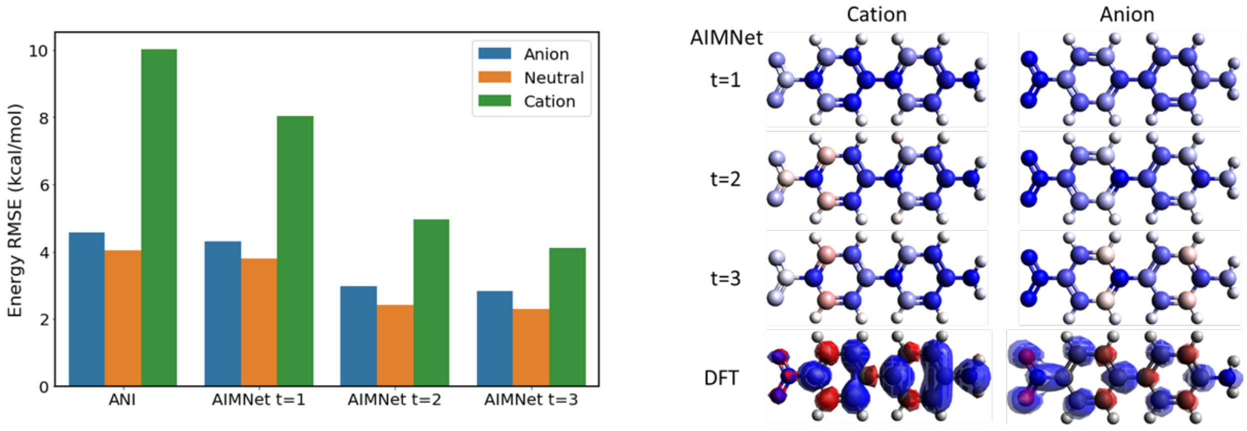

Figure 4. A) Comparison of AIMNet and ANI model performance (RMSE, kcal/mol) for Ions-16 dataset with different t values. All models were trained on exactly the same datasets of total energies of neutral molecules, cations, and anions. B) Automatic redistribution of α and β spin-charges by the AIMNet on 4-amino-4'-nitrobiphenyl molecule with a different number of iterative passes t. For comparison, DFT (PBE0/ma-def2-SVP) spin-density is depicted on the bottom of the panel.

Let us consider 4-amino-4'-nitrobiphenyl molecule as an illustrative example (Figure 4b). This is a prototypical optoelectronic system, where a π-conjugated system separates the electron-donating (NH_2) and accepting (NO_2) groups. These polar moieties underpin an increase in the transition dipole moment upon electronic excitation leading to two-photon absorption. The effect of donor-acceptor substitution is apparent from the ground state calculations of the charge species where electron and hole in cation and anion, respectively, are centered on the substituent groups with strong delocalization across π orbitals of the aromatic rings as illustrated by the DFT calculations (Figure 4b). Capturing such charge delocalization constitutes an extreme challenge to all ML with local geometric descriptors, including AIMNet with $t = 1$. The latter predicts all spin-charges for both cation and anion to be practically the same. The correct behavior could be recovered with either an increase of the cutoff radius for the local environment or some kind of "message-passing". AIMNet with $t = 2$ starts to redistribute charges and predicts alternation of an excess of alpha and beta spins. At $t = 3$, the charge redistribution in the AIMNet model correctly reproduces spin-density wave-like behavior with opposite phases for cation and anion as predicted by DFT (Figure 4b). Notably, 4-amino-4'-nitrobiphenyl molecule was neither part of the training nor test data, exemplifying convergence and reproduction of quantum-mechanical properties through iterative updates.

The previously described neural charge equilibration could be an attractive alternative to popular charge equilibration schemes like EEM,[46] QEq,[47] and QTPIE[48] that use simple physical relationships. They often suffer from transferability issues and might produce unphysical results. To our knowledge, this is a primary example where the ML model provides a consistent and qualitatively correct physical behavior between molecular geometry, energy, integral molecular charge, and partial atomic charges. Other schemes like BP,[15] TensorMol,[18] HIP-NN,[49,50] and PhysNet[21] typically employ auxiliary neural network that predicts atomic charges from a local geometrical descriptor. Electrostatic interactions are computed with Coulomb's law based on those charges. In principle, many effects can be captured by a geometrical descriptor, but it does not depend on the *total charge and spin multiplicity* of the molecule. Following the basic principles of quantum mechanics to incorporate such information successfully, the model should adapt according to changes in the electronic structure, preferably in a *self-consistent* way. This is exemplified here through the case of the AIMNet family of models.

Case study for chemical reactivity and reaction prediction.

As a practical application of AIMNet models, we demonstrate a case study on chemical reactivity and prediction of reaction outcomes. The robust prediction of the products of chemical reactions is of central importance to the chemical sciences. In principle, chemical reactions can be described by the stepwise rearrangement of electrons in molecules, which is also known as a reaction mechanism.[51] Understanding this reaction mechanism is crucial because it provides an atomistic insight into how and why the specific products are formed.

DFT has shown to be a powerful interpretative and computational tool for mechanism elucidation.[52–55] In particular, conceptual DFT (c-DFT) popularized many intuitive chemical concepts like electronegativity (χ) and chemical hardness.[56] In c-DFT, reactive indexes measure the energy (E) change of a system when it is a subject to a perturbation in its number of electrons (N). The foundations of c-DFT were laid by Parr et al.[57] with the identification of the electronic chemical potential μ and hardness η as the Lagrangian multipliers in the Euler equation. In the finite-difference formulation, the these quantities could be derived from EA and IP values as

$$\mu = -\chi = \left(\frac{\partial E}{\partial N}\right) \approx -\frac{1}{2}(IP + EA) \qquad (7)$$

$$\eta = \left(\frac{\partial^2 E}{\partial N^2}\right) \approx -\frac{1}{2}(IP - EA) \qquad (8)$$

The Fukui function *f(r)* is defined as a derivative of the electron density on the total number of electrons in the system. These global and condensed-to-atom local indexes were successfully applied to a variety of problems in chemical reactivity.[58,59] Using finite difference approximation and condensed to atoms representation, Fukui functions for electrophilic (f_a^-), nucleophilic (f_a^+), and radical (f_a^0) reactions are defined as:

$$f_a^- = q_N - q_A;\ f_a^+ = q_C - q_N\ ;\ f_a^0 = \frac{1}{2}(q_C + q_A) \qquad (9)$$

Another useful c-DFT reactivity descriptor is electrophilicity index given by

$$\omega = {}^{\mu^2}\!/_{2\eta} \tag{10}$$

as well as it's condensed to atoms variants for electrophilic (ω_a^e), nucleophilic (ω_a^n) and radical (ω_a^r) attacks:[60]

$$\omega_a^e = \omega f_a^-; \quad \omega_a^n = \omega f_a^+ \, ; \quad \omega_a^r = \omega f_a^0 \tag{11}$$

On the basis of the predicted with AIMNet-MT vertical IPs, EAs, and charges, we could *directly compute* all listed c-DFT indexes. Figure 5 displays the correlation plots for all nine quantities. The AIMNet-ME model achieves an excellent quality of prediction of three global indexes with R^2 ranging from 0.93 to 0.97. Condensed indexes are more challenging to predict, with electrophilic ones being the hardest (R^2 is 0.53 and 0.62). This is related to the overall more substantial errors in the cation energy predictions. Here we would like to emphasize again that *none* of these properties were part of the cost function or training data. The values were derived from the pre-trained neural network and therefore this opens a possibility to a direct modeling *fully bypassing c-DFT calculations and wavefunction analysis.*

We observe relatively poor prediction for the electrophilic and radical Fukui functions and philicity indexes. This is a consequence of larger prediction errors in atomic charges for anions. The most probable reason is the high polarizability of anions and thus more delocalized charge distribution. However, the accuracy of nucleophilic indexes appears to be suitable to make a reliable prediction of reaction outcomes.

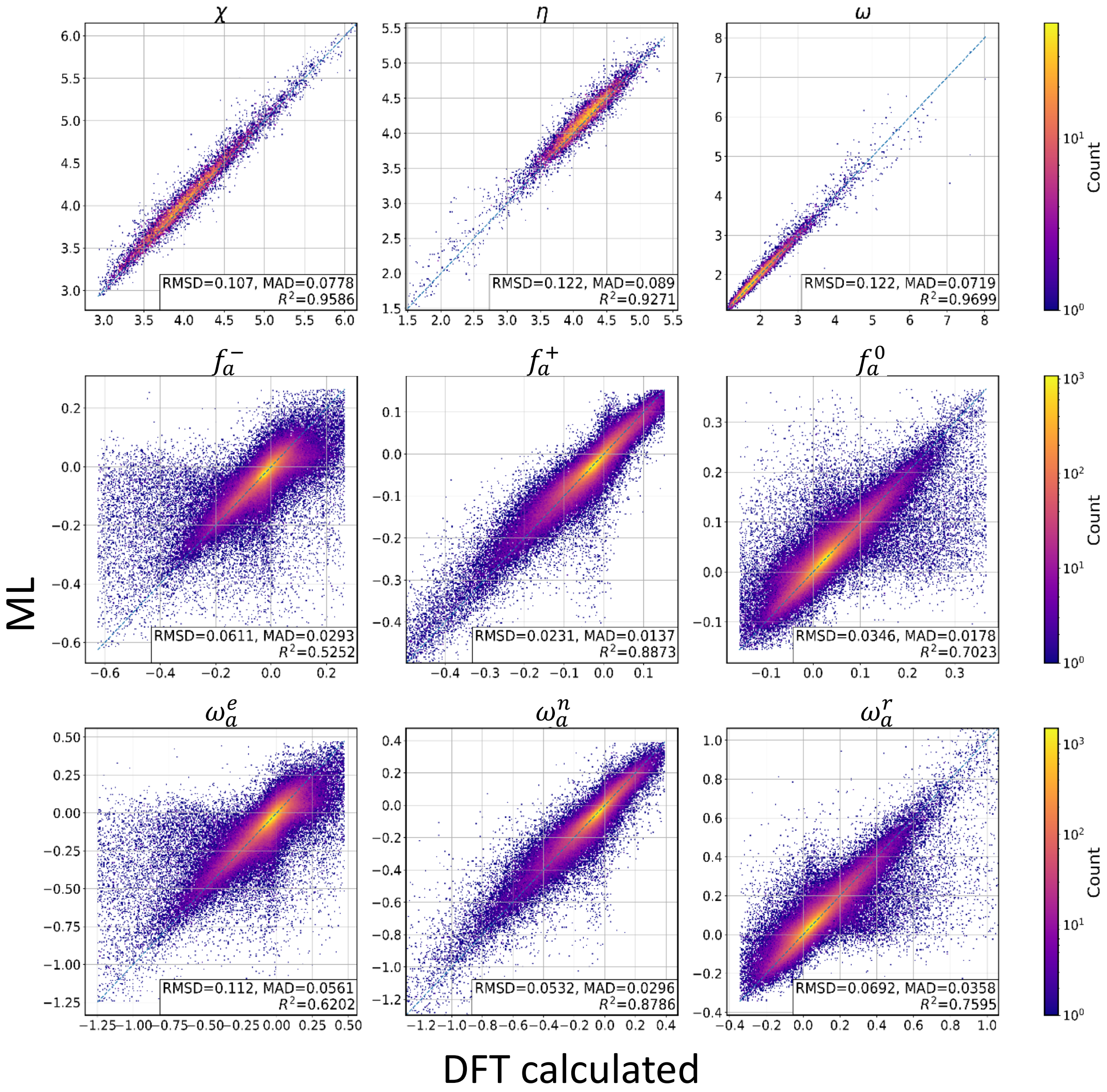

Figure 5. Correlation between DFT PBE0/ma-def2-SVP and AIMNet-MT predictions for electronegativity (χ), chemical hardness (η) and electrophilicity index (ω), Fukui coefficients for nucleophilic (f_a^+), for electrophilic (f_a^-) and radical (f_a^0) attacks and three corresponding condensed philicity indexes (ω_a).

Let us exemplify prediction of site selectivity for aromatic C–H bonds using electrophilic aromatic substitution (EAS) reaction. The EAS reaction is a standard organic transformation. Its mechanism involves the addition of an electrophile to the aromatic ring to form a σ-complex (Wheland intermediate) followed by deprotonation to yield the observed substitution product

(Figure 6). The reactivity and regioselectivity of EAS would generally depend on the ability of the substituents to stabilize or destabilize a σ-complex.

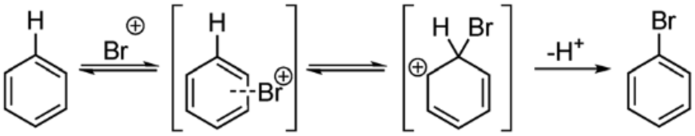

Figure 6. General mechanism of electrophilic aromatic substitution reaction.

Recently EAS attracted significant attention from computational studies due to its importance in late-stage functionalization (LSF) for the drug development process.[61] A direct and numerically very expensive approach to EAS selectivity predictions is to calculate all transition states on the complete path from reactants to products. A popular approach called RegioSQM achieves high site prediction accuracy based on enumeration and calculation of σ-complex with semi-empirical quantum mechanical calculations.[62]

Table 2 lists the accuracy of regioselectivity prediction with recently published methods using data from ref [61]. A random forest (RF) model with DFT TPSSh/Def2-SVP derived descriptors like charges (q), bond orders (BO), Fukui indexes, and solvent accessible surface (SAS) achieves 90% accuracy on the validation data (note different DFT methodology used for this study and for training our DNNs). This model relies on QM calculations of reagents but does not require searching σ-complexes. When QM descriptors are combined with RegioSQM, the RF classifier exhibits an excellent performance of 93%. While the RegioSQM model is accurate, it is slow for high throughput screening. A modest dataset of a few hundred molecules takes about two days to complete on a multicore compute node. Very recently, Weisfeiler-Lehman Neural Network (WLNN) was suggested to predict site selectivity in aromatic C-H functionalization reactions.[63]

This model was trained on 58,000 reactions from the Reaxys database and used RDKit molecular descriptors. WLNN achieves an accuracy approaching 90% for the prediction of EAS regioselectivity.

Table 2. Compilation of results for EAS regioselectivity prediction with different approaches.

Descriptors	**ML Model**	**Validation accuracy**	**Test accuracy**
q, BO, SAS, f_-	RF[1]	0.899	
q, BO, SAS, f_-, RegioSQM	RF[1]	0.931	0.876
Reaxis data, molecular descriptors	Weisfeiler-Lehman Neural Net [2]	0.895	0.836
ω, ω_a^-, *AIM vector*	RF *(present work)*	0.905	0.849

[1] Results from ref. [61]
[2] Results from ref. [63]

We used AIMNet-MT to calculate Fukui coefficients and atomic philicity indexes. We also added the AIM layer as an additional set of descriptors. As we argued before[26] a multimodal knowledge residing inside the AIM layer could be exploited as an information-rich feature representation. The RF classifier trained with AIMNet-MT descriptors displays an excellent performance of 90% on the validation set and 85% on the test set. Therefore, AIMNet models could provide a competitive universal alternative to QM targeting not only geometry minimization but also a prediction of reaction outcomes with several orders of magnitude speedup.

Conclusions

We recently witnessed that machine learning models trained to quantum-mechanical data achieve formidable success in quantitative predictions of ground-state energies and interatomic potentials for common, typically charge-neutral organic molecules. Nevertheless, a quantitative description of complex chemical processes involving reactions, bond breaking, charged species, and radicals remains an outstanding problem for data science. The conceptual challenge is a proper description of spatially delocalized electronic density (which strongly depends on molecular conformation) and accounting for long-range Coulombic interactions stemming from the inhomogeneously distributed charges. These phenomena appear as a consequence of the quantum-mechanical description of delocalized electronic wavefunctions. Consequently, representation of spatially non-local, frequently intensive molecular properties is problematic for common neural nets adapting local geometric descriptors. The recently developed AIMNet neural network architecture addresses this challenge via an iterative message passing-based process, which ultimately captures complex latent relationships across atoms in the molecule.

In the present work, we extended the AIMNet architecture to learn a transferrable potential for organic molecules in three different charge states (neutral, cation-radical and anion-radical species). AIMNet achieves consistent 3-4 kcal/mol accuracy in predicting energies of larger molecules (Ions-16 dataset), even though it was only trained to non-equilibrium DFT data for small molecular species (Ions-12 dataset). In addition to energy, the AIMNet models achieve a state of the art performance in prediction of intensive properties. It demonstrates accuracy ~3 kcal/mol for vertical electron affinities and about 4 kcal/mol for vertical ionization potentials across a broad chemical and conformational space.

The key ingredients that allow the AIMNet family of models to achieve such high level of accuracy are i) multimodal learning, ii) joint information-rich representation of atom in a molecule that is shared across multiple modalities, and iii) iterative "SCF-like" passes for an accounting of long-range interactions. In contrast to the standard geometric descriptors, we have highlighted an importance of incorporating adaptable *electronic* information into ML models. Essentially the

AIMNet method could serve as a neural charge equilibration scheme. As a side benefit, it can be used for a high-quality estimate of reactive indexes based on conceptual DFT and reliable prediction of reaction outcomes. Overall, demonstrated flexible incorporation of quantum mechanical information into AIMNet structure and data fusion (underpinning jointly trained AIMNet-MT and AIMNet-ME models) exemplify a step toward developing a universal single neural net architecture capable of quantitative prediction of multiple properties of interest.

ACKNOWLEDGMENTS

O.I. acknowledges support from NSF CHE-1802789. This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science. The authors acknowledge Extreme Science and Engineering Discovery Environment (XSEDE) award DMR110088, which is supported by NSF grant number ACI-1053575. This research in part was done using resources provided by the Open Science Grid[64,65], which is supported by the award 1148698, and the U.S. DOE Office of Science. We gratefully acknowledge the support and hardware donation from NVIDIA Corporation and express our special gratitude to Mark Berger. The work at Los Alamos National Laboratory (LANL) was supported by the Laboratory Directed Research and Development (LDRD) program and was done in part at the Center for Nonlinear Studies (CNLS) and the Center for Integrated Nanotechnologies (CINT), a U.S. Department of Energy and Office of Basic Energy Sciences user facility, at LANL. J.S.S., R.Z. and O.I. thank CNLS and CINT for their support and hospitality.

REFERENCES

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for

Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2.

(2) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11* (6), 2336–2347. https://doi.org/10.1021/acs.jpclett.9b03664.

(3) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K. R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9* (8), 3404–3419. https://doi.org/10.1021/ct400195d.

(4) Handley, C. M.; Behler, J. Next Generation Interatomic Potentials for Condensed Systems. *Eur. Phys. J. B* **2014**, *87* (7), 152. https://doi.org/10.1140/epjb/e2014-50070-0.

(5) Podryabinkin, E. V; Shapeev, A. V. Active Learning of Linear Interatomic Potentials. **2016**, No. Ml, 1–19. https://doi.org/10.1016/j.commatsci.2017.08.031.

(6) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8* (0), 13890. https://doi.org/10.1038/ncomms13890.

(7) Tsubaki, M.; Mizoguchi, T. Fast and Accurate Molecular Property Prediction: Learning Atomic Interactions and Potentials with Neural Networks. *J. Phys. Chem. Lett.* **2018**, *9* (19), 5733–5741. https://doi.org/10.1021/acs.jpclett.8b01837.

(8) Rowe, P.; Csányi, G.; Alfè, D.; Michaelides, A. Development of a Machine Learning Potential for Graphene. *Phys. Rev. B* **2018**, *97* (5), 1–20. https://doi.org/10.1103/PhysRevB.97.054303.

(9) 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* (5),

058301. https://doi.org/10.1103/PhysRevLett.108.058301.

(10) Basdogan, Y.; Groenenboom, M. C.; Henderson, E.; De, S.; Rempe, S.; Keith, J.; Keith, J. A. Machine Learning Guided Approach for Studying Solvation Environments. **2019**, No. 1. https://doi.org/10.26434/chemrxiv.8292362.v1.

(11) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18* (20), 13754. https://doi.org/10.1039/C6CP00415F.

(12) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.

(13) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2015**, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

(14) Hornik, K. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks* **1991**, *4* (2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T.

(15) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401. https://doi.org/10.1103/PhysRevLett.98.146401.

(16) 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* (4), 3192–3203. https://doi.org/10.1039/C6SC05720A.

(17) 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* (1), 2903. https://doi.org/10.1038/s41467-019-10827-4.

(18) 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* (8), 2261–2269. https://doi.org/10.1039/c7sc04934j.

(19) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148* (24), arXiv:1710.00017. https://doi.org/10.1063/1.5011181.

(20) 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* (24), 241722. https://doi.org/10.1063/1.5019779.

(21) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15* (6), 3678–3693. https://doi.org/10.1021/acs.jctc.9b00181.

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

(23) 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* (24), 241733. https://doi.org/10.1063/1.5023802.

(24) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci. Data* **2020**, *7* (1), 134. https://doi.org/10.1038/s41597-020-0473-z.

(25) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; Ong, S. P. A Performance and Cost Assessment of Machine

Learning Interatomic Potentials. **2019**.

(26) 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* (8), eaav6490. https://doi.org/10.1126/sciadv.aav6490.

(27) Redlich, O. Intensive and Extensive Properties. *J. Chem. Educ.* **1970**, *47* (2), 154. https://doi.org/10.1021/ed047p154.2.

(28) Tolman, R. C. The Measurable Quantities of Physics. *Phys. Rev.* **1917**, *9* (3), 237–253.

(29) Pronobis, W.; Sch, K. T. Capturing Intensive and Extensive DFT / TDDFT Molecular Properties with Machine Learning. **2018**.

(30) Westermayr, J.; Gastegger, M.; Menger, M. F. S. J.; Mai, S.; González, L.; Marquetand, P. Machine Learning Enables Long Time Scale Molecular Photodynamics Simulations. *Chem. Sci.* **2019**, *10* (35), 8100–8107. https://doi.org/10.1039/c9sc01742a.

(31) Chen, W. K.; Liu, X. Y.; Fang, W. H.; Dral, P. O.; Cui, G. Deep Learning for Nonadiabatic Excited-State Dynamics. *J. Phys. Chem. Lett.* **2018**, *9* (23), 6702–6708. https://doi.org/10.1021/acs.jpclett.8b03026.

(32) Dral, P. O.; Barbatti, M.; Thiel, W. Nonadiabatic Excited-State Dynamics with Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9* (19), 5660–5663. https://doi.org/10.1021/acs.jpclett.8b02469.

(33) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**. https://doi.org/10.1021/cr990029p.

(34) Chattaraj, P. K. *Chemical Reactivity Theory*; 2009. https://doi.org/10.1201/9781420065442.

(35) Cohen, M. H.; Wasserman, A. On the Foundations of Chemical Reactivity Theory. *J. Phys.*

Chem. A **2007**, *111* (11), 2229–2242. https://doi.org/10.1021/jp066449h.

(36) Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System. *J. Cheminform.* **2013**, *5* (1), 1–9. https://doi.org/10.1186/1758-2946-5-3.

(37) Grimme, S. A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.* **2014**, *10* (10), 4497–4514. https://doi.org/10.1021/ct500573f.

(38) 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 Comput.* **2019**, *15* (3), 1652–1671. https://doi.org/10.1021/acs.jctc.8b01176.

(39) Landrum, G. RDkit: Open-source Cheminformatics.

(40) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (1), 73–78. https://doi.org/10.1002/wcms.81.

(41) Verstraelen, T.; Vandenbrande, S.; Heidar-Zadeh, F.; Vanduyfhuys, L.; Van Speybroeck, V.; Waroquier, M.; Ayers, P. W. Minimal Basis Iterative Stockholder: Atoms in Molecules for Force-Field Development. *J. Chem. Theory Comput.* **2016**, *12* (8), 3894–3912. https://doi.org/10.1021/acs.jctc.6b00456.

(42) Verstraelen, T.; Tecmer, P.; Heidar-Zadeh, F.; González-Espinoza, C. E. .; Chan, M.; Kim, T. D. .; Boguslawski, K.; Fias, S.; Vandenbrande, S.; Berrocal, D.; Ayers, P. W. HORTON 2.1.0. 2017.

(43) Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. *arXiv:1711.05101*

2017.

(44) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; others. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in neural information processing systems*; 2019; pp 8026–8037.

(45) Devereux, C.; Smith, J.; Davis, K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. **2020**. https://doi.org/10.26434/CHEMRXIV.11819268.V1.

(46) Mortier, W. J.; Van Genechten, K.; Gasteiger, J. Electronegativity Equalization: Application and Parametrization. *J. Am. Chem. Soc.* **1985**, *107* (4), 829–835. https://doi.org/10.1021/ja00290a017.

(47) Rappé, A. K.; Goddard III, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95* (8340), 3358–3363. https://doi.org/10.1021/j100161a070.

(48) Chen, J.; Martínez, T. J. QTPIE: Charge Transfer with Polarization Current Equalization. A Fluctuating Charge Model with Correct Asymptotics. *Chem. Phys. Lett.* **2007**. https://doi.org/10.1016/j.cplett.2007.02.065.

(49) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9* (16), 4495–4501. https://doi.org/10.1021/acs.jpclett.8b01939.

(50) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Transferable Dynamic Molecular Charge Assignment Using Deep

Neural Networks. *J. Chem. Theory Comput.* **2018**, *14* (9), 4687–4698. https://doi.org/10.1021/acs.jctc.8b00524.

(51) Herges, R. Organizing Principle of Complex Reactions and Theory of Coarctate Transition States. *Angewandte Chemie International Edition in English*. 1994. https://doi.org/10.1002/anie.199402551.

(52) Houk, K. N. Frontier Molecular Orbital Theory of Cycloaddition Reactions. *Acc. Chem. Res.* **1975**, *8* (11), 361–369. https://doi.org/10.1021/ar50095a001.

(53) Houk, K.; Paddon-Row, M.; Rondan, N.; Wu, Y.; Brown, F.; Spellmeyer, D.; Metz, J.; Li, Y.; Loncharich, R. Theory and Modeling of Stereoselective Organic Reactions. *Science (80-.).* **1986**, *231* (4742), 1108–1117. https://doi.org/10.1126/science.3945819.

(54) Jones, G. O.; Liu, P.; Houk, K. N.; Buchwald, S. L. Computational Explorations of Mechanisms and Ligand-Directed Selectivities of Copper-Catalyzed Ullmann-Type Reactions. *J. Am. Chem. Soc.* **2010**, *132* (17), 6205–6213. https://doi.org/10.1021/ja100739h.

(55) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**. https://doi.org/10.1038/s41586-019-1384-z.

(56) Ayers, P. W.; Levy, M. Perspective on "Density Functional Approach to the Frontier-Electron Theory of Chemical Reactivity." *Theoretical Chemistry Accounts*. 2000. https://doi.org/10.1007/s002149900093.

(57) Parr, R. G.; Yang, W. Density Functional Approach to the Frontier-Electron Theory of Chemical Reactivity. *J. Am. Chem. Soc.* **1984**. https://doi.org/10.1021/ja00326a036.

(58) Chermette, H. Chemical Reactivity Indexes in Density Functional Theory. *J. Comput. Chem.* **1999**, *20*, 129–154. https://doi.org/10.1002/(SICI)1096-

987X(19990115)20:1<129::AID-JCC13>3.0.CO;2-A.

(59) Chattaraj, P. K. Chemical Reactivity Theory: A Density Functional View. *Chem.Duke.Edu* **2009**. https://doi.org/10.1201/9781420065442.

(60) Chattaraj, P. K.; Maiti, B.; Sarkar, U. Philicity: A Unified Treatment of Chemical Reactivity and Selectivity. *J. Phys. Chem. A* **2003**, *107* (25), 4973–4975. https://doi.org/10.1021/jp034707u.

(61) Tomberg, A.; Johansson, M. J.; Norrby, P. O. A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning. *J. Org. Chem.* **2019**, *84* (8), 4695–4703. https://doi.org/10.1021/acs.joc.8b02270.

(62) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. Fast and Accurate Prediction of the Regioselectivity of Electrophilic Aromatic Substitution Reactions. *Chem. Sci.* **2018**, *9* (3), 660–665. https://doi.org/10.1039/c7sc04156j.

(63) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask Prediction of Site Selectivity in Aromatic C–H Functionalization Reactions. *React. Chem. Eng.* **2020**, *5* (5), 896–902. https://doi.org/10.1039/D0RE00071J.

(64) Sfiligoi, I.; Bradley, D. C.; Holzman, B.; Mhashilkar, P.; Padhi, S.; Würthwein, F. The Pilot Way to Grid Resources Using GlideinWMS. In *2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009*; IEEE, 2009; Vol. 2, pp 428–432. https://doi.org/10.1109/CSIE.2009.950.

(65) Pordes, R.; Petravick, D.; Kramer, B.; Olson, D.; Livny, M.; Roy, A.; Avery, P.; Blackburn, K.; Wenaus, T.; Würthwein, F.; Foster, I.; Gardner, R.; Wilde, M.; Blatecky, A.; McGee, J.; Quick, R. The Open Science Grid. In *Journal of Physics: Conference Series*; IOP Publishing, 2007; Vol. 78, p 012057. https://doi.org/10.1088/1742-6596/78/1/012057.