JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

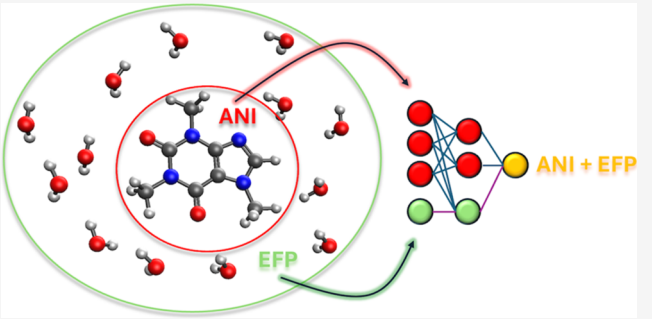
ANI/EFP: Modeling Long-Range Interactions in ANI Neural Network with Effective Fragment Potentials

Shahed Haghiri, Claudia Viquez Rojas, Sriram Bhat, Olexandr Isayev, and Lyudmila Slipchenko*

Cite This: https://doi.org/10.1021/acs.jctc.4c01052 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Deep learning Neural Networks (NN) have been developed in the field of molecular modeling for the purpose of circumventing the high computational cost of quantum-mechanical calculations while rivaling their accuracies. Although these networks have found great success, they generally lack the ability to accurately describe long-range interactions, which makes them unusable for extended molecular systems. Herein, we provide a method for partially retraining the deep learning general-use neural network ANI, in which the long-range interactions are represented via atomic electrostatic potentials. The electrostatic potentials, generated with polarizable effective fragment potentials (EFP), are used as an additional input feature for the network. This new ANI/EFP network can predict solute−solvent interaction energies on a trained data set with a kcal/mol accuracy. It also shows promise in predicting the interaction energies of a solute in solvent environments that have not been included in a training data set. The proposed protocol can be taken as an example and further developed, leading to highly accurate and transferable neural network potentials capable of handling long-range interactions and extended molecular systems.



1. INTRODUCTION

Machine learning has proven to be a useful problem-solving tool in almost any field. In computational chemistry, scientists have developed neural networks (NN) to replace computationally expensive quantum-mechanical calculations.[1] Graph neural networks have been developed to predict chemical reactivity.[2,3] Even tasks as daunting as predicting protein folding have been tackled with success by groups like AlphaFold.[4] Deep molecular neural networks, such as MEGNet[5] and SchNet,[6] have achieved computational accuracy rivaling correlated electronic structure methods,[7] while boasting the computational cost of classical force fields. One of the most successful NN of the latter type is ANAKIN-ME[1,8] (ANI), an open-source general-purpose neural network-based atomistic potential for organic molecules. To date, several ANI models have been published, focusing on organic molecules in the gas phase. The original ANI-1 model was developed by random sampling conformational space of 57,000 organic molecules composed of C, N, O, and H atoms and then running wB97x-D/6-31G(d,p) density functional theory (DFT) calculations to obtain potential energies for training. The subsequent ANI-1x model was improved through an active learning scheme.[9] The latest ANI-2x and ANI-2xt were trained at the same level of theory, but now include elements S, F, and Cl.[10,11] Unfortunately, these neural networks have their own set of problems.

ANI uses a modified version of the symmetry functions proposed by Behler and Parrinello in 2007 for representing molecules in a vectorized form.[12] As the method circumvents the need for excess feature engineering, it has been adopted by many different projects.[13−15] However, it is also a source of one of the greatest limitations of these networks. The symmetry functions act like radial distribution functions, i.e., they represent the environment around each atom through pairwise distances and three-bodied angles.[12] Effectively, the symmetry functions create a fingerprint for each atom. To ensure computational efficiency and avoid overfitting, both radial and angular symmetry functions utilize a distance cutoff; the atoms past the cutoff distance are ignored.[12] For example, the radial and angular cutoffs in the original implementation of ANI are 5.2 and 3.5 Å, respectively.[8] Because of this, in the eyes of the network, each atom only "sees" other atoms within this 5.2 Å range, meaning that any long-range effects beyond two overlapping environments are neglected. Thus, while providing significant computational savings, the distance cutoffs result in the accumulation of errors in modeling larger systems where long-range effects might originate from electrostatic, polarization, and van der Waals terms. The

Received: August 11, 2024
Revised: September 18, 2024
Accepted: September 19, 2024

inclusion of long-range terms in ANI potentials was previously explored using additive dispersion corrections and Coulomb interactions via partial charges trained in a complementary charge network.[16] The goal of this work is to develop an alternative computationally efficient way to account for the long-range effects by coupling ANI potentials with an effective fragment potential (EFP) polarizable force field.

EFP is a model potential designed for systematic first-principles-based description of intermolecular interactions.[17−21] EFP represents a molecular system as a collection of effective fragments that interact via electrostatic, polarization, dispersion, exchange-repulsion, and charge-transfer terms derived from the Rayleigh−Schrödinger perturbation theory. The parameters of the effective fragments, which are typically chosen to be either individual molecules or functional units of polymers and other macromolecules, are precomputed in separate quantum-mechanical calculations and reused in all subsequent evaluations of the system properties, providing significant computational savings. Additionally, EFP can be used in conjunction with quantum mechanics (QM) methods in hybrid quantum mechanics/molecular mechanics (QM/MM) or, more precisely, QM/EFP calculations, where the chemically active region is modeled quantum-mechanically, while the remainder of a system is modeled with EFP.[22−28] In QM/EFP models, EFP provides polarizable embedding for the QM region, which has been shown to be essential for describing electronic excited and ionized states in solvated chromophores and photoactive proteins.[27−31] However, the cost of QM/EFP calculations is only slightly more expensive than the cost of the analogous gas-phase calculation of the QM region.

In this work, we explore the ANI/EFP hybrid model in which ANI describes properties of the chemically active region, i.e., formally substitutes the QM calculation in QM/EFP, while the EFP embedding accounts for long-range interactions. To accomplish this task, we augment the ANI network with additional input parameters that represent interactions of ANI atoms with the EFP environment. The modified network is retrained on the interaction energies between the QM and EFP regions, as obtained from the QM/EFP calculations. The performance of the ANI/EFP network is evaluated on the interaction energies in several systems with simple molecular solvents (water, methanol, ammonia, and methane) and a subset of water-solvated molecules from the FreeSolv database.[32] Prospects and limitations of the developed NN, along with an outlook for future work, are also discussed.

2. THEORETICAL DETAILS

2.1. ML/MM Model. Our machine learning/molecular mechanics (ML/MM) models closely mimic the electrostatic or polarizable embedding QM/MM. In QM/MM, the QM Hamiltonian is augmented by coupling terms responsible for interactions with the classical environment, such that the quantum wave function "feels" the electrostatic field produced by partial charges (or, generally, electrostatic multipoles) of surrounding classical atoms and, in the case of the polarizable embedding, also by induced dipoles. Such coupling between the quantum and classical regions is essential for the efficient modeling of extended systems. Since electrostatic interactions decay slowly with distance, one needs to include at least several nanometers of the classical environment around the chemically active region.[27,33] On the other hand, typical ML potentials are based on a local atomic environment, such that each atom is aware of its neighbors within a ∼1 nm radius.[8] While training an NN based on the local atomic environment protects it from overparameterization, the resulting ML potentials effectively account only for the short- and midrange interactions, such that the long-range effects need to be incorporated separately. In this work, we address this shortcoming of local atomic environment networks by training the NN in the presence of an additional parameter, an electrostatic potential (EP) specified on each atom described by the NN. Effectively, the electrostatic potential provides a compact representation of a classical environment on a given atom in the ML region. Moreover, the EP is a scalar property that will not destroy the NN's translational and rotational invariance. The NN augmented by EPs as additional input parameters will ensure that long-range electrostatic interactions are accounted for. In principle, such ML potentials can be trained in the presence of electrostatic charges only, mimicking electrostatic embedding QM/MM or in a self-consistently polarized system, which would correspond to polarizable embedding models.

The QM/MM system is described with the Hamiltonian $\hat{H}$

$$\hat{H} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM\text{-}MM} \tag{1}$$

The total ground-state energy in the electrostatic embedding is defined by

$$E_{QM/MM} = \langle \Psi | \hat{H}_{QM} + \hat{V}^{coul} | \Psi \rangle + E^{vdW}_{QM\text{-}MM} + E^{coul}_{MM} + E^{vdW}_{MM} + E^{bond}_{MM} \tag{2}$$

where Ψ is the wave function of the quantum system, E^{coul}_{MM}, E^{vdW}_{MM}, and E^{bond}_{MM} are electrostatic, van der Waals, and covalent energies of the classical region defining the Hamiltonian $\hat{H}_{MM}$. The energy of van der Waals interactions between the QM and MM subsystems, E^{vdW}_{QM-MM}, and the electrostatic perturbation of the quantum Hamiltonian by the classical region, $\hat{V}^{coul}$, define the coupling term $\hat{H}_{QM-MM}$. $\hat{H}_{QM}$ is the electronic Hamiltonian of a pristine QM system. In the simplest form, $\hat{V}^{coul} = \sum_i^{atoms} \hat{q}_i / |r - R_i|$, where $\hat{q}_i$ and R_i are partial charges and positions of classical atoms i, and r is the electronic coordinate.

In the ML/MM model, the ML potentials can be trained to replicate the energies and gradients of the $\langle \Psi | \hat{H}_{QM} + \hat{V}^{coul} | \Psi \rangle$ integral. Thus, similarly to how ANI potentials return an effective atomic energy following the training to molecular energies $\langle \Psi | \hat{H}_{QM} | \Psi \rangle$, the potentials in ANI/MM are trained to reproduce atomic energies of a molecule embedded in the EP given by $\hat{V}^{coul}$. The total system energy and gradient in ANI/MM can be obtained by evaluating the remaining terms in eq 2 classically, in complete analogy to QM/MM models.

The effective fragment potential (EFP) is designed to describe intermolecular interactions.[19−21] EFP represents a molecular system as a collection of effective fragments with predefined parameters that interact with each other through electrostatic, polarization, dispersion, and exchange-repulsion terms

$$E_{EFP} = E^{elec}_{EFP} + E^{pol}_{EFP} + E^{disp}_{EFP} + E^{ex\text{-}rep}_{EFP} \tag{3}$$

The parameters of the effective fragments, including distributed multipoles, distributed static and dynamic polarizability tensors, localized wave function, and Fock matrix, can be either computed in the GAMESS electronic structure software[34−36] or obtained from a database of the precomputed parameters.[37]

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

The total ground-state energy in the polarizable QM/EFP model $E_{\mathrm{QM/EFP}}$ is

$$E_{\mathrm{QM/EFP}} = \langle\Psi|\hat{H}_{\mathrm{QM}} + \hat{V}^{\mathrm{coul}} + \hat{V}^{\mathrm{pol}}|\Psi\rangle + E^{\mathrm{coul}}_{\mathrm{QMnuc\text{-}EFP}} + E^{\mathrm{pol}} + E^{\mathrm{disp}}_{\mathrm{QM\text{-}EFP}} + E^{\mathrm{ex\text{-}rep}}_{\mathrm{QM\text{-}EFP}} + E^{\mathrm{coul}}_{\mathrm{EFP}} + E^{\mathrm{disp}}_{\mathrm{EFP}} + E^{\mathrm{ex\text{-}rep}}_{\mathrm{EFP}} \quad (4)$$

where the last three terms represent electrostatic, dispersion, and exchange-repulsion interactions in the EFP subsystem; $E^{\mathrm{disp}}_{\mathrm{QM-EFP}}$ and $E^{\mathrm{ex-rep}}_{\mathrm{QM-EFP}}$ are dispersion and exchange-repulsion interactions between the QM and EFP subsystems.

The electrostatic operator $\hat{V}^{\mathrm{coul}}$ includes contributions from point charges $\hat{q}_m$, dipoles $\hat{\mu}_m$, quadrupoles $\hat{\Theta}_m$, and octupoles $\hat{\Omega}_m$ located at the atoms and bond midpoints, referred to as multipole expansion points m positioned at R_m, of the EFP fragments

$$\hat{V}^{\mathrm{coul}} = \sum_{m}^{\mathrm{mult.points}} \left[\hat{q}_m T(|r - R_m|) - \sum_{a}^{x,y,z} \hat{\mu}_{a,m} T_a(|r - R_m|) + \frac{1}{3} \sum_{a,b}^{x,y,z} \hat{\Theta}_{ab,m} T_{ab}(|r - R_m|) - \frac{1}{15} \sum_{a,b,c}^{x,y,z} \hat{\Omega}_{abc,m} T_{abc}(|r - R_m|) \right] \quad (5)$$

T, T_a, T_{ab}, and T_{abc} are the electrostatic tensors of ranks zero to three.[38] $E^{\mathrm{coul}}_{\mathrm{QMnuc-EFP}}$ is an electrostatic interaction energy between solute nuclei Z_{A} and EFP multipoles. This interaction can be expressed as an interaction of classical point charges with the electrostatic potential of a multipole expansion $P^{\mathrm{coul}}_{\mathrm{A}}$

$$E^{\mathrm{coul}}_{\mathrm{QMnuc\text{-}EFP}} = \sum_{A}^{\mathrm{QMnuclei}} Z_{\mathrm{A}} P^{\mathrm{coul}}_{\mathrm{A}} \quad (6)$$

$$P^{\mathrm{coul}}_{\mathrm{A}} = \sum_{m}^{\mathrm{mult.points}} \left[\hat{q}_m T(|R_{\mathrm{A}} - R_m|) - \sum_{a}^{x,y,z} \hat{\mu}_{a,m} T_a(|R_{\mathrm{A}} - R_m|) + \frac{1}{3} \sum_{a,b}^{x,y,z} \hat{\Theta}_{ab,m} T_{ab}(|R_{\mathrm{A}} - R_m|) - \frac{1}{15} \sum_{a,b,c}^{x,y,z} \hat{\Omega}_{abc,m} T_{abc}(|R_{\mathrm{A}} - R_m|) \right] \quad (7)$$

where electrostatic tensors T, T_a, etc., are computed between the positions of multipole expansion points R_m and solute nuclei R_{A}.

The polarization perturbation due to the EFP fragments $\hat{V}^{\mathrm{pol}}$ is

$$\hat{V}^{\mathrm{pol}} = \frac{1}{2} \sum_{p}^{\mathrm{pol.points}} \sum_{a}^{x,y,z} \frac{(\hat{\mu}_{a,p} + \hat{\tilde{\mu}}_{a,p})a}{|r - R_p|^3} = -\frac{1}{2} \sum_{p}^{\mathrm{pol.points}} \sum_{a}^{x,y,z} (\hat{\mu}_{a,p} + \hat{\tilde{\mu}}_{a,p}) T_a\left(|r - R_p|\right) \quad (8)$$

where $\hat{\mu}_p$ and $\hat{\tilde{\mu}}_p$ are the induced and conjugated induced dipoles on points p with positions R_p, converged self-consistently with the electronic ground-state wave function. Polarization energy

$$E^{\mathrm{pol}} = \frac{1}{2} \sum_{p}^{\mathrm{pol.points}} \sum_{a}^{x,y,z} \left[-\hat{\mu}_{a,p} F^{\mathrm{mult},p}_{a} + \left(\hat{\tilde{\mu}}_{a,p} F^{\mathrm{QM}}_{a,p} - \hat{\mu}_{a,p} F^{\mathrm{QMnuc}}_{a,p}\right)\right] \quad (9)$$

includes the polarization energy of the EFP subsystem (the first term, $F^{\mathrm{mult},p}_{\mathrm{a}}$, is the field due to static multipoles on effective fragments) and an additional contribution to the QM-EFP polarization energy (two terms in parentheses, $F^{\mathrm{QM}}_{a,p}$ and $F^{\mathrm{QMnuc}}_{a,p}$, are the fields due to the QM electronic wave function and the QM nuclei, respectively).

The interaction energy between the QM solute and EFP solvent, where the "solute" and "solvent" are broadly defined, is given as

$$E_{\mathrm{inter}} = E_{\mathrm{QM/EFP}} - E_{\mathrm{QM}} - E_{\mathrm{EFP}} = \langle\Psi|\hat{H}_{\mathrm{QM}} + \hat{V}^{\mathrm{coul}} + \hat{V}^{\mathrm{pol}}|\Psi\rangle + E^{\mathrm{coul}}_{\mathrm{QMnuc\text{-}EFP}} - E_{\mathrm{QM}} + (E^{\mathrm{pol}} - E^{\mathrm{pol}}_{\mathrm{EFP}}) + E^{\mathrm{disp}}_{\mathrm{QM\text{-}EFP}} + E^{\mathrm{ex\text{-}rep}}_{\mathrm{QM\text{-}EFP}} \quad (10)$$

The functional form of $E^{\mathrm{disp}}_{\mathrm{QM-EFP}}$ and $E^{\mathrm{ex-rep}}_{\mathrm{QM-EFP}}$ terms have been discussed and developed in refs 18,39−42. However, in polarizable embedding models, it is common to represent these energy contributions classically, i.e., as dispersion and exchange-repulsion terms in a purely EFP system (eq 3).[43−45] We utilized this approach in this work. Additionally, assuming that polarization (and induced dipoles) of the solvent is not strongly affected by the solute, the solute−solvent interaction energy becomes

$$E_{\mathrm{inter}} \sim \langle\Psi|\hat{H}_{\mathrm{QM}} + \hat{V}^{\mathrm{coul}} + \hat{V}^{\mathrm{pol}}|\Psi\rangle + E^{\mathrm{coul}}_{\mathrm{QMnuc\text{-}EFP}} - E_{\mathrm{QM}} + \frac{1}{2} \sum_{p}^{\mathrm{pol.points}} \sum_{a}^{x,y,z} (\hat{\tilde{\mu}}_{a,p} F^{\mathrm{QM}}_{a,p} - \hat{\mu}_{a,p} F^{\mathrm{QMnuc}}_{a,p}) + [E^{\mathrm{disp}}_{\mathrm{EFP\text{-}EFP}} + E^{\mathrm{ex\text{-}rep}}_{\mathrm{EFP\text{-}EFP}}] \quad (11)$$

where the polarization solute−solvent interactions are split between the integral term and the two terms in parentheses, while the dispersion and exchange-repulsion interactions between the solute and solvent are modeled at the EFP level (two terms in square brackets).

The ANI potentials in the ANI/EFP model are trained to replicate the energies of the Coulomb and polarization components of the interaction energy in eq 11, i.e.

$$E^{\mathrm{train}}_{\mathrm{inter}} = \langle\Psi|\hat{H}_{\mathrm{QM}} + \hat{V}^{\mathrm{coul}} + \hat{V}^{\mathrm{pol}}|\Psi\rangle + E^{\mathrm{coul}}_{\mathrm{QMnuc\text{-}EFP}} - E_{\mathrm{QM}} + \frac{1}{2} \sum_{p}^{\mathrm{pol.points}} \sum_{a}^{x,y,z} \left(\hat{\tilde{\mu}}_{a,p} F^{\mathrm{QM}}_{a,p} - \hat{\mu}_{a,p} F^{\mathrm{QMnuc}}_{a,p}\right) \quad (12)$$

After the NN returns this interaction energy, the total solute−solvent interaction energy can be evaluated using eq 11, and the total system energy can be obtained following eq 4, in which all nonintegral terms are computed at the EFP level. To provide the NN with information on the polarizable solvent environment, we consider the following classical electrostatic

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

potential EP on the atoms of the solute subsystem A with positions R_A

$$\begin{aligned}\mathrm{EP_A} &= P_\mathrm{A}^\mathrm{coul} + P_\mathrm{A}^\mathrm{pol} \\ &= \sum_m^{\text{mult.points}} \Bigg[\hat{q}_m T(|R_\mathrm{A} - R_m|) - \sum_a^{x,y,z} \hat{\mu}_{a,m} T_a(|R_\mathrm{A} - R_m|) \\ &\quad + \frac{1}{3} \sum_{a,b}^{x,y,z} \hat{\Theta}_{ab,m} T_{ab}(|R_\mathrm{A} - R_m|) \\ &\quad - \frac{1}{15} \sum_{a,b,c}^{x,y,z} \hat{\Omega}_{abc,m} T_{abc}(|R_\mathrm{A} - R_m|) \Bigg] \\ &\quad - \frac{1}{2} \sum_p^{\text{pol.points}} \sum_a^{x,y,z} \left(\hat{\mu}_{a,p} + \hat{\tilde{\mu}}_{a,p} \right) T_a\left(\left| R_\mathrm{A} - R_p \right| \right)\end{aligned} \tag{13}$$

where summation goes over all EFP solvent multipole and polarizability points. This potential is a representation of the Coulomb and polarization contributions of the EFP environment to the quantum region. The values of the electrostatic potential at all solute nuclei are evaluated in the EFP calculation of the (solute + solvent) system and provided as additional parameters to the NN. The electrostatic potential is the only information that the NN defined for the solute system knows about the solvent.

2.2. Symmetry Function Modification. When developing a neural network that takes molecular coordinates as input, the final output needs to be invariant to the transformations of translation, rotation, and permutation of the same types of atoms.[46] This makes working with raw coordinates as input to a network more difficult, and usually, some transformation is done when vectorizing a molecule to prepare it. Behler and Parinello proposed the symmetry functions that satisfy the preservation of the required symmetries.[12] These functions were called an atomic environmental vector (AEV), $\vec{G}_\mathrm{i}^\mathrm{Z} = \{G_1, G_2, G_3, \cdots, \text{and } G_\mathrm{M}\}$, and for each atom, an associated AEV is computed and passed through its own atomic network to return the atomic energy. The total energy is obtained as a sum of all-atom energies. The ANI model, built upon the work by Behler and Parrinello, borrows the radial symmetry function (G_m^R) and uses a modified version of the angular symmetry function ($G_m^{A_\mathrm{mod}}$)[8]

$$G_m^R = \sum_{j \neq i}^{\text{all atoms}} \mathrm{e}^{-\eta(R_{ij} - R_\mathrm{S})^2} f_\mathrm{C}(R_{ij}) \tag{14}$$

$$\begin{aligned}G_m^{A_\mathrm{mod}} &= 2^{1-\zeta} \sum_{j,k \neq i}^{\text{all atoms}} (1 + \cos(\theta_{ijk} - \theta_\mathrm{s}))^\zeta \exp\Bigg[-\eta \\ &\quad \left(\frac{R_{ij} + R_{ik}}{2} - R_\mathrm{S} \right)^2 \Bigg] f_\mathrm{C}(R_{ij}) f_\mathrm{C}(R_{ik})\end{aligned} \tag{15}$$

Here, R_{ij} and R_{ik} are the distances between atoms i and j and i and k, respectively, and θ_{ijk} is the angle between atoms i, j, k centered on atom i. R_s and η are the center and width of the Gaussian peak for the radial distribution. Different R_s and η values, indexed by m, set the part of the radial environment to be probed by the symmetry function. Similarly, θ_s and ζ determine the center and width of the angular probe. The cutoff function f_C is defined as

$$f_\mathrm{C}(R_{ij}) = \begin{cases} 0.5 \times \cos\left(\dfrac{\pi R_{ij}}{R_\mathrm{C}} \right) + 0.5 \text{ for } R_{ij} \leq R_\mathrm{C} \\ 0.0 \text{ for } R_{ij} > R_\mathrm{C} \end{cases} \tag{16}$$

where R_C is the cutoff radius.

The symmetry functions encode N interatomic distances with ith atom neighbors within a cutoff radius R_c into invariant fixed-length AEV. To adapt these symmetry functions to chemical environments that consist of more than one or two atom types, an explicit-type differentiation is performed whereby AEV consists of multiple radial vectors for each atom type and multiple angular vectors for each atom type pair. A separate neural network model is used for each element. This explicit atom type differentiation leads to a quadratic growth in the size of the AEV with the number of atomic species included in the training set and a different network for each atomic species.

Each $\vec{G}_\mathrm{i}^\mathrm{Z}$ for the ith atom with atomic number Z is used as input into a single neural network potential (NNP). With an invariant AEV $\vec{G}_\mathrm{i}^\mathrm{Z}$, the total energy of a molecule m is expressed as

$$E_\mathrm{total}(\mathrm{m}) = \sum_i^{\text{all atoms}} \mathrm{NNP}_{Z_i}(\vec{G}_\mathrm{i}^\mathrm{Z}) \tag{17}$$

In the ANI/EFP network, the ANI AEV is computed for each atom. The associated electrostatic potential for the atom is concatenated to the end of the AEV and passed through the network. The EFP electrostatic potentials are computed in the LibEFP software library.[47,48]

2.3. Neural Network Architecture. Development and training of the ANI/EFP network was done via TorchANI.[1] The design of the ANI/EFP network is as follows: The original ANI-1x network is used for the base, and an additional node, termed the "EFP node" and schematically shown with orange circles in Figure 1, is added to each layer except for the final. The EFP nodes have no connections to the original ANI network, but the ANI nodes do have new connections to the

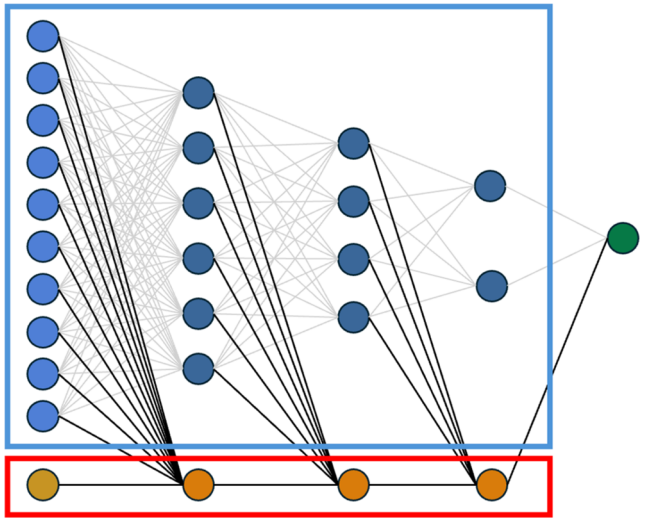


Figure 1. Schematic representation of the ANI/EFP network architecture. Blue: original ANI network nodes. Red: added EFP nodes. The weights of the connections linking to the EFP nodes, shown schematically with dark gray lines, are retrained.

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

EFP nodes whose weights are updated through training. The biases for each layer are set to zero. The weights of the original ANI network are frozen during training so as not to retrain the ANI portion of the network.

The EFP node of the initial input layer contains the electrostatic potential EP (eq 13) felt by the atom due to its EFP environment. This value is passed through the network alongside the ANI network and then combined in the final layer. This final number is the atomic energy that accounts for both the atom's local ANI environment and the long-range EFP environment. In this work, we trained the ANI/EFP network to the Coulomb and polarization components of the QM/EFP interaction energy $E_{\text{inter}}^{\text{train}}$ (eq 12). The loss function L is

$$L = \sum_{j}^{\text{all molecules}} (E_j^{\text{ANI/EFP}} - E_{\text{inter},j}^{\text{train}}) \tag{18}$$

where $E_j^{\text{ANI/EFP}}$ is the energy returned by the network. In the present implementation of the ANI/EFP model, the forces were not included in the loss function.

The architecture of the ANI/EFP network was modeled after that of the ANI-1x network. The network consists of four individual networks for each supported atom type: H, C, N, and O. Each atom network varies slightly in the width of each layer. The input layer for each network is 385 nodes wide, composed of 384-length ANI-1x AEV and an additional node containing the EFP electrostatic potential. The number of nodes per layer is a hyperparameter tuned during the development and training of ANI; it is kept the same in ANI/EFP. A CELU activation function is used for all nodes. There are three hidden layers in the network for a total of five layers, including the input and output layers.

3. COMPUTATIONAL DETAILS

3.1. Data Set. The data set is composed of 5964 randomly sampled structures from the molecular dynamics trajectories of the following systems: ammonia solvated in ammonia, methanol solvated in methanol, methane solvated in methane, and water solvated in water. Additionally, 1364 structures of small organic molecules from the FreeSolv[32] data set solvated in water are included. Molecular dynamics simulations were conducted in the GROMACS software.[49] Each system was first prepared by inserting the molecule into a 3 nm wide cubic box and solvating the box. Depending on the system, the simulation cell contained roughly 300 to over 800 solvent molecules. First, an energy minimization step was performed, followed by a 300 ps NVT step, then a 1.8 ns NPT step, and finally, a 10 ns production run with a 2 fs time step, all performed at 298 K and ambient pressure, using the OPLS all-atom force field for all molecules including water.[50] LINCS algorithm was used to constrain the bonds involving H atoms.[51] 2500 snapshots were randomly extracted from the trajectories of the production runs.

Each molecular structure is split into two parts: a central molecule as the solute region and the remaining molecules as the solvent region. EFP parameters for each solute and solvent molecule were prepared using the MAKEFP run in the GAMESS electronic structure package[34−36] in a 6-31G* basis set. Then, the following calculations were conducted for each molecular system: (1) a full-system EFP calculation provided electrostatic potentials due to the EFP environment on each atom in the solute region (also represented as the EFP fragment); (2) a gas-phase QM energy calculation of the solute region ignoring the solvent; (3) an EFP energy calculation of the solvent region ignoring the solute region; and (4) a full-system QM/EFP energy calculation in which the solute is represented with QM and the solvent is modeled with EFP. Using this data, the QM/EFP interaction energy was evaluated following eq 12 and used as the training label. Note that the interaction energies do not account for the solute and solvent deformation energies.

EFP-only calculations were conducted in the LibEFP software library.[47,48] The QM and QM/EFP calculations, conducted in the Q-Chem electronic structure package,[52,53] were performed with the wB97x[54]/6-31G* level of theory to match the level of theory used for training ANI-1x. The data set was randomly shuffled and split into a ratio of 80:20 for training and testing, respectively.

3.2. Additional Structures for Testing ANI/EFP Transferability. To test the transferability of the retrained ANI/EFP network, additional molecular structures were generated. None of these structures were used for training the network. These additional structures can generally be split into three categories: (1) known solute, known solvent; (2) known solute, unknown solvent; and (3) unknown solute, known solvent. For the first category, one ammonia molecule (known solute) is solvated in ten or hundred ammonia molecules (known solvent). For the second category, one ammonia molecule (known solute) is solvated in about ten solvent molecules of the following: acetic acid, acetonitrile, benzene, and toluene (unknown solvent). For the last category, random small organic molecules from the FreeSolv data set[32] that were not a part of the original training/testing data set were taken (unknown solute) and solvated in 10 water molecules (known solvent). All of these structures were obtained from the *ab initio* MD simulations performed in the Q-Chem electronic structure package[52,53] using the wB97x level of theory with a 6-31G* basis set. Each system was sampled for 5 ps, with 0.48 fs time step and with initial velocities corresponding to 298 K, in the NVE ensemble. 1000 structures were sampled from each trajectory. This was also to test if energies of the structures sampled from *ab initio* MD could be predicted by the network, since the network was only trained on structures sampled from classical MD simulations originally.

3.3. Code Availability. Scripts for generating the data set and for training the ANI/EFP network, as well as the data set and the best-performing network, are available in the GitHub repository.[55] The repository also contains a tutorial on how to mimic the workflow of this project. Most of the code to retrain ANI-1x was taken from the TorchANI tutorials stored on the TorchANI Web site.[1,56]

4. RESULTS AND DISCUSSION

4.1. Performance of the ANI/EFP Network on the Original Data Set. First, we analyze the performance of the ANI/EFP network in predicting electrostatic and polarization components of the solute−solvent interaction energies in the original training data set. Figure 2 shows the regression plot and the histogram of the network errors versus the QM/EFP interaction energies of eq 12 on the training data set. Note that differently from the original ANI-1x, the ANI/EFP network produces the solute−solvent interaction energy as the output. For reference, Figure S1 in the Supporting Information provides a comparison of the total interaction energies of all

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

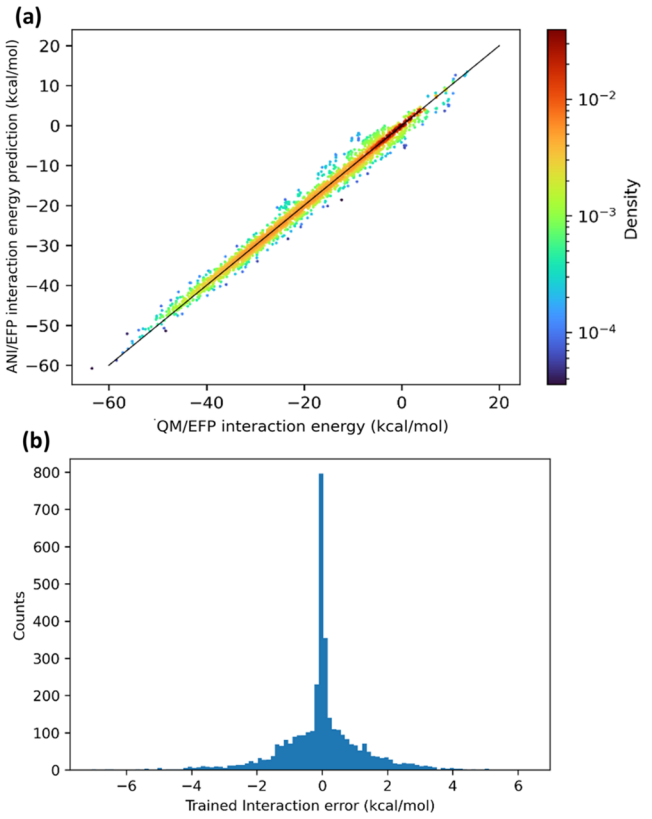


Figure 2. (a) Regression plot of the ANI/EFP predicted interaction energies versus the QM/EFP interaction energies for the training data set. $y = x$ asymptote is shown in black. The density of points is color-coded to a log scale, with red being the densest region of points and blue being the least dense. (b) Histogram plot of the ANI/EFP interaction energy errors for the training data set.

systems in the training data set computed with the original ANI-1x versus the QM/EFP model.

The new ANI/EFP network has an RMSE of 1.318 kcal/mol. However, this result is quite expected because the predictions are run over the original data set used for training/testing. In order to have a less biased analysis, the following sections analyze the performance of ANI/EFP on the structures that the network had never seen before.

4.2. Transferability of the ANI/EFP Network. It is essential to test the transferability of the developed network. We explored the performance of the ANI/EFP network applied to a different ensemble of structures, as well as to the systems in different solvent environments and containing new solvated molecules. For that, three sets of new structures were prepared, as explained in Section 3.2. These new structures, not used for network training, can generally be split into three categories: (1) known solute, known solvent; (2) known solute, unknown solvent; and (3) unknown solute, known solvent. The first set of structures tested whether the network can accurately predict interaction energies for an ensemble of structures different from the molecules from the training set. The second set explores whether training of a solute in some solvent environments is sufficient to predict its interaction energies with different types of solvents. The third set examines whether a solute molecule needs to be included in the training set for an accurate prediction of its interaction with the solvent.

The data for the first category, where one ammonia molecule (known solute) is solvated by ten ammonia molecules (known solvent), are shown in Figure 3.

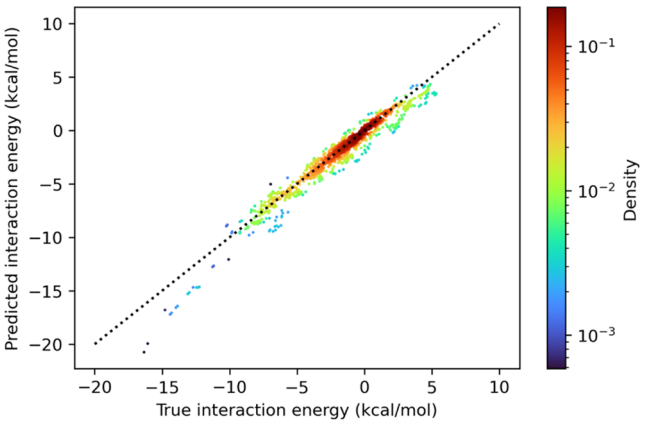


Figure 3. Regression plot of the ANI/EFP network predicted interaction energies versus the QM/EFP interaction energies for one ammonia molecule solvated in ten other ammonia molecules. $y = x$ Asymptote is shown in black. The density of points is color-coded to a log scale, with red being the densest region of points and blue being the least dense.

RMSE for the ammonia in the ammonia system is 0.680 kcal/mol. The small RMSE value shows that the network can handle predictions for structures outside the data set. Apart from the different number of solvent molecules between the trained and tested structures, the additional disparity was in the level of theory used to prepare the structures. The former ones were extracted from classical MD trajectories, while the latter ones were prepared using *ab initio* MD simulations. This suggests that the network can accurately predict the interaction energies of the structures extracted from the QM dynamics based on the training set prepared with structures from classical simulations.

To test the second case, i.e., the interaction of a known solute with an unknown solvent, we computed interaction energies of systems consisting of an ammonia molecule (solute) sequestered within a small bubble of less than ten solvent molecules of either acetic acid, acetonitrile, benzene, or toluene. The hypothesis is that using a different solvent should not be a significant challenge because, in the eyes of the network, the solvent is represented by only a single number: the electrostatic potential. Therefore, a specific type of solvent should not matter if the training data set includes data points with similar electrostatic potentials. To this end, Figure 4 compares the values of electrostatic potentials at N and H atoms in ammonia in all structures of the training data set (shown in red) and the testing data set (shown in blue). As demonstrated in Figure 4, the training data set provides a broader distribution of the electrostatic potential values for both N and H compared to the data set with new solvents, meaning that the network is expected to accurately predict the interaction energies in this test set.

The regression plot representing ANI/EFP versus QM/EFP electrostatic and polarization interaction energies of ammonia solvated in different solvents is shown in Figure 5. The data in Figure 5, with an RMSE of 0.56 kcal/mol, confirm the hypothesis that the network can handle the effects of novel solvent environments if the electrostatic potential calculated on

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

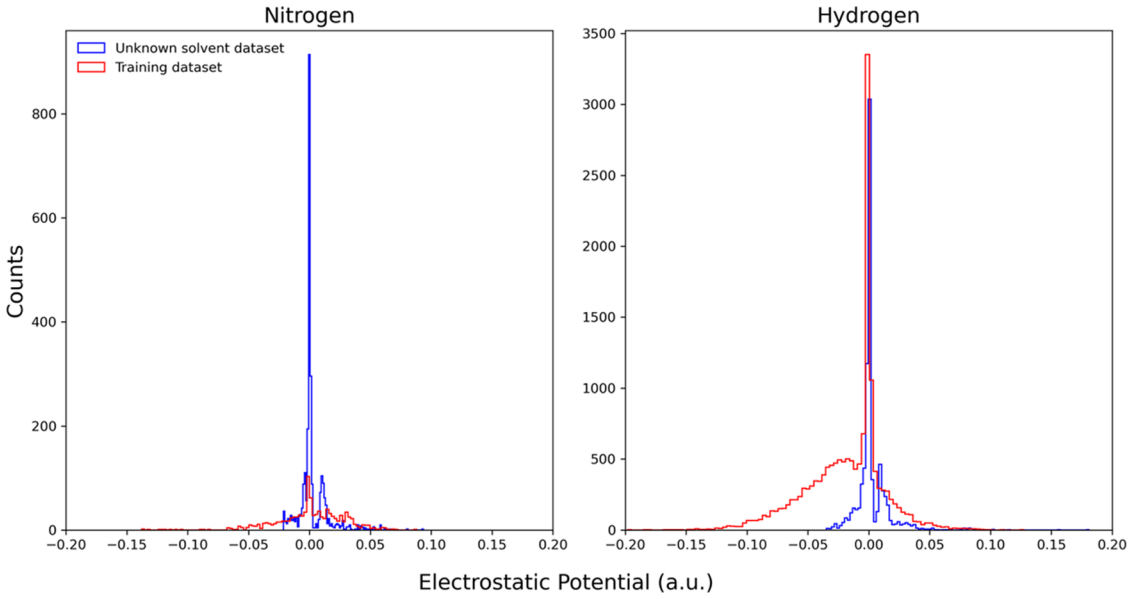


Figure 4. Histogram plot of the electrostatic potentials on the nitrogen and hydrogen atoms of ammonia in the data set used during training (red) and in the data set of ammonia solvated by small clusters of acetic acid, acetonitrile, benzene, or toluene (blue).

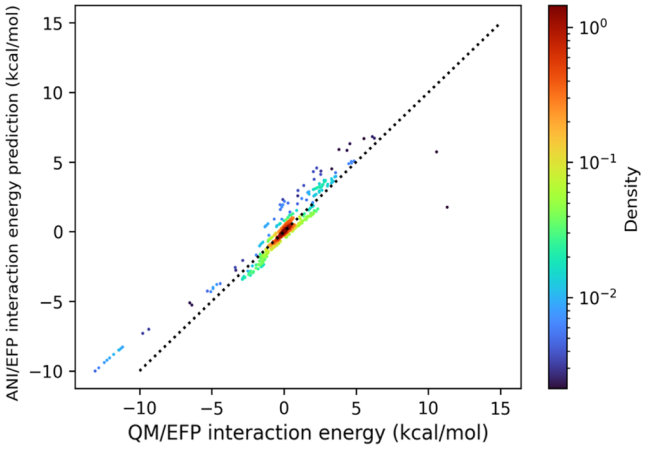


Figure 5. Regression plot of the ANI/EFP network predicted interaction energies vs the QM/EFP interaction energies for one ammonia molecule solvated in small clusters of acetic acid, acetonitrile, benzene, or toluene. $y = x$ Asymptote is shown in black. The density of points is color-coded to a log scale, with red being the densest region of points and blue being the least dense.

the atoms is within the range of the training data. We consider the results of this test as a major success of the ANI/EFP network, which makes it suitable for modeling a solute in an arbitrary environment.

As a side note, larger deviations between ANI/EFP and QM/EFP results for complexes with large (by magnitude) electrostatic and polarization interaction energies observed in Figures 3 and 5 are somewhat expected. It should be noted that such complexes are expected to correspond to structures with short intermolecular distances that might be on the repulsive part of the potential energy surface. However, for these structures, the errors due to the representation of the electronic density by classical point multipoles are often significant. Similarly, utilizing electrostatic potentials only at the solute nuclei in ANI/EFP becomes less accurate for strongly interacting molecules compared to the QM/EFP treatment, in which the electrostatic potential due to solvent is available and interacts with the electron density at the whole solute molecule. Additionally, strongly interacting complexes often exhibit significant polarization contributions. In the present ANI/EFP model, the solvent polarization contribution to EP is computed at the pure EFP level, i.e., when both solute and solvent are modeled as EFP fragments. However, in QM/EFP, the solute is modeled quantum-mechanically such that it induces a somewhat different field to the solvent compared to the pure EFP scheme (see discussion near eq 10 and 11). The discrepancies in the polarization of EFP/EFP vs QM/EFP representations, and, by formulation, the corresponding discrepancies of ANI/EFP and QM/EFP, become more noticeable for strongly interacting complexes.

Finally, we apply the ANI/EFP network to solute molecules that it has never seen. For that, we generated structures of random small organic molecules from the FreeSolv data set,[32] not included in the original training/testing data set, that were solvated by 10 water molecules. This data are shown in Figure 6.

Figure 6 demonstrates that the network's transferability to solute molecules it has not been trained on is poor. The RMSE for this trial is 55.47 kcal/mol. This test provides an important lesson that for better transferability, the training data must be expanded to include a larger and more diverse set of solute molecules. Another plausible solution to the problem of the network's transferability to different solute molecules is expanding the training set by including a variety of chemical functional groups and binding motifs. The validity of this approach will be explored in future work.

5. CONCLUSIONS

In this work, we developed the ANI/EFP machine-learned neural network potential that allows a description of long-range effects and solute−solvent interaction energies by combining the short-range ANI-1x network with the EFP potentials for describing noncovalent interactions. Electrostatic potentials on solute atoms are used as an additional parameter in ANI/EFP. The network is retrained on a relatively small data set of solvated systems while keeping the weights of the original ANI-

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

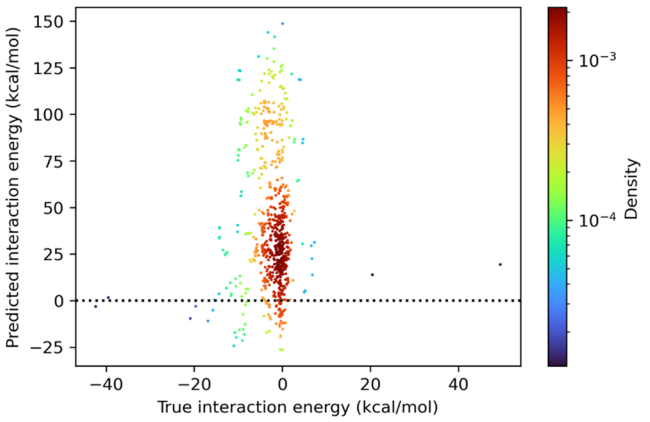


Figure 6. Regression plot of the ANI/EFP network predicted interaction energies versus the QM/EFP interaction energies for the case of small organic molecules from the FreeSolv data set not previously trained on. $y = x$ Asymptote is shown in black. The density of points is color-coded to a log scale, with red being the densest region of points and blue being the least dense.

1x network fixed. The resulting network successfully predicts solute−solvent interaction energies for the solute molecules for which it was trained. Thus, this work stands as a proof of concept that it is possible to retrain a general atomistic neural network potential to describe long-range interactions of specific solute molecules with an arbitrary solvent. However, such a minimally retrained network fails to predict interaction energies for the solute molecules not included in the retraining data set. To achieve generality, the network should be expanded and include a larger diversity of solute molecules in the training data.

Alternatively, the electrostatic embedding energies in QM/MM could be accounted for by a classical polarizable model in which the solute (QM region) is represented by atomic charges and polarizabilities.[57] In this scheme, the atomic charges are learned using an electron-equilibration scheme, and the Thole polarizability model relates the electronic structure to the response of the solvent electric field.[57] On the other hand, long-range effects could be learned with data-driven models like AIMNet2 utilizing a neural charge equilibration (NQE) module and message passing iterations.[58,59]

Two plausible future directions for this work are to improve the generality of the network by generating a more diverse data set of solvated systems, especially expanding the solute functional space. Another less computationally demanding task is to train the ANI/EFP network for a specific class of solute−solvent interactions, such as protein−ligand interactions for a set of related ligands, effectively using ANI as an accurate and computationally affordable potential incorporated in multiscale modeling of biological systems.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c01052.

Comparison of ANI-1x and QM/EFP interaction energies (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Lyudmila Slipchenko − *Department of Chemistry, Purdue University, West Lafayette, Indiana 47907-2084, United States;* orcid.org/0000-0002-0445-2990; Email: lslipchenko@purdue.edu

Authors

Shahed Haghiri − *Department of Chemistry, Purdue University, West Lafayette, Indiana 47907-2084, United States*

Claudia Viquez Rojas − *Department of Chemistry, Purdue University, West Lafayette, Indiana 47907-2084, United States*

Sriram Bhat − *Department of Computer Science, The University of Texas at Dallas, Richardson, Texas 75080, United States*

Olexandr Isayev − *Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0001-7581-8497

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c01052

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.H. and L.V.S. acknowledge the support of the National Science Foundation (award number CHE-2102639). O.I. acknowledges the support from the National Science Foundation (award number: CHE-2154447). 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 was supported in part through computational resources provided by Information Technology at Purdue, West Lafayette, Indiana.

■ REFERENCES

(1) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60* (7), 3408−3415.

(2) Fine, J.; Liu, J. K.-Y.; Beck, A.; Alzarieni, K. Z.; Ma, X.; Boulos, V. M.; Kenttämaa, H. I.; Chopra, G. Graph-Based Machine Learning Interprets and Predicts Diagnostic Isomer-Selective Ion−Molecule Reactions in Tandem Mass Spectrometry. *Chem. Sci.* **2020**, *11* (43), 11849−11858.

(3) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370−377.

(4) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(5) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31* (9), 3564−3572.

(6) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; et al. Schnet−a Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, No. 241722.
(7) 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), No. 2903.
(8) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192−3203.
(9) 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), No. 241733.
(10) Liu, Z.; Zubatiuk, T.; Roitberg, A.; Isayev, O. Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2022**, *62* (22), 5373−5382.
(11) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16* (7), 4192−4202.
(12) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), No. 146401.
(13) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121* (16), 10037−10072.
(14) Huang, S.-D.; Shang, C.; Zhang, X.-J.; Liu, Z.-P. Material Discovery by Combining Stochastic Surface Walking Global Optimization with a Neural Network. *Chem. Sci.* **2017**, *8* (9), 6327−6337.
(15) Hart, G. L. W.; Mueller, T.; Toher, C.; Curtarolo, S. Machine Learning for Alloys. *Nat. Rev. Mater.* **2021**, *6* (8), 730−755.
(16) Inizan, T. J.; Plé, T.; Adjoua, O.; Ren, P.; Gökcan, H.; Isayev, O.; Lagardère, L.; Piquemal, J.-P. Scalable Hybrid Deep Neural Networks/Polarizable Potentials Biomolecular Simulations Including Long-Range Effects. *Chem. Sci.* **2023**, *14* (20), 5438−5452.
(17) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112* (1), 632−672.
(18) Gordon, M. S.; Smith, Q. A.; Xu, P.; Slipchenko, L. V. Accurate First Principles Model Potentials for Intermolecular Interactions. *Annu. Rev. Phys. Chem.* **2013**, *64* (1), 553−578.
(19) Slipchenko, L. V. Effective Fragment Potential Method. In *Many-Body Effects and Electrostatics in Biomolecules*, 1st ed.; Cui, Q.; Meuwly, M.; Ren, P., Eds.; Jenny Stanford Publishing: New York, 2016; pp 147−190.
(20) Slipchenko, L. V.; Gurunathan, P. K. Effective Fragment Potential Method: Past, Present, and Future. In *Fragmentation*; Gordon, M. S., Ed.; Wiley, 2017; pp 183−208.
(21) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. The Effective Fragment Potential Method: A QM-Based MM Approach to Modeling Environmental Effects in Chemistry. *J. Phys. Chem. A* **2001**, *105* (2), 293−307.
(22) Nanda, K. D.; Krylov, A. I. The Effect of Polarizable Environment on Two-Photon Absorption Cross Sections Characterized by the Equation-of-Motion Coupled-Cluster Singles and Doubles Method Combined with the Effective Fragment Potential Approach. *J. Chem. Phys.* **2018**, *149* (16), No. 164109.
(23) Sen, R.; Dreuw, A.; Faraji, S. Algebraic Diagrammatic Construction for the Polarisation Propagator in Combination with Effective Fragment Potentials. *Phys. Chem. Chem. Phys.* **2019**, *21* (7), 3683−3694.
(24) Arora, P.; Slipchenko, L. V.; Webb, S. P.; DeFusco, A.; Gordon, M. S. Solvent-Induced Frequency Shifts: Configuration Interaction Singles Combined with the Effective Fragment Potential Method. *J. Phys. Chem. A* **2010**, *114* (25), 6742−6750.
(25) DeFusco, A.; Minezawa, N.; Slipchenko, L. V.; Zahariev, F.; Gordon, M. S. Modeling Solvent Effects on Electronic Excited States. *J. Phys. Chem. Lett.* **2011**, *2* (17), 2184−2192.
(26) Slipchenko, L. V. Solvation of the Excited States of Chromophores in Polarizable Environment: Orbital Relaxation versus Polarization. *J. Phys. Chem. A* **2010**, *114* (33), 8824−8830.
(27) Ghosh, D.; Isayev, O.; Slipchenko, L. V.; Krylov, A. I. Effect of Solvation on the Vertical Ionization Energy of Thymine: From Microhydration to Bulk. *J. Phys. Chem. A* **2011**, *115* (23), 6028−6038.
(28) Kosenkov, D.; Slipchenko, L. V. Solvent Effects on the Electronic Transitions of *p* -Nitroaniline: A QM/EFP Study. *J. Phys. Chem. A* **2011**, *115* (4), 392−401.
(29) Kim, Y.; Mitchell, Z.; Lawrence, J.; Morozov, D.; Savikhin, S.; Slipchenko, L. V. Predicting Mutation-Induced Changes in the Electronic Properties of Photosynthetic Proteins from First Principles: The Fenna−Matthews−Olson Complex Example. *J. Phys. Chem. Lett.* **2023**, *14* (31), 7038−7044.
(30) Tazhigulov, R. N.; Gurunathan, P. K.; Kim, Y.; Slipchenko, L. V.; Bravaya, K. B. Polarizable Embedding for Simulating Redox Potentials of Biomolecules. *Phys. Chem. Chem. Phys.* **2019**, *21* (22), 11642−11650.
(31) Kim, Y.; Morozov, D.; Stadnytskyi, V.; Savikhin, S.; Slipchenko, L. V. Predictive First-Principles Modeling of a Photosynthetic Antenna Protein: The Fenna−Matthews−Olson Complex. *J. Phys. Chem. Lett.* **2020**, *11* (5), 1636−1643.
(32) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput.-Aided Mol. Des.* **2014**, *28* (7), 711−720.
(33) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48* (7), 1198−1229.
(34) Gordon, M. S.; Schmidt, M. W. Advances in Electronic Structure Theory: GAMESS a Decade Later. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E.; Frenking, G.; Kim, K. S.; Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 1167−1189.
(35) Barca, G. M. J.; Bertoni, C.; Carrington, L.; Datta, D.; De Silva, N.; Deustua, J. E.; Fedorov, D. G.; Gour, J. R.; Gunina, A. O.; Guidez, E.; Harville, T.; Irle, S.; Ivanic, J.; Kowalski, K.; Leang, S. S.; Li, H.; Li, W.; Lutz, J. J.; Magoulas, I.; Mato, J.; Mironov, V.; Nakata, H.; Pham, B. Q.; Piecuch, P.; Poole, D.; Pruitt, S. R.; Rendell, A. P.; Roskop, L. B.; Ruedenberg, K.; Sattasathuchana, T.; Schmidt, M. W.; Shen, J.; Slipchenko, L.; Sosonkina, M.; Sundriyal, V.; Tiwari, A.; Vallejo, J. L. G.; Westheimer, B.; Włoch, M.; Xu, P.; Zahariev, F.; Gordon, M. S. Recent Developments in the General Atomic and Molecular Electronic Structure System. *J. Chem. Phys.* **2020**, *152*, No. 154102.
(36) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14* (11), 1347−1363.
(37) Kim, Y.; Bui, Y.; Tazhigulov, R. N.; Bravaya, K. B.; Slipchenko, L. V. Effective Fragment Potentials for Flexible Molecules: Transferability of Parameters and Amino Acid Database. *J. Chem. Theory Comput.* **2020**, *16* (12), 7735−7747.
(38) Stone, A. J. *The Theory of Intermolecular Forces*, 2nd ed.; Oxford University Press: Oxford, 2013.
(39) Rojas, C. I. V.; Slipchenko, L. V. Exchange Repulsion in Quantum Mechanical/Effective Fragment Potential Excitation Energies: Beyond Polarizable Embedding. *J. Chem. Theory Comput.* **2020**, *16* (10), 6408−6417.
(40) Rojas, C. I. V.; Fine, J.; Slipchenko, L. V. Exchange-Repulsion Energy in QM/EFP. *J. Chem. Phys.* **2018**, *149* (9), No. 094103.
(41) Smith, Q. A.; Ruedenberg, K.; Gordon, M. S.; Slipchenko, L. V. The Dispersion Interaction between Quantum Mechanics and Effective Fragment Potential Molecules. *J. Chem. Phys.* **2012**, *136* (24), No. 244107.
(42) Slipchenko, L. V.; Gordon, M. S.; Ruedenberg, K. Dispersion Interactions in QM/EFP. *J. Phys. Chem. A* **2017**, *121* (49), 9495−9507.

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(43) Jensen, J. H.; Gordon, M. S. An Approximate Formula for the Intermolecular Pauli Repulsion between Closed Shell Molecules. *Mol. Phys.* **1996**, *89* (5), 1313−1325.
(44) Jensen, J. H.; Gordon, M. S. An Approximate Formula for the Intermolecular Pauli Repulsion between Closed Shell Molecules. II. Application to the Effective Fragment Potential Method. *J. Chem. Phys.* **1998**, *108* (12), 4772−4782.
(45) Adamovic, I.; Gordon, M. S. Dynamic Polarizability, Dispersion Coefficient C6 and Dispersion Energy in the Effective Fragment Potential Method. *Mol. Phys.* **2005**, *103* (2−3), 379−387.
(46) Behler, J. Constructing High-Dimensional Neural Network Potentials: A Tutorial Review. *Int. J. Quantum Chem.* **2015**, *115* (16), 1032−1050.
(47) Kaliman, I. A.; Slipchenko, L. V. Hybrid MPI/OpenMP Parallelization of the Effective Fragment Potential Method in the *Libefp* Software Library. *J. Comput. Chem.* **2015**, *36* (2), 129−135.
(48) Kaliman, I. A.; Slipchenko, L. V. LIBEFP: A New Parallel Implementation of the Effective Fragment Potential Method as a Portable Software Library. *J. Comput. Chem.* **2013**, *34* (26), 2284−2292.
(49) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.
(50) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225−11236.
(51) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463−1472.
(52) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Landau, A.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, M. A.; Steele, R. P.; Sundstrom, E. J.; Woodcock, H. L.; Zimmerman, P. M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, K. B.; Brown, S. T.; Casanova, D.; Chang, C.-M.; Chen, Y.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Diedenhofen, M.; DiStasio, R. A.; Do, H.; Dutoi, A. D.; Edgar, R. G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.-C.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V.; Levchenko, S. V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.-P.; Mardirossian, N.; Marenich, A. V.; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; O'Neill, D. P.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Prociuk, A.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.-C.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vogt, L.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; You, Z.-Q.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, G. K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; Schaefer, H. F.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, J.-D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Gwaltney, S. R.; Hsu, C.-P.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Slipchenko, L. V.; Subotnik, J. E.; Van Voorhis, T.; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113* (2), 184−215.
(53) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L.; Gan, Z.; Hait, D.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Kussmann, J.; Lange, A. W.; Lao, K. U.; Levine, D. S.; Liu, J.; McKenzie, S. C.; Morrison, A. F.; Nanda, K. D.; Plasser, F.; Rehn, D. R.; Vidal, M. L.; You, Z.-Q.; Zhu, Y.; Alam, B.; Albrecht, B. J.; Aldossary, A.; Alguire, E.; Andersen, J. H.; Athavale, V.; Barton, D.; Begam, K.; Behn, A.; Bellonzi, N.; Bernard, Y. A.; Berquist, E. J.; Burton, H. G. A.; Carreras, A.; Carter-Fenk, K.; Chakraborty, R.; Chien, A. D.; Closser, K. D.; Cofer-Shabica, V.; Dasgupta, S.; de Wergifosse, M.; Deng, J.; Diedenhofen, M.; Do, H.; Ehlert, S.; Fang, P.-T.; Fatehi, S.; Feng, Q.; Friedhoff, T.; Gayvert, J.; Ge, Q.; Gidofalvi, G.; Goldey, M.; Gomes, J.; González-Espinoza, C. E.; Gulania, S.; Gunina, A. O.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A.; Herbst, M. F.; Vera, M. H.; Hodecker, M.; Holden, Z. C.; Houck, S.; Huang, X.; Hui, K.; Huynh, B. C.; Ivanov, M.; Jász, Á.; Ji, H.; Jiang, H.; Kaduk, B.; Kähler, S.; Khistyaev, K.; Kim, J.; Kis, G.; Klunzinger, P.; Koczor-Benda, Z.; Koh, J. H.; Kosenkov, D.; Koulias, L.; Kowalczyk, T.; Krauter, C. M.; Kue, K.; Kunitsa, A.; Kus, T.; Ladjánszki, I.; Landau, A.; Lawler, K. V.; Lefrancois, D.; Lehtola, S.; Li, R. R.; Li, Y.-P.; Liang, J.; Liebenthal, M.; Lin, H.-H.; Lin, Y.-S.; Liu, F.; Liu, K.-Y.; Loipersberger, M.; Luenser, A.; Manjanath, A.; Manohar, P.; Mansoor, E.; Manzer, S. F.; Mao, S.-P.; Marenich, A. V.; Markovich, T.; Mason, S.; Maurer, S. A.; McLaughlin, P. F.; Menger, M. F. S. J.; Mewes, J.-M.; Mewes, S. A.; Morgante, P.; Mullinax, J. W.; Oosterbaan, K. J.; Paran, G.; Paul, A. C.; Paul, S. K.; Pavošević, F.; Pei, Z.; Prager, S.; Proynov, E. I.; Rák, Á.; Ramos-Cordoba, E.; Rana, B.; Rask, A. E.; Rettig, A.; Richard, R. M.; Rob, F.; Rossomme, E.; Scheele, T.; Scheurer, M.; Schneider, M.; Sergueev, N.; Sharada, S. M.; Skomorowski, W.; Small, D. W.; Stein, C. J.; Su, Y.-C.; Sundstrom, E. J.; Tao, Z.; Thirman, J.; Tornai, G. J.; Tsuchimochi, T.; Tubman, N. M.; Veccham, S. P.; Vydrov, O.; Wenzel, J.; Witte, J.; Yamada, A.; Yao, K.; Yeganeh, S.; Yost, S. R.; Zech, A.; Zhang, I. Y.; Zhang, X.; Zhang, Y.; Zuev, D.; Aspuru-Guzik, A.; Bell, A. T.; Besley, N. A.; Bravaya, K. B.; Brooks, B. R.; Casanova, D.; Chai, J.-D.; Coriani, S.; Cramer, C. J.; Cserey, G.; DePrince, A. E., III; DiStasio, R. A., Jr.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Goddard, W. A., III; Hammes-Schiffer, S.; Head-Gordon, T.; Hehre, W. J.; Hsu, C.-P.; Jagau, T.-C.; Jung, Y.; Klamt, A.; Kong, J.; Lambrecht, D. S.; Liang, W.; Mayhall, N. J.; McCurdy, C. W.; Neaton, J. B.; Ochsenfeld, C.; Parkhill, J. A.; Peverati, R.; Rassolov, V. A.; Shao, Y.; Slipchenko, L. V.; Stauch, T.; Steele, R. P.; Subotnik, J. E.; Thom, A. J. W.; Tkatchenko, A.; Truhlar, D. G.; Van Voorhis, T.; Wesolowski, T. A.; Whaley, K. B.; Woodcock, H. L., III; Zimmerman, P. M.; Faraji, S.; Gill, P. M. W.; Head-Gordon, M.; Herbert, J. M.; Krylov, A. I. Software for the Frontiers of Quantum Chemistry: An Overview of Developments in the Q-Chem 5 Package. *J. Chem. Phys.* **2021**, *155* (8), No. 084801.
(54) Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128* (8), No. 084106.
(55) Haghiri, S.; Rojas, C. V.; Slipchenko, L. V. ANI-EFP neural network. https://github.com/Slipchenko-Group/ANIEFP (accessed September 15, 2024).
(56) TorchANI. https://aiqm.github.io/torchani/ (accessed September 1, 2024).
(57) Zinovjev, K. Electrostatic Embedding of Machine Learning Potentials. *J. Chem. Theory Comput.* **2023**, *19* (6), 1888−1897.
(58) Zubatyuk, R.; Smith, J. S.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a Neural Network to Attach and Detach Electrons from Molecules. *Nat. Commun.* **2021**, *12* (1), No. 4870.
(59) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet Your Neutral, Charged, Organic, and Elemental-Organic Needs *ChemRxiv* 2023.

https://doi.org/10.1021/acs.jctc.4c01052
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX