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 machine learning framework toward open-shell anions and cations. We introduce

AIMNet-NSE (Neural Spin Equilibration) architecture, which being properly trained, could predict atomic and molecular properties for an arbitrary combination of molecular charge and spin multiplicity. This model explores a new dimension of transferability by adding the charge-spin space. The AIMNet-NSE model is capable of reproducing reference QM energies for cations, neutrals, and anions with errors of about 2-3 kcal/mol, compared to the reference QM simulations. The spin-charges have errors ~0.01 electrons for small organic molecules containing nine chemical elements {H, C, N, O, F, Si, P, S and Cl}. The AIMNet-NSE model allows 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 with a speed up to 10^4 molecules per second on a single modern GPU. We show that these descriptors, along with learned atomic representations, could be used to model chemical reactivity through an example of regioselectivity 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 $\sim O(N^3)$ 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.[3–8]

Deep neural networks (DNN)[9,10] are a particular class of ML algorithms proven to be universal function approximators.[11] 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)[12] and its modifications ANI[13,14] and TensorMol[15] 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,[16] DTNN,[4] SchNet,[17] and PhysNet[18] use non-invariant radial symmetry functions or interatomic distances and iteratively construct a representation of the atomic environment through message-passing techniques.[19]

The ANAKIN-ME (ANI) method[13,20] is one example of a technique for building transferable DNN-based molecular potentials. The key components of ANI models are the diverse training dataset[21] and BP type descriptors[12] with modified symmetry functions.[13] 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.[21] 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.[14] 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.[8]

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.[22] The notation has been introduced by Tolman over one hundred years ago.[23] Only a few reports have attempted to use ML for *intensive* properties.[24–29] independent of the system size, which pose a challenge ML techniques due to spatial non-locality and long-range interactions.

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.[26] 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[24,30] 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.[31,32] c-DFT, or Chemical Reactivity Theory, is a powerful tool for the prediction, analysis, and interpretation of chemical reactions.[33,34] 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)[12] 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 $\boldsymbol{g_i^{(r)}}$ and $\boldsymbol{g_i^{(a)}}$ 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:

$$\boldsymbol{G_i} = \left\{ \boldsymbol{g_i^{(r)\top} A^{(r)}}, \boldsymbol{g^{(a)\top} A^{(a)}} \right\} \quad (1)$$

This definition of the HDNN models suffer from the "curse of dimensionality" problem. Namely, the size of ***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 ***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:

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

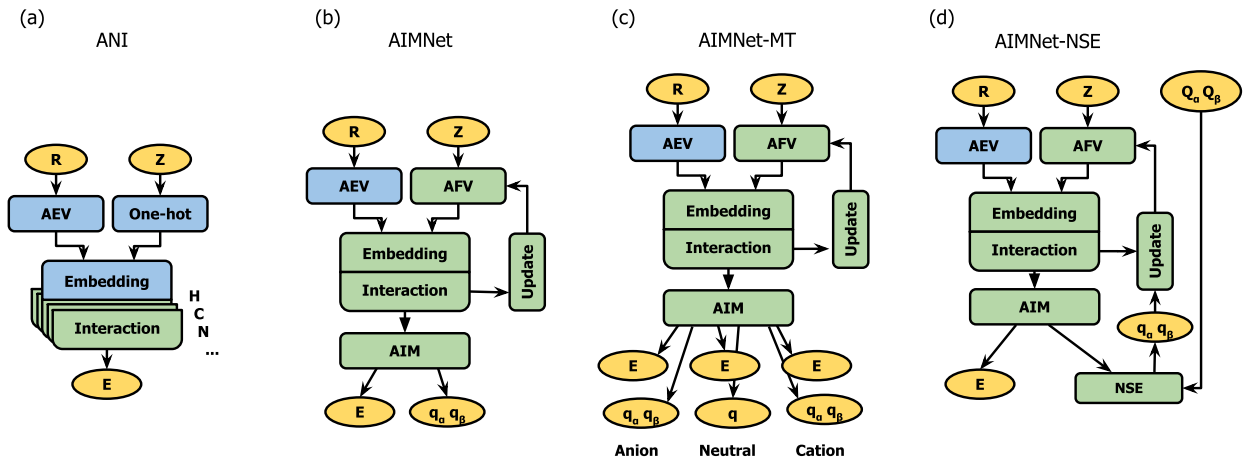

Figure 1. Neural network architectures explored in this work. Models from literature: a) ANI[13], b) AIMNet[8]; Here each model is separately trained for neutral species, cations and ions. Models introduced in this work: c) AIMNet-MT: a multitask model jointly trained on all data which concurrently predicts energies and charges for neutral species as well as cations and ions; and d) AIMNet-NSE, a Neural Charge Equilibration model which is capable to redistribute spin-polarized atomic charges according to a given molecular spin-charges, and predicts energy for specified (arbitrary) spin state of the molecule. The yellow blocks show input data (coordinates **R**, atomic numbers **Z** and total molecular spin 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 dimensionality issue 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 $\mathcal{F}$:

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

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.

The AIMNet and ANI models does not use total molecular charge and therefore could not discriminate between different charge states of the same conformer. The straightforward way to obtain reasonable predictions for is to train separate models for neutral, anionic and cationic species. Since AIMNet model works well in multi-task regime,[8] we also build the AIMNet model that simultaneously predicts energies and spin-polarized atomic charges with multiple output heads from same **AIM** layer for predefined set of charge states (AIMNet-MT, Fig. 1c). 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 this paper we introduce an extension to the AIMNet architecture which allows the model to predict energy, properties and partial atomic charges for a specified state based on total molecular charge and spin multiplicity (or, alternatively, total α and β spin-charges) given as input for the model. The key component of the new model is Neural Spin-charge Equilibration unit (NSE, Fig. 1d), which makes prediction of partial spin-polarized atomic charges $\tilde{q}^s$ and atomic

weight factors f^s (conseptually related to atomic Fukui functions, ∂q/∂Q) from the **AIM** layer using fully-connected NN output head. The factors f^s are used to re-distribute atomic spin-charges such as their sum is equal to the specified total molecular spin-charges:

$$q_i^s = \tilde{q}_i^s + \frac{f_i^s}{\sum_{j=1}^N f_j^s}\left(Q^s - \sum_{j=1}^N \tilde{q}_j^s\right) \quad (4)$$

where index *s* corresponds to spin-component of the charge density, $\tilde{q}$ and q are initial and re-normalized charges, *N* is number of atoms and Q total is the total charge of the molecule. The consequent **Update** block injects normalized atomic charges into the **AFV** vector. This way, during the next AIMNet iteration, the information about charge distribution will be used in the **Embedding** block. We should note, that for AIMNet and AIMNet-MT models sum of atomic charges integer, but rather is very close to the total integer molecular charge and reflects uncertainty in charge prediction. However, by design of the AIMNet-NSE model, the charges are conserved and add up to the total charge.

Dataset construction. For the training dataset, we randomly selected about 200k neutral molecules from the UNICHEM database[35] 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 thermal fluctuations allow to model sufficiently well conformational structures of molecules near equilibrium as was exploited in previous reports.[37,38]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[39]) as an efficient method to construct system-specific and charge-specific mechanistic potential for a molecule. We relied on the GFN2-xTB[40] tight-binding model to obtain minimum conformation, force constants, charges, and bond orders that are needed for the QMDFF model.

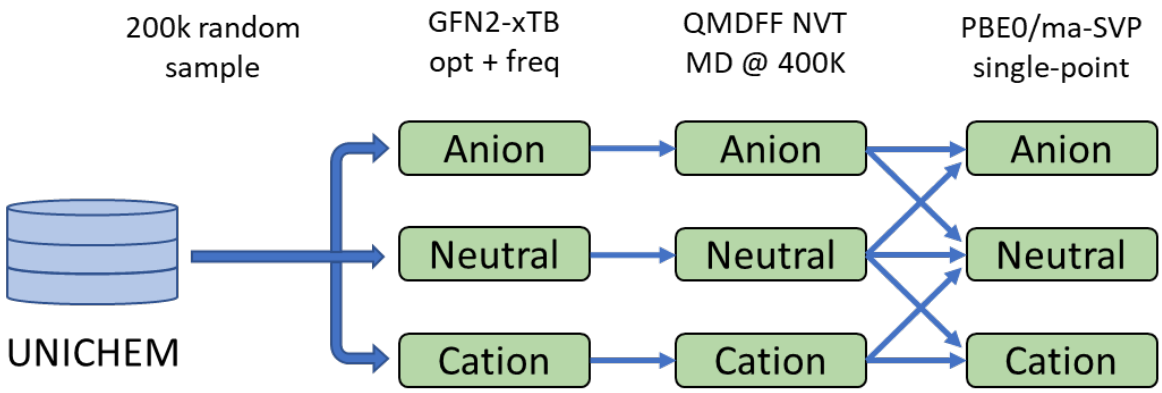

Figure 2. The overall workflow targeting dataset generation for the energetics of neutral and charged molecular species.

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[41] 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 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. For DFT calculations we selected PBE0/ma-def2-SVP level of theory as a reasonable compromise between accuracy and computational expenses. PBE0 is a non-empirical hybrid DFT that is widely used to compute molecular properties. Exact exchange and diffuse functions in the basis set are need in order to describe anionic species. All DFT calculations were performed using ORCA 4.0 package.[42] Atomic spin-polarized charges were calculates NBO-7 software package[43] for DFT wavefunction.s

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. Please see supplementary information (Table S1, Figures S1, S2) for more details. 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.

For the further evaluation of model performance, transferability, and extensibility we compiled a dataset which should be close to real-world application. We randomly selected 800 of organic molecules from ChEMBL database[44,45] with 13-20 non-hydrogen atoms, 100 per molecular size. Neutral state of each molecule was optimized with B97-3c composite DFT method[46], and on that geometry energy calculation were performed for anion and cation radicals. We call this dataset as ChEMBL-20 and it covers equilibrium conformations of "drug-like" molecules.

Training protocol.

The ANI model and AIMNet variants were trained using minibatch gradient descent powered by the Adam optimizer.[47] For training performance considerations, all minibatches were composed of molecules with the same number of atoms, to avoid padding. Proper data feed shuffling was achieved with mutli-GPU Data-parallel approach: gradients on model weights were averaged after 8 random batches were evaluated in parallel. The effective combined batch size was 2048. Training was performed on 8 Nvidia V100 GPUs, with about 200s for AIMNet-MT model and 130s for AIMNet-NSE model per epoch of Ions-12 dataset with 6.4M data points. With reduce-on-plateau learning rate schedule training typically converged within 400-500 epochs.

The training objective was minimization of weighted multi-target mean squared error (MSE) loss function with included errors in energy and charge predictions. The AIMNet architecture shares weights of Embedding, Interaction blocks and fully-connected output heads for all "SCF-like" iterative passes. The models were trained with 3 passes. The outputs from each pass were included into weight function, except for AIMNet-NSE model. Due to architecture of the AIMNet-NSE model during first pass it makes predictions without use of information about total spin charge. Therefore, for this model only, outputs from the two last passes were included in the loss function. 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. Additional details about the loss function are given in the SI.

The baseline ANI and AIMNet models were trained independently for each of the three charge states of the molecules. For AIMNet-MT and AIMNet-NSE, joint training for all charge states was performed, and errors for each charge state were included 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[48] and is available in a public code repository at https://github.com/isayevlab/aimnet-nse.

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:

$$IP = E_{cation} - E_{neutral}; EA = E_{neutral} - E_{anion} \quad (6)$$

The prediction errors are evaluated on the Ions-12 (up to 12 non-H atoms) dataset which provides 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 (13-16 non-H atoms) 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). Further, we evaluate performance of the models on the dataset of equilibrium conformations of neutral drug-like molecules ChEMBL-20 (13-20 non-H atoms) as a realistic example application of the model. We report root-mean-square errors (RMSE), rather than more popular in the field[5,17,49] mean absolute errors (MAE). MAE is less sensitive to severe prediction errors and could often mislead about the generalization capabilities of the models.

Table 1. Root mean square errors (RMSEs) in kcal/mol for individual models and ensemble of 5 models (ens5) on validation subset of Ions-12 dataset, Ions-16 and ChEMBL-20 external test sets. The resulting RMSEs for vertical ionization potentials (IP) and electron affinities (EA) are computed from the respective total energies.

Model	Test Dataset	Total energy RMSE			IP	EA
		Cation	Neutral	Anion	RMSE	RMSE
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	4.1	3.7	3.0	4.7	4.4
	Ions-16	6.3	3.2	3.4	6.5	4.0
	Ions-16 (ens5)	5.3	**2.6**	**2.8**	5.3	**3.1**
	ChEMBL-20 (ens5)	12.8	5.3	6.0	9.2	2.9
AIMNet-MT	Ions-12	3.5	3.4	2.8	4.1	3.9
	Ions-16	5.4	3.0	3.2	5.5	3.5
	Ions-16 (ens5)	4.9	2.5	2.7	5.0	**3.0**
	ChEMBL-20 (ens5)	13.0	4.3	5.4	10.3	3.0
AIMNet-NSE	Ions-12	3.6	3.4	2.9	4.1	3.9
	Ions-16	3.9	3.1	3.1	4.1	3.6
	Ions-16 (ens5)	**3.4**	**2.5**	**2.6**	**3.5**	**3.0**
	ChEMBL-20 (ens5)	**4.0**	**3.4**	**3.8**	**2.7**	**2.4**

While ANI models are known to achieve state-of-the-art performance[14,50] 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,[8] 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 and has no capability to perform charge equilibration either explicitly or implicitly, we use it as a baseline for comparison. 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 molecule.

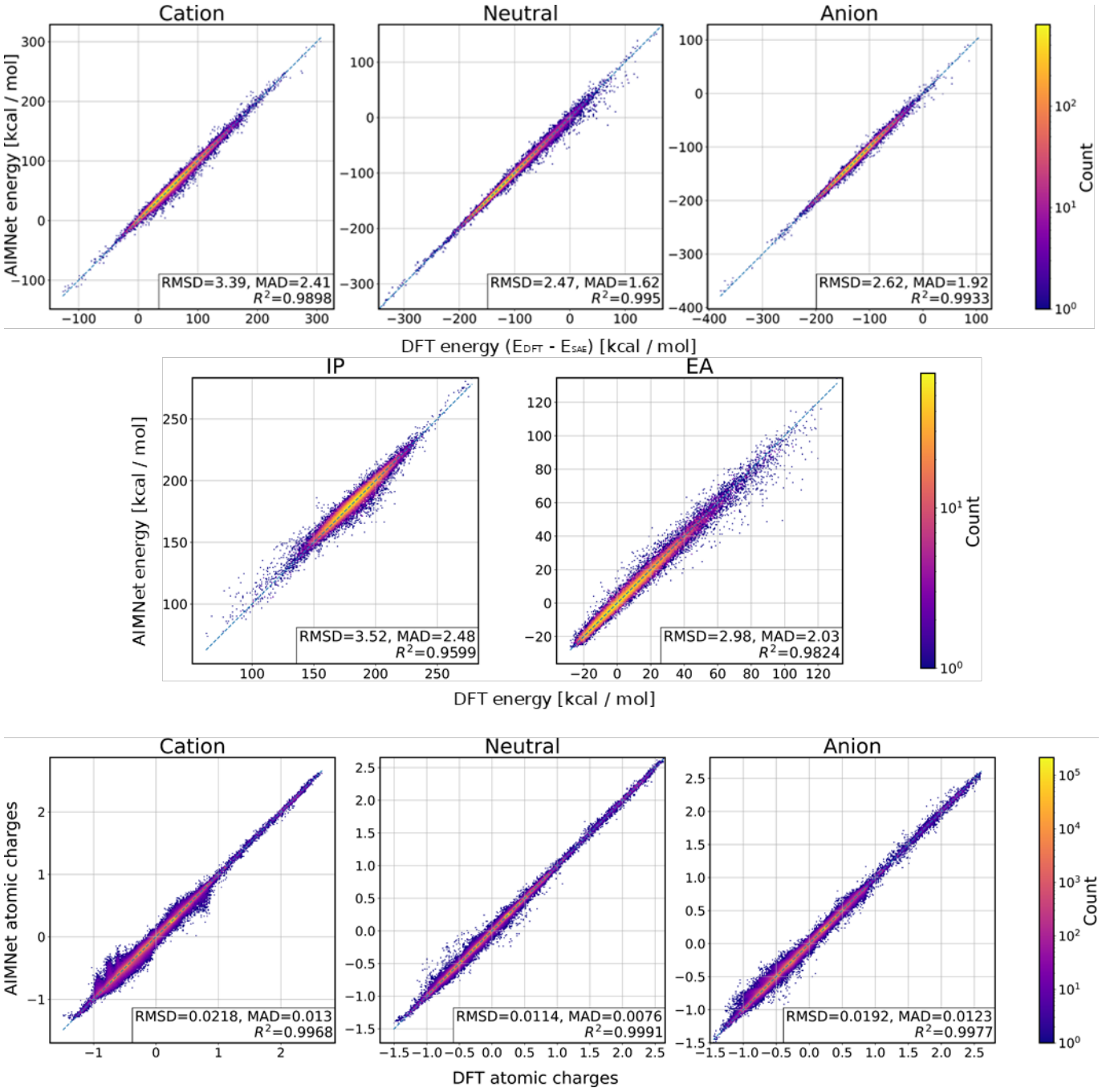

Figure 3. Correlation between DFT PBE0/ma-def2-SVP and AIMNet-NSE predictions for total molecular energies (top row), non-equilibrium vertical ionization potentials and electron affinities (middle row) and NBO atomic charges (bottom row) calculated for three charge states for Ions-16

dataset. DFT total energies were shifted by sum of atomic self-energies (E_{SAE}) to allow comparison for molecules with different composition. Element-specific E_{SAE} calculated using linear regression, correspond to average atomic energies in the entrie training dataset that include all charge states. See also Figure S4 and S5 for all models results.

While the AIMNet and AIMNet-MT models show reasonable accuracy for neutral and anionic species, the errors for cations are few times larger, especially for ChEMBL dataset. This indicates the shortcoming in extensibility of implicit charge equilibration with "SCF-like" passes. Overall, data-fused AIMNet-MT model performs marginally better then separate AIMNet models for each charge state. Contrary, the AIMNet-NSE model with explicit charge equilibration shows consistent performance across charge states and molecule sizes, both for near and off-equilibrium conformers. The RMSE errors on IP and EA values are approaching to 0.1 eV for optimized structures and to 0.15 eV for off-equilibrium geometries. Figure 3 provides overall correlation plots for energies and charges as predicted by AIMNet-NSE model for Ions-16 dataset. 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. The AIMNet-NSE models consistently provide the same level of performance across the energy range of 400 kcal/mol (~17 eV) without noticeable outliers. The model is able to learn atomic charges up to 0.01e (electron, elementary charge) for neutral molecules and 0.02e for ions as shown in Figure 3 (See also Table S2. 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.

The AIMNet-NSE model has a superb utility for high throughput applications. In this sense, it is interesting to compare this model with excellent semi-empirical IPEA-xTB method[51]. The IPEA-xTB is a re-parametrization of GFN-XTB Hamiltonian to predict EA and IP values of

organic and inorganic molecules. The re-parametrization aimed to reproduce PW6B95/def2-TZVPD results. The IPEA-xTB method was successfully used to make accurate predictions of electron ionization mass spectra[51] and for high-throughput screening of polymers.[52,53] For the medium-sized organic molecules, AIMNet-NSE model raises accuracy/computational performance ratio to the a new level. For the ChEMBL-20 dataset, the RMSE of IPEA-xTB EA and IP vs PBE0/ma-def2-SVP are 4.6 and 10.6 kcal/mol, compared to AIMNet-NSE errors of 2.7 and 2.4 kcal/mol, respectively. Therefore, being at least two orders of magnitude faster, the AIMNet-NSE model could be two to four times more accurate.

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.[8]

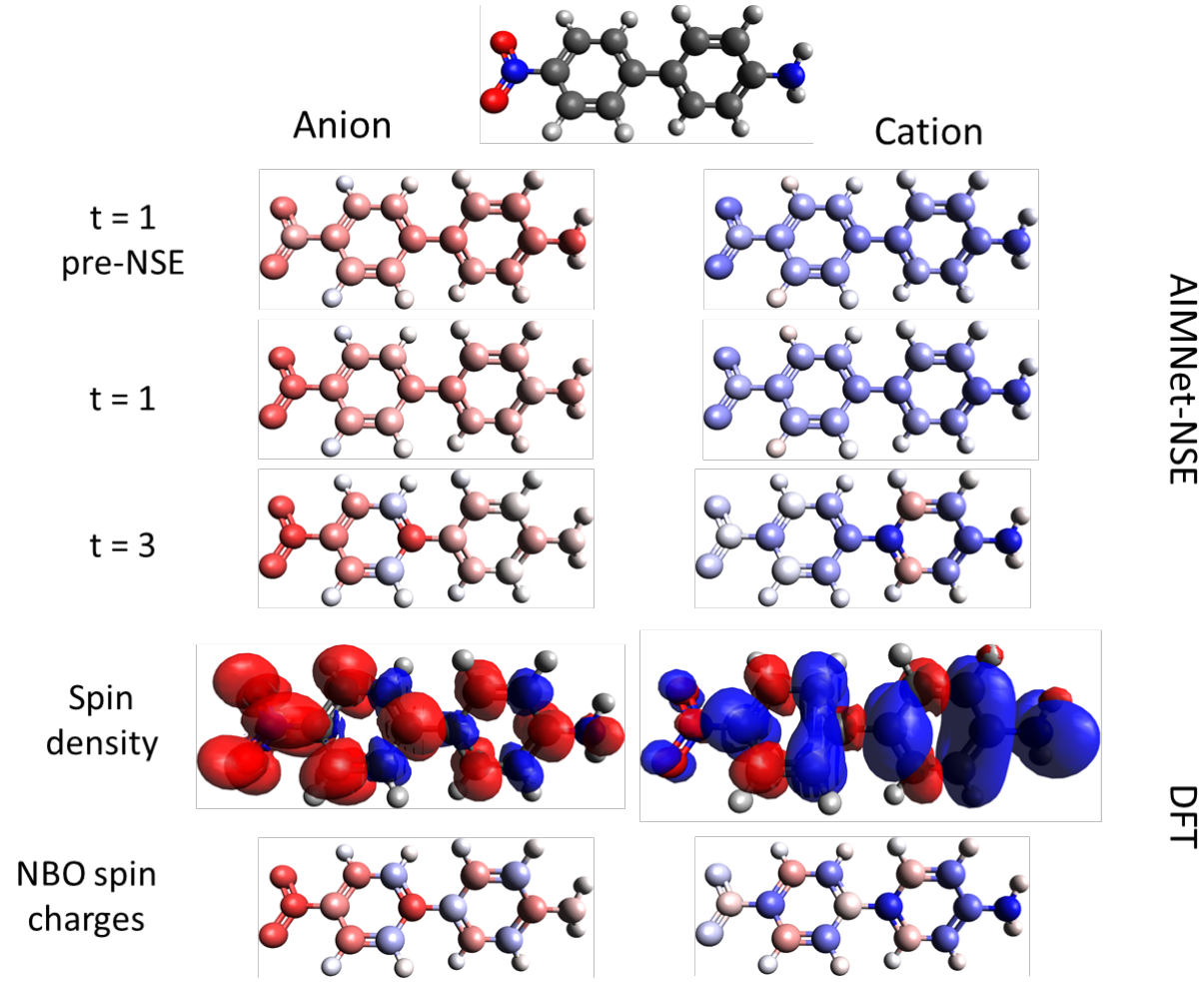

Figure 4. Spin charge re-distribution due to Neural Spin Equilibration for electron attached and detached states of 4-amino-4'-nitrobiphenyl molecule. For anion colors correspond to spin electron charge or density (α - β), while for cation to spin hole density (β - α), with red color corresponding to negative spin-charge. For comparison, DFT (PBE0/ma-def2-SVP) spin-density and charges is depicted on the bottom of the panel.

For 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 shifted towards the substituent

groups with strong delocalization across π orbitals of the aromatic rings. Figure 4 illustrated the charge equilibration procedure in AIMNet-NSE models and compares it to DFT results. During first pass, before charge normalization, the predicted densities are the same for anion and cation (note inverse color codes for anion and cation on the Fig. 4), but after weighted normalization spin charge density is already slightly shifted towards nitro group in anion and amino group in cation. At the same time spin charges on the hydrogen atoms does not change. After three iterations the AIMNet-NSE model correctly reproduces spin-density wave-like behavior with opposite phases for cation and anion as predicted by DFT. There is no sign alternation for spin charge for 4, 4' positions, however the absolute value of spin charge difference for these atoms is high. Overall, the AIMNet-NSE model predicts spin charges for non-hydrogen atoms of this molecule with MAE 0.03e for anion and 0.02e for anion. Notably, 4-amino-4'-nitrobiphenyl molecule was neither part of the training nor validation data, exemplifying convergence and reproduction of quantum-mechanical properties through iterative updates.

In AIMNet-NSE, the physical meaning of the weights f (see eq, 4) is related to atomic Fukui functions, $\partial q_i/\partial Q$, e.g. how much would atomic charge q_i change with the change of total charge Q. In practice, the model would assign higher values of f to the atoms which trend to have different charges in different charge states of the molecule, for example, to aromatic and hereto atoms. The value of f also reflects the uncertainty in charge distribution predicted by neural network. Somewhat related approach for weighted charge re-normalization was used previously.[54] It was based on charge prediction uncertainty estimated with ensemble of random forests, however without noticeable improvement in charge prediction accuracy. Our neural spin-charge equilibration method provides simple and affordable alternative to other ML charge equilibration approaches[55–57] based on QEq method which finds charge distribution by minimization molecular

Coulomb energy. While the QEq solution impose physics-based constraints for the obtained charge distribution, it is limited by the approximate form of Coulomb integral and could be computationally demanding due to matrix inversion operation.

The described neural charge equilibration could be an attractive alternative to popular charge equilibration schemes like EEM,[58] QEq,[59] and QTPIE[60] 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. Upon submitting this manuscript we learned about work by Xie,[61] where ML model built to predict energy as function of electron populations in prototypical LiH clusters. Other schemes like BP,[12] TensorMol,[15] HIP-NN,[62,63] and PhysNet[18] 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-NSE model.

Case study for chemical reactivity and reaction prediction.

As a practical application of AIMNet-NSE model, 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.[64] 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.[65–68] In particular, conceptual DFT (c-DFT) popularized many intuitive chemical concepts like electronegativity (χ) and chemical hardness.[69] 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.[70] 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

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

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

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.[32,71] 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:

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

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

$$\omega = {\mu^2}/{2\eta} \quad (10)$$

as well as it's condensed to atoms variants for electrophilic (ω_a^-), nucleophilic (ω_a^+) and radical ($\omega_a^\pm$) attacks:[72]

$$\omega_a^- = \omega f_a^-; \quad \omega_a^+ = \omega f_a^+ ; \quad \omega_a^\pm = \omega f_a^\pm \tag{11}$$

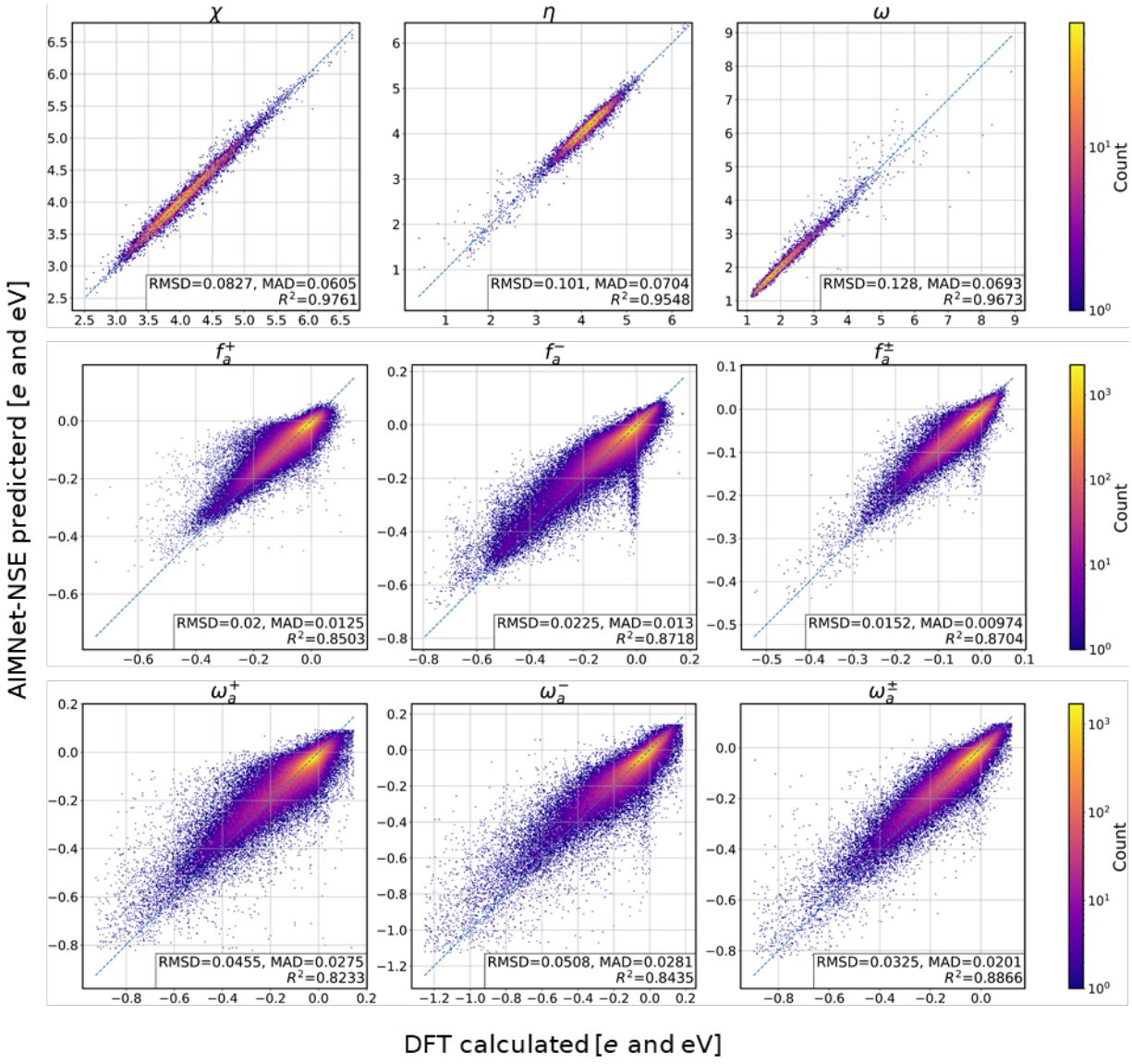

Figure 5. Correlation between DFT PBE0/ma-def2-SVP and AIMNet-NSE predictions for electronegativity (χ), chemical hardness (η) and electrophilicity index (ω), Fukui coefficients for nucleophilic (f_a^+), for electrophilic (f_a^-) and radical (f_a^0) attacks and three corresponding condensed philicity indexes (ω_a) for Ions-16 dataset.

On the basis of the predicted with AIMNet-NSE 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-NSE 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 philicity index (ω_a^+) being the hardest (R^2 is 0.82). This is related to the overall larger errors in the cation energy predictions. Here we would like to emphasize again that <u>*none*</u> 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.* The accuracy of condensed indexes appears to be suitable to make a reliable prediction of reaction outcomes.

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.

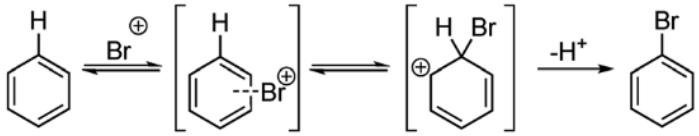

Figure 6. General mechanism of electrophilic aromatic substitution reaction.

Recently EAS attracted significant attention from computational studies due to its importance in late-stage functionalization (LSF) for the drug development process.[73] 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.[74]

Table 2 lists the accuracy of regioselectivity prediction with recently published methods using data from ref [73]. 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.[75] 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)*	0.906	0.850

[1] Results from ref. [73]
[2] Results from ref. [75]

We used AIMNet-NSE to calculate Fukui coefficients and atomic philicity indexes. We also added the AIM layer of the query atom in cation-radical form of the molecule as an additional set of descriptors. The size of AIM layer is smaller (144 elements) then the training dataset size (602 data points). The use of cross-validation scores and random forest method generally mitigates any overfitting issues. As we argued before[8] a multimodal knowledge residing inside the AIM layer could be exploited as an information-rich feature representation. The RF classifier trained with AIMNet-NSE descriptors displays an excellent performance of 90% on the validation set and 85% on the test set. While obtained predictions for the electrophilic aromatic substitution reaction are marginally better than previously reported values, our model achieve six orders of magnitude computational speedup since no quantum mechanical simulations are necessary.

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 introduced the AIMNet-NSE architecture to learn a transferrable potential for organic molecules in arbitrary charge states. For neutral, cation-radical and anion-radical species, the AIMNet-NSE achieves consistent 3-4 kcal/mol accuracy in predicting energies of larger molecules (13-20 non-H atoms), even though it was only trained small molecular up to 12 non-H atoms. In addition to energy, the AIMNet-NSE model achieve a state of the art performance in prediction of intensive properties. It demonstrates accuracy about 0.10-0.15 eV for vertical electron affinities and ionization potentials across a broad chemical and conformational space.

The key ingredients that allow the AIMNet-NSE model 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) Neural Spin-charge Equilibration (NSE) block inside the neural network. In contrast to the standard geometric descriptors, we have highlighted an importance of incorporating adaptable *electronic* information into ML models. Essentially the AIMNet-NSE model serves as a charge equilibration scheme. AIMNet-NSE 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 and spin states.

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 exemplify a step toward developing a universal single neural net architecture capable of quantitative prediction of multiple properties of interest. As we show in our case studies the the

AIMNet-NSE model appears as a fast and reliable method to compute multiple properties like ionization potential, electron affinity, spin polarized charges and a wide variety of Conceptual DFT indexes. It potentially emerges as a drop-in replacement calculator in a myriad of potential applications where high computational accuracy and throughput are required.

Data availability
All test data used in this study are publicly available from the project GitHub.

Code availability
The code to reproduce this study is available on GitHub at https://github.com/isayevlab/aimnet-nse

ACKNOWLEDGMENTS

O.I. acknowledges support from NSF CHE-1802789 and CHE-2041108. 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 is part of the Frontera computing project at the Texas Advanced Computing Center. Frontera is made possible by the National Science Foundation award OAC-1818253. This research in part was done using resources provided by the Open Science Grid[76,77], 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 Jonathan Lefman. 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) 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.

(4) 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.

(5) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3* (5), e1603015. https://doi.org/10.1126/sciadv.1603015.

(6) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104* (13), 136403. https://doi.org/10.1103/PhysRevLett.104.136403.

(7) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3* (12),

e1701816. https://doi.org/10.1126/sciadv.1701816.

(8) 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.

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

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

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

(20) 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.

(21) 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.

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

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

(24) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine Learning of Molecular Electronic

Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15* (9), 095003. https://doi.org/10.1088/1367-2630/15/9/095003.

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

(26) 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.

(27) 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.

(28) 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.

(29) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at near Chemical Accuracy with Sub-Second Computational Cost. *Nat. Commun.* **2020**, *11* (1), 2328. https://doi.org/10.1038/s41467-020-16201-z.

(30) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett.* **2020**, 3828–3834. https://doi.org/10.1021/acs.jpclett.0c00527.

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

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

(33) 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.

(34) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. https://doi.org/10.1016/j.chempr.2020.02.017.

(35) 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.

(36) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1* (1), 140022. https://doi.org/10.1038/sdata.2014.22.

(37) 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.

(38) 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.

(39) 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.

(40) 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.

(41) Landrum, G.; others. RDKit: Open-Source Cheminformatics. **2006**.

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

(43) Glendening, E. D.; Badenhoop, J, K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Karafiloglou, P.; Landis, C. R.; Weinhold, F. NBO 7.0. Theoretical Chemistry Institute, University of Wisconsin: Madison 2018.

(44) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43* (W1), W612–W620. https://doi.org/10.1093/nar/gkv352.

(45) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(46) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A Revised Low-Cost Variant of the B97-D Density Functional Method. *J. Chem. Phys.* **2018**, *148* (6), 064104. https://doi.org/10.1063/1.5012601.

(47) Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. *arXiv:1711.05101* **2017**.

(48) 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.

(49) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Glowacki, D. R.; von Lilienfeld, O. A. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. **2019**, 1–16.

(50) 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.

(51) Ásgeirsson, V.; Bauer, C. A.; Grimme, S. Quantum Chemical Calculation of Electron Ionization Mass Spectra for General Organic and Inorganic Molecules. *Chem. Sci.* **2017**, *8* (7), 4879–4895. https://doi.org/10.1039/C7SC00601B.

(52) Heath-Apostolopoulos, I.; Wilbraham, L.; Zwijnenburg, M. A. Computational High-Throughput Screening of Polymeric Photocatalysts: Exploring the Effect of Composition, Sequence Isomerism and Conformational Degrees of Freedom. *Faraday Discuss.* **2019**, *215*, 98–110. https://doi.org/10.1039/C8FD00171E.

(53) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers. *J. Chem. Inf. Model.* **2018**, *58* (12), 2450–2459. https://doi.org/10.1021/acs.jcim.8b00256.

(54) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58* (3), 579–590. https://doi.org/10.1021/acs.jcim.7b00663.

(55) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic Potentials for Ionic

Systems with Density Functional Accuracy Based on Charge Densities Obtained by a Neural Network. *Phys. Rev. B* **2015**, *92* (4), 045131. https://doi.org/10.1103/PhysRevB.92.045131.

(56) Faraji, S.; Ghasemi, S. A.; Rostami, S.; Rasoulkhani, R.; Schaefer, B.; Goedecker, S.; Amsler, M. High Accuracy and Transferability of a Neural Network Potential through Charge Equilibration for Calcium Fluoride. *Phys. Rev. B* **2017**, *95* (10), 1–11. https://doi.org/10.1103/PhysRevB.95.104105.

(57) 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* (1), 398. https://doi.org/10.1038/s41467-020-20427-2.

(58) 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.

(59) 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.

(60) 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.

(61) Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *J. Chem. Theory Comput.* **2020**, *16* (7), 4256–4270. https://doi.org/10.1021/acs.jctc.0c00217.

(62) 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.

(63) 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.

(64) 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.

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

(66) 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.

(67) 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.

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

(69) 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.

(70) 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.

(71) 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.

(72) 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.

(73) 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.

(74) 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.

(75) 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.

(76) 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.

(77) 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.