JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

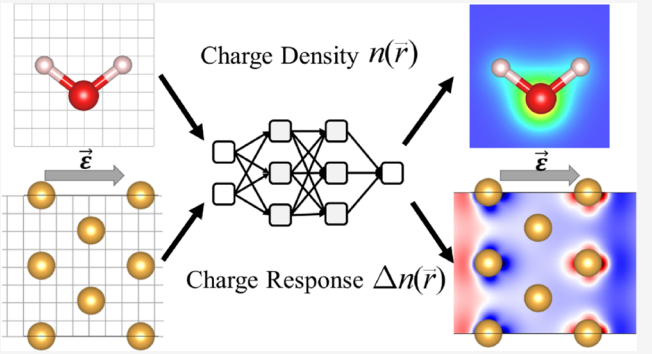
Efficient Sampling for Machine Learning Electron Density and Its Response in Real Space

Chaoqiang Feng, Yaolong Zhang, and Bin Jiang*

Cite This: J. Chem. Theory Comput. 2025, 21, 691−702 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Electron density is a fundamental quantity that can in principle determine all ground state electronic properties of a given system. Although machine learning (ML) models for electron density based on either an atom-centered basis or a real-space grid have been proposed, the demand for a number of high-order basis functions or grid points is enormous. In this work, we propose an efficient grid-point sampling strategy that combines targeted sampling favoring a large density and a screening of grid points associated with linearly independent atomic features. This new sampling strategy is integrated with a field-induced recursively embedded atom neural network model to develop a real-space grid-based ML model for the electron density and its response to an electric field. This approach is applied to a QM9 molecular data set, a H_2O/Pt(111) interfacial system, an Au(100) electrode, and an Au nanoparticle under an electric field. The number of training points is found to be much smaller than previous models, while yielding comparably accurate predictions for the electron density of the entire grid. The resultant machine-learned electron density model enables us to properly partition partial charge onto each atom and analyze the charge variation upon proton transfer in the H_2O/Pt(111) system. The machine-learning electronic response model allows us to predict charge transfer and the electrostatic potential change induced by an electric field applied to an Au(100) electrode or an Au nanoparticle.



1. INTRODUCTION

Electron density, $\mathbf{n}(\mathbf{r})$, is one of the most important and fundamental variables of physical and chemical systems. According to density functional theory (DFT), electron density can in principle uniquely determine all ground state properties of a given system, such as total energy, dipole moment, and higher-order multipoles.[1] Computing $\mathbf{n}(\mathbf{r})$ for a given configuration often requires to solve an electronic structure problem, e.g., by DFT. It is typically implemented in a self-consistent way, and a reasonable estimate of an initial $\mathbf{n}(\mathbf{r})$ can in turn accelerate the solution of the Kohn−Sham (KS) equation in DFT.[2] Additionally, an accurate $\mathbf{n}(\mathbf{r})$ is a prerequisite for applying a certain charge partitioning scheme to analyze the electron transfer in redox reactions or proton transfer processes.

Machine learning (ML) has become increasingly important in modern chemistry.[3−9] For instance, machine-learned potentials (MLPs) have been widely used in molecular dynamics simulations of molecules, condensed phase materials, and interfacial systems.[10−31] Most of these MLPs learn the relationship between chemical structure and potential energy by generating atomic descriptors that satisfy translational, rotational, and permutational invariance with respect to energy and expressing total energy as a sum of atomic contributions for better scaling and generalizability.[32,33]

Electron density is inherently more information-rich than total energy, so that a range of important electronic properties beyond total energy can be obtained directly from ML models that learn from electron density. Several machine-learned electron density (MLED) models have been developed.[34−61] In particular, $\mathbf{n}(\mathbf{r})$ is a three-dimensional probability distribution function in space whose value is dependent on the relative spatial position and nuclear configuration. A key inductive bias for these MLED models is the choice of the representation of $\mathbf{n}(\mathbf{r})$.[62] For example, Brockherde et al. and Bogojeski et al. first used a plane-wave basis representation of $\mathbf{n}(\mathbf{r})$ and independent kernel ridge models to regress individual basis function coefficients.[34,43] This scheme has later been extended to predict excited-state electron density.[63] While the choice of a plane-wave basis representation in these studies allows a systematic convergence of the electron density with an increase in the number of basis functions, it is less efficient for

Received: October 10, 2024
Revised: December 17, 2024
Accepted: December 19, 2024
Published: January 3, 2025

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

anisotropic systems and not easily transferable to other molecules. Alternatively, Grisafi et al. chose to expand $\mathbf{n}(\mathbf{r})$ in localized atom-centered basis functions and developed a transferable model based on symmetry-adapted Gaussian process regression (SA-GPR), which is named as the symmetry-adapted learning of three-dimensional electron densities (SALTED).[36,38,49,56,59,64] Specifically, the atom-centered basis function is a combination of radial functions and spherical harmonics, whose corresponding coefficients are correlated with its local environment and learned by SA-GPR to preserve the correct symmetry of $\mathbf{n}(\mathbf{r})$. This SA-GPR representation for electron density has been employed successfully in constructing MLED models for small molecules,[64] noncovalent systems,[36] and condensed phases.[49] However, it requires an initial decomposition of the charge density onto predefined basis sets, which may introduce additional errors. del Rio et al. have partially overcome this limitation by learning an optimal basis set together with coefficients to minimize the differences between the ML-predicted densities, which are obtained by projecting atomic bases onto a real-space grid, with the actual DFT charge densities.[57] In addition, Rackers et al. proposed an equivariant graph convolutional neural network (GCNN) to predict the electron density of water clusters[53] and biomolecules.[52] It was found necessary to increase the order of the atom-centered basis set for accurate prediction, which led to a rapid increase in computational cost.

On the other hand, most DFT codes directly output electron density values in a real-space grid.[65,66] As a result, one can conveniently learn these discrete values in the same spirit of learning potential energies and obtain a smooth function for $\mathbf{n}(\mathbf{r})$. The grid-based MLED models were first proposed by us in representing the embedded electron density to compute electronic friction at metal surfaces by viewing each grid point as a virtual atom.[67,68] More recently, Jørgensen and Bhowmik constructed a grid-based equivariant message passing neural network (MPNN) model based on a uniform grid-point sampling scheme, referred to as DeepDFT, which exhibits superior accuracy than typical basis-based models.[51] Sunshine et al. further improved the DeepDFT model by enforcing charge balance and directly assigning zero density to grid points far away from any atoms.[58] The modified DeepDFT model was trained on a subset of the very large Open Catalyst 2020 (OC20) data set[69] containing 56 different chemical elements, demonstrating the generalizability of grid-based MLED models.

In addition to learning the electron density of isolated systems, describing the response of electron density to an external electric field, i.e., the charge difference with and without the field, is also of great importance, for example, in electrochemistry. Very recently, Grisafi et al.[59] and Lewis et al.[70] independently developed similar ML models for electron density response (MLEDR) to a finite field based on the SA-GPR atomic basis representation for electron density.[49] Both models introduce the field-dependence to kernels, and Grisafi et al. further include a long-distance equivariant descriptor to capture the nonlocal feature of the electron density response in metal electrodes.[59] Thus far, however, no real-space MLEDR models have been reported.

Indeed, a disadvantage of real-space grid-based MLED models is that they typically rely on a huge number of data points, causing substantial memory demand. This issue becomes more severe when the configuration space or the chemical space of the target system(s) is more complicated. To overcome this difficulty, Focassio et al. proposed a targeted sampling (TS) strategy, which assigns a probability of point selection based on a normal distribution of the inverse of the charge density, which effectively samples more points with larger densities.[60] Based on this TS strategy, they were able to train an accurate model with just about 0.1% of the available grid points for individual molecular and material systems.

The purpose of this work is 2-fold. The first goal is to further reduce the required number of grid points for training a grid-based MLED model to make it suitable for a large group of molecules or a complicated system involving a broad chemical or configuration space. The second goal is to extend the previously developed field-induced recursive embedded atom neural network (FIREANN) potential model[71−74] to describe electron density in the absence and presence of an electric field and thus its response. To this end, we leverage the concept of feature or structure selection in constructing MLPs[56,75,76] and propose an efficient two-step grid-point sampling strategy. First, the entire charge density mesh is sparsified using the density value-based TS. Second, the remaining points are screened based on their structural similarity in terms of associated grid-centered structural features. This sampling strategy is coupled with our FIREANN framework to construct an accurate MLED model using an extremely low fraction of data, 0.005 −0.015% of the entire charge density mesh. Numerical tests in the well-known QM9 molecular data set and a liquid−solid interfacial H_2O/Pt(111) system validate this strategy, and the resultant MLED model is found useful to serve as the foundation of the charge analysis. Moreover, the FIREANN model naturally introduces the field dependence of charge density and the nonlocal effect, enabling an accurate description of the charge density response of an Au(100) electrode.

2. METHOD

2.1. Field-Induced Recursively Embedded Atom Neural Network Model for Electron Density and Response. As mentioned above, $\mathbf{n}(\mathbf{r})$ is a three-dimensional spatial function, which in real-space is typically represented by discrete and scalar values on a dense grid in the simulation box in DFT codes. As proposed in our previous studies, the three-dimensional function $\mathbf{n}(\mathbf{r})$ of a system with N atoms is formally analogous to a potential energy surface (PES) of a system with $N + N_g$ atoms, where the extra N_g atoms are ghost atoms.[67,68] These ghost atoms have no physical meaning but represent the grid point position. As a result, the grid-centered local environments for both realistic atoms and virtual atoms can be described by conventional many-body atomic features, which are frequently used for constructing atomistic MLPs. Note that the local environment of ghost atoms is only related to the surrounding realistic atoms and is mapped to the charge density or response values through an atomic neural network. Here, we adopt the FIREANN approach,[72,73] which employs field-induced embedded atom density (FI-EAD) features to describe the atomic environment and system-field interactions.[71] Specifically, the nth FI-EAD feature of a central atom (including the virtual atom) i is defined by the square of the linear combination of all neighboring atomic orbitals and field-dependent orbital,

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

$$\rho_i^n = \sum_{l=0}^{L} \sum_{l_x,l_y,l_z}^{l_x+l_y+l_z=l} \frac{l!}{l_x!l_y!l_z!} \left[\sum_{j\neq i}^{N_\varphi} d_m^n \left(\sum_{m=1}^{N_c} c_j \varphi_{l_x l_y l_z}^m(\vec{r}_{ij}) + c_\varepsilon \varphi_{l_x l_y l_z}(\vec{\varepsilon}_i) \right) \right]^2 \quad (1)$$

where N_c is the number of neighbor atoms, c_j is the combination coefficient of the jth neighbor atom, and its corresponding atomic orbital is obtained by the contraction (the second linear combination in eq 1 of N_φ primitive Gaussian-type orbitals (GTO, $\varphi_{l_x l_y l_z}^m(\hat{r}_{ij})$), with d_m^n being the contraction coefficient of the nth FI-EAD feature and mth GTO. A primitive GTO at atom j is characterized by its center (r_s), width (σ), and angular momentum ($l = l_x + l_y + l_z$) as,

$$\varphi_{l_x l_y l_z}^m(\hat{r}_{ij}) = (x_{ij})^{l_x}(y_{ij})^{l_y}(z_{ij})^{l_z} \exp\left[-\frac{(r_{ij} - r_s)^2}{2\sigma^2} \right] f_c(r_{ij}) \quad (2)$$

where $\hat{r}_{ij} = \hat{r}_i - \hat{r}_j$ is the relative position vector, with r_{ij}, x_{ij}, y_{ij}, and z_{ij} being its norm and three Cartesian components, and $f_c(r_{ij})$ is a cosine-type cutoff function. The virtual field vector-dependent is defined as,

$$\varphi_{l_x l_y l_z}(\vec{\varepsilon}) = (\varepsilon_x)^{l_x}(\varepsilon_y)^{l_y}(\varepsilon_z)^{l_z} \quad (3)$$

Note that in practice, the summation order is exchanged in eq 1 for faster evaluation. Varying these hyperparameters forms an array of FI-EAD features, which encode three-body interactions implicitly by the summation of $l > 0$ terms with $\sim O(N_c)$ scaling. To incorporate higher-order and more nonlocal interactions, c_j can be also the output of an atomic NN that depends on the jth atom's local environment and can be updated iteratively, namely,

$$c_j^t = g_j^{t-1}[\boldsymbol{\rho}_j^{t-1}(\mathbf{c}_j^{t-1}, \mathbf{r}_j^{t-1})] \quad (4)$$

where g_j^{t-1} represents the jth atomic NN module in the tth iteration. This leads to a message-passing NN architecture, which has been successfully applied to learn PESs and response properties of molecular and condensed phase systems.[72,73] It should be noted that the learning target of the MLED model is the electron densities in the absence of an electric field, and that of the MLEDR model is defined as the difference between the perturbed electron densities and the electron densities of the corresponding isolated electrodes, i.e., $\Delta n_e = n_e - n_0$.

Note that the FI-EAD feature vector here is used to describe the atomic environment and does not represent the true electron density of the atom system. For clarity, we use $\{\boldsymbol{\rho}\}$ to denote the FI-EAD features (or set) and $\boldsymbol{n}$ to denote the charge density of systems. The corresponding hyperparameters of FI-EAD features are optimized iteratively during the cycle of sampling and training.

2.2. Linearly Independent Feature-Based Grid-Point Sampling. A remarkable difference between training a PES and $\mathbf{n}(\mathbf{r})$ in real space is that every single configuration on the PES provides a dense 3D grid of density data. In practice, this leads to highly redundant data points, and using the entire mesh is neither efficient nor necessary. Because $\mathbf{n}(\mathbf{r})$ is near zero in a wide range of space, e.g., a position far from any nuclei, a random sampling (RS) of grid points would sample too many near-zero density values, leading to inefficient coverage of the most probable region of $\mathbf{n}(\mathbf{r})$. In this regard, Jørgensen and Bhowmik randomly sampled 1000 grid points for each configuration to train their equivariant MLED model and found that the removal of some low-density points can be beneficial.[51] Later, Focassio et al. proposed a probabilistic TS procedure that tends to sample more grid points with high densities, by which an accurate MLED model can be constructed with just 0.1% of the entire mesh.[60] However, significant redundancy remains in the resulting data set as many of these grid points would have similar local environments due to symmetry, especially for small molecules with high symmetry and crystals. As a result, we propose to select only those grid points that have their atomic features linearly uncorrelated with others. A similar strategy was previously applied to select a minimal number of linearly independent atomic features that best represent the local environment in constructing EANN potentials.[76]

The proposed two-step grid-point sampling workflow is illustrated in Figure 1. The three-dimensional grid generated

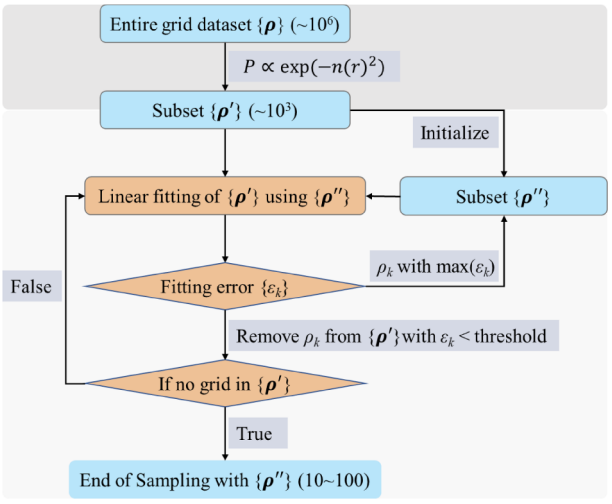


Figure 1. A schematic workflow of the proposed two-step grid-point sampling strategy.

by DFT typically consists of a few million points, which is our initial data set. We first perform the TS strategy of Focassio et al. to assign a higher probability for sampling points with large densities.[60] For a grid point candidate $\boldsymbol{r}_g$, the relative sampling probability is given by,

$$P(\boldsymbol{r}_g) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1/\mathbf{n}(\boldsymbol{r}_g))^2}{2\sigma^2} \right) \quad (5)$$

where the hyperparameter σ controls the width of the probability distribution. The smaller the σ, the higher the concentration of sampled grid points is in the high-density region. After collecting a subset with, say, N_g grid points by TS, we calculate their grid-centered FI-EAD features and perform a second-step linearly independent feature-based sampling (LIFS) to judge their similarity. The first point is randomly selected out from the subset $\{\boldsymbol{\rho}'\}$ and add into the $\{\boldsymbol{\rho}''\}$, then every later candidate's FI-EAD feature vector will be linearly fitted by these feature vector(s) of the selected data point(s), and only candidates with linearly independent features are accepted and added into $\{\boldsymbol{\rho}''\}$. For example, for the kth candidate, the deviation of the linear least-squares by feature vectors of these accepted points (ε_k) is normalized to be used

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

as the criterion of linear independence of the candidate's feature vector $\boldsymbol{\rho}_{k'}$

$$\varepsilon_k = \frac{\left\| \rho_k - \sum_{N_s}^{i} \rho_i \gamma_i \right\|}{\|\rho_k\|} \tag{6}$$

where γ_i represents the least-squares coefficient. A candidate with a large ε_k implies that its atomic feature is dissimilar to those of selected points and should be thus added. In contrast, the grid point with a small ε_k indicates that its atomic feature can be well represented by the linear combination of other features and is thus not necessary. This procedure can be performed iteratively until all points with linearly independent features are sorted out. In practice, we find that on average merely 10−100 grid points for each nuclear configuration are necessary, which makes it possible to train an MLED model for a complex system or a mix of many subsystems at an affordable cost.

2.3. Data Sets and Training Setup. We have tested the performance of the proposed sampling strategy in learning electron density with distinct data sets. The first one is the QM9 data set,[77,78] which contains 133,885 small organic molecules and is widely used for benchmarking ML models for predicting molecular properties. The electron density of this QM9 data set was taken from previous work.[51,79] In addition, we prepared an independent test data set comprising 1,000 molecules, each containing 10 heavy atoms and consisting of four elements (C, H, O, N), to evaluate the model's generalization capability beyond the QM9 data set. The second one is a liquid−solid interfacial H_2O/Pt(111) system, where the surface was represented by a four-layer Pt(111) periodic slab using a (4×4) supercell, initially separated by an 18 Å vacuum space in the vertical direction. The electrochemical interface was then built up by fully filling the vacuum space between repeated slabs with 60 water molecules. A total of 2000 configurations were generated by running ab initio molecular dynamics (AIMD) simulations with a thermostat temperature of 300 K. For all three data sets, the electron densities were calculated with Vienna Ab initio Simulation Package (VASP),[80,81] and the Perdew−Burke−Ernzerhof (PBE) density functional[82] and projector augmented wave (PAW) method[83] were employed. Moreover, we used the PBE0 density functional[84] for an additional test data set to investigate the charge density difference by using different DFT functionals. The wave function of valence electrons was expanded using plane waves with energy cutoffs of 400, 400, and 600 eV for the QM9 data set, additional test data set, and the H_2O/Pt(111) data set, respectively. The first Brillouin zone was sampled only at the Gamma center for the QM9 data set and additional data set, and with a $3 \times 3 \times 1$ Monkhorst−Pack k-point mesh for the H_2O/Pt(111) data set.[85] The charge densities were output on a three-dimensional grid evenly spaced by 0.08−0.10 Å. Besides, the Au(100) electrode slab data set consists of 260 slab configurations of Au(100) taken from previous work,[59] with 3−17 metal layers in a (2×2) supercell. The charge responses were computed with the open-source CP2K package with the PBE density functional, Goedecker−Teter−Hutter (GTH) pseudopotentials, and double-ζ basis sets with one set of polarization functions (DZVP).[66,86−90] The charge transfer at the two metal surfaces was calculated under a uniform electric field along z, and the field intensity was set to −1.0 V/Å. Furthermore, to validate our method on a metallic system with a more complex geometry, we constructed a data set of Au nanoparticles. Each nanoparticle was generated using the Wulff construction to minimize the surface energy consisting of 85 Au atoms. Random displacements were introduced to the atomic positions and sampled from a normal distribution with a standard deviation of 1% of the lattice constant along the three Cartesian directions, resulting in a total of 10 nanoparticle configurations. The charge response of these nanoparticles was calculated using the same parameter settings as for the Au electrode data set, but the direction of the electric field was reversed.

Note that all of the electron density and response used for training are limited to the valence electron density. The all-electron charge density, possessing sharp peaks near nucleus regions, is more challenging to learn. One potential solution is to use Gaussian basis functions to approximate the density distribution of inner electrons and then subtract their contributions from the all-electron charge density so that machine learning for the valence charge density remains unchanged. Afterward, all-electron charge density can be recovered by adding the inner contribution back to the machine-learned valence charge density. The average number of grid points per configuration is ∼0.7 million for the QM9 data set and ∼5.4 million for the H_2O/Pt(111) data set. In practice, before sampling grid points for training MLED models, 10,000 molecules from the QM9 data set and 200 configurations from the H_2O/Pt(111) data set were selected randomly as the test set, respectively. We performed the TS in the residual configurations to select 1000 grid points per molecule in the QM9 data set and 5000 grid points per configuration in the H_2O/Pt(111) data set with a width parameter σ of 30 and 80 Å^3 e^{-1}, respectively. A subsequent LIFS was performed on these collections of grid points. The final data set sampled by this two-step procedure was divided into the training and validation sets with a ratio of 9:1. See Table S1 for a detailed description of the data set generation process and segmentation scheme. The training process was optimized using the AdamW algorithm.[91] The learning rate, initially set to 0.001, decayed by a factor of 0.5 whenever the validation error did not decrease for 50 epochs. Training was stopped when the learning rate dropped below 1×10^{-6}. More information about the architecture of the FIREANN model is given in Table 1.

3. RESULTS AND DISCUSSION

3.1. Performance of the MLED Model. *3.1.1. QM9 Data Set.* Let us first discuss the prediction accuracy of the FIREANN models for $\mathbf{n}(\mathbf{r})$ of molecules in the QM9 data set that cover a vast chemical space. The proposed TS + LIFS strategy was used to sample ∼100 grid points per molecule, collecting ∼13 million grid points. Compared to the entire grid representing $\mathbf{n}(\mathbf{r})$ for all molecules in the data set, the FIREANN model was thus trained with only about 0.014% of available grid points. The correlations between DFT calculated electron densities and FIREANN predictions for different data sets are illustrated in Figure 2. Note in the test set that grid points are more distributed in the low-density region, which is a natural consequence of the even distribution of grid points and a large area filled with near-zero electron density. Indeed, more than half of the grid points in the QM9 data set have charge densities lower than 10^{-3} e·Å^{-3}. While the TS algorithm favors more high-density points, the LIFS algorithm excludes these grid points with similar local environments, and as a

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

Table 1. NN Structures (the Number of Neurons in Each Hidden Layer), Cutoff Radii (in Å), Maximum Angular Momentum, the Number of Message Passing Iterations, Batch Size, and the Number of FI-EAD Features Used in Training FIREANN Electron Density Models for QM9 and H_2O/Pt(111) Datasets and Electron Response Model for Au Electrode and Au Nanoparticle Dataset

	System		
Parameter	QM9	H_2O/Pt(111)	Au electrode and Au nanoparticle
NN structure for atomic properties	32 × 32	64 × 64	128 × 128
NN structure for orbital coefficients	16 × 16	32 × 32	64 × 64
Cutoff radius/Å	4.0	4.0	6.0
Maximum angular momentum	2	2	2
Number of messages-passing iterations	2	2	6
Batch size	16	8	8
Number of features	108	108	234

result, the charge densities in the training set and validation set become more evenly distributed over the whole range between 0 and 14 e·Å^{-3}. The optimized FIREANN model yields a very low root-mean-square error (RMSE) of 0.0011 e·Å^{-3} for the test set with an extremely high correlation coefficient (R^2 = 0.999986). This is even lower than the RMSE for the validation set (0.0016 e·Å^{-3}) and for the training set (0.0019 e·Å^{-3}), indicating that the FIREANN model can accurately predict $\mathbf{n}(\mathbf{r})$ in the entire mesh, including low-density regions that are uncommon in the training data set.

Figure 3 displays the error iso-surfaces for several representative molecules in the QM9 data set. These molecules comprise common chemical elements (i.e., carbon, hydrogen, oxygen, and nitrogen) and prevalent chemical interactions in the QM9 data set. The prediction error for each molecule in the entire simulation box is evaluated with the normalized mean absolute error (MAE), which was introduced by Jørgensen and Bhowmik[51] and defined as,

$$\varepsilon_{\mathrm{mae}} = \frac{\int_{\vec{r}\in V} |\boldsymbol{n}_{\mathrm{DFT}}(\boldsymbol{r}) - \boldsymbol{n}_{\mathrm{NN}}(\boldsymbol{r})|}{\int_{\vec{r}\in V} |\boldsymbol{n}_{\mathrm{DFT}}(\boldsymbol{r})|} \tag{7}$$

where $\boldsymbol{n}_{\mathrm{DFT}}(\boldsymbol{r})$ represents the DFT electron density, and $\boldsymbol{n}_{\mathrm{NN}}(\boldsymbol{r})$ is the FIREANN predicted electron density at the grid. As seen in Figure 3, large prediction errors primarily appear around atoms or chemical bonds, i.e., in the regions with large charge densities. This indicates that sampling less data points in low-density regions does not reduce the prediction ability of the FIREANN model there. Additionally, larger molecules have generally smaller normalized MAEs, as they possess a higher average charge density.

In Figure 4, the accuracy of the FIREANN model is further investigated through a scan along a central line across the propyne and phenol molecules. The variation of charge density at the grid along this line is essentially indistinguishable by DFT and FIREANN. The maximum difference between them never exceeds 0.02 e·Å^{-3}, which is ∼1% of the maximum density. It should be noted that the ML model here only learns the density of valence electrons as core electrons are described by pseudopotentials. In some circumstances, for example, the aggressive pseudopotential removes some charges from carbon atomic centers, corresponding to some local minima in the charge density variation in Figure 4. The total electron density can, in principle, be recovered by adding the core charge to the ML-predicted valence charge density.

To assess the generalization capability of the model beyond the QM9 data set, we have predicted the charge density of molecules in the independent test data set. Moreover, previous studies have highlighted that the performance of density prediction can vary significantly depending on the chosen evaluation metric.[92,93] To provide a comprehensive assessment of the FIREANN-density model's prediction accuracy, we calculated three distinct error metrics for the independent test data set: (1) RMSE of pointwise charge density, (2) normalized MAE as defined in eq 6, and (3) infinity norm of the difference in dipole moment.[93] Figure S1 shows that the RMSE and MAE of the FIREANN-density model on this independent test data set are comparable to those of the QM9 data set, demonstrating the high generalization capability of our model. It also shows that the RMSE and MAE distributions of the model are comparable to the charge density difference distributions obtained using different DFT functionals, namely, PBE and PBE0. However, the model error, i.e., $n_{\mathrm{PBE}} - n_{\mathrm{NN}}$, is larger than that between the difference between two functionals, i.e., $n_{\mathrm{PBE}} - n_{\mathrm{PBE0}}$, when using the infinity norm of the dipole moment as the metric. This discrepancy is likely because we treat charge density in each grid point independently in the loss function, and reproducing dipole moment from the charge density distribution is not our purpose for developing this model. One can, however, easily learn dipole moment directly by the FIREANN model, as discussed in our previous work.[74]

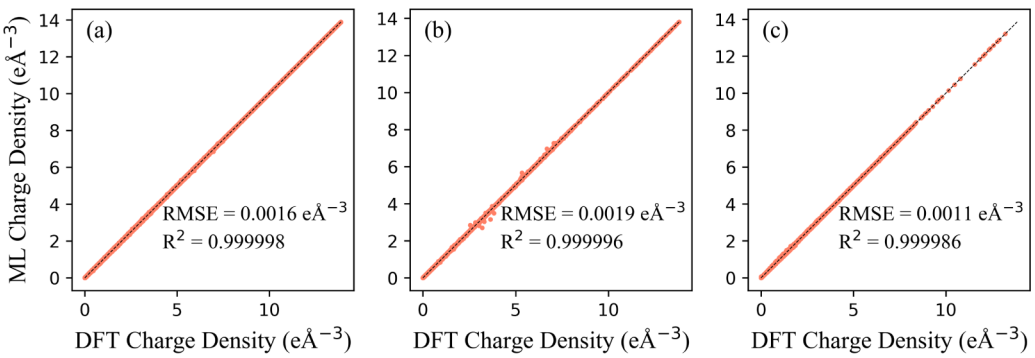


Figure 2. Correlation between DFT and ML charge densities in the training set (a), validation set (b), and test set (c) as part of the QM9 data set. The black dashed lines represent a perfect linear correlation.

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

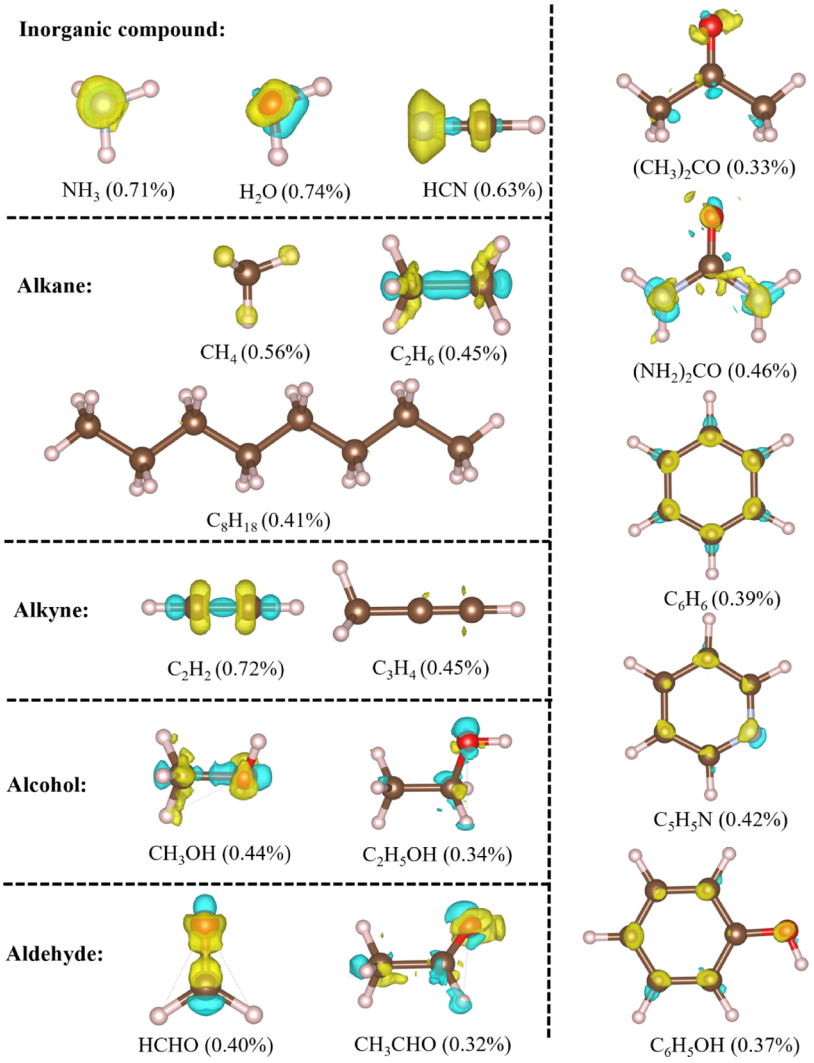


Figure 3. Prediction error iso-surfaces ($\pm$0.001 e·Å^{-3}) for several exemplary molecules in QM9 data set containing H (white), C (brown), O (red), and N (gray) atoms. Numbers in parentheses denote the normalized MAE values for each molecule, as defined in eq 6.

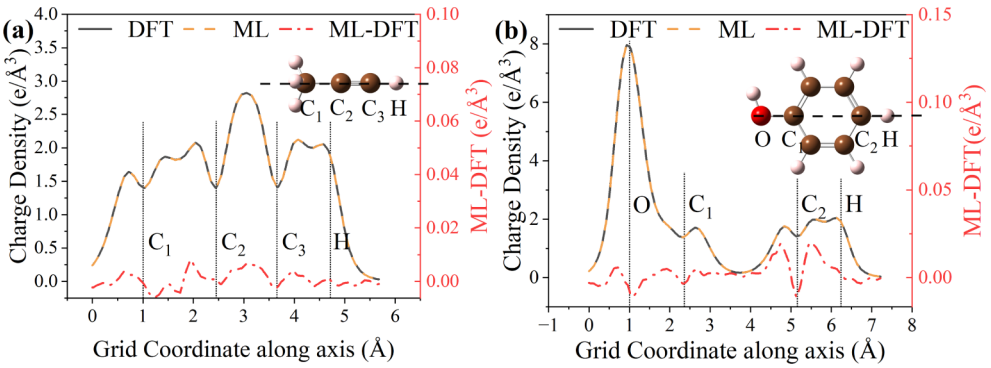


Figure 4. DFT and ML-predicted charge density for propyne and phenol molecules in the test set computed along the line indicated in the inset; the prediction error is also provided on the right-hand side scale (red).

3.1.2. H_2O/Pt(111) Data Set. Next, we turn our attention to the liquid−solid H_2O/Pt(111) interfacial system. Compared to the QM9 data set, this system contains many more electrons and a periodic electron density distribution. Additionally, the H_2O/Pt(111) data set involves more complex chemical interactions, including the metallic bonds within the platinum slab and hydrogen bonds between water molecules. Using a similar procedure as for the QM9 data set, the TS + LIFS strategy selected only 200 grid points out of the raw ∼500 million grid points for each configuration generated by VASP, yielding a data set with 90,000 grid points and representing ∼0.004% of the entire mesh. The RMSEs for the training set and validation set are 0.0031 and 0.0036 e Å^{-3}, respectively, corresponding to ∼1% of the average density value. Similar to

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

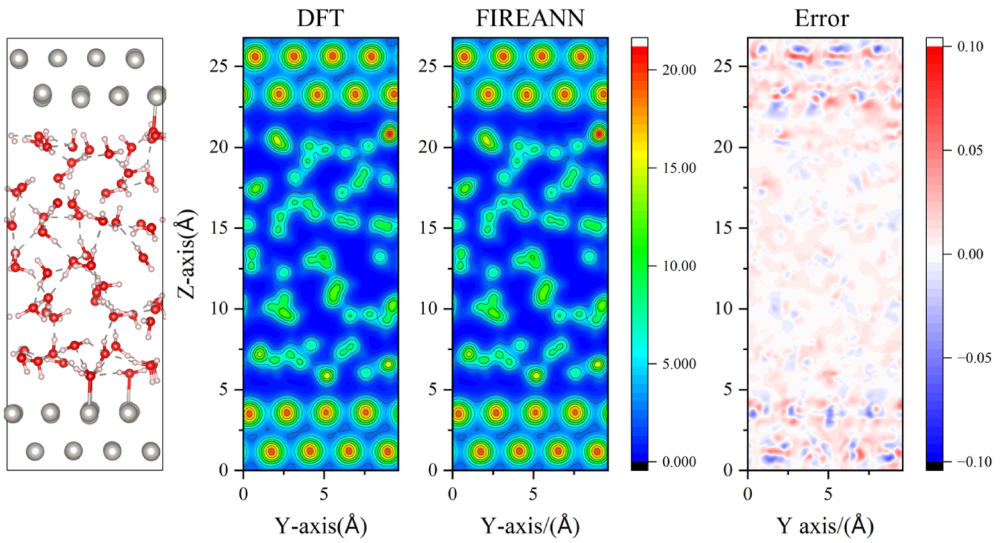


Figure 5. Comparison of electron density of an exemplary configuration in the test set of H_2O/Pt(111) obtained from DFT and FIREANN. From left to right: (a) a snapshot of the H_2O/Pt(111) trajectory; contour plots of DFT calculated (b) and FIREANN predicted (c) electron density integrated along the *X*-axis; (d) the difference between (c) and (d). The units of the color bar are e·Å^{-2}.

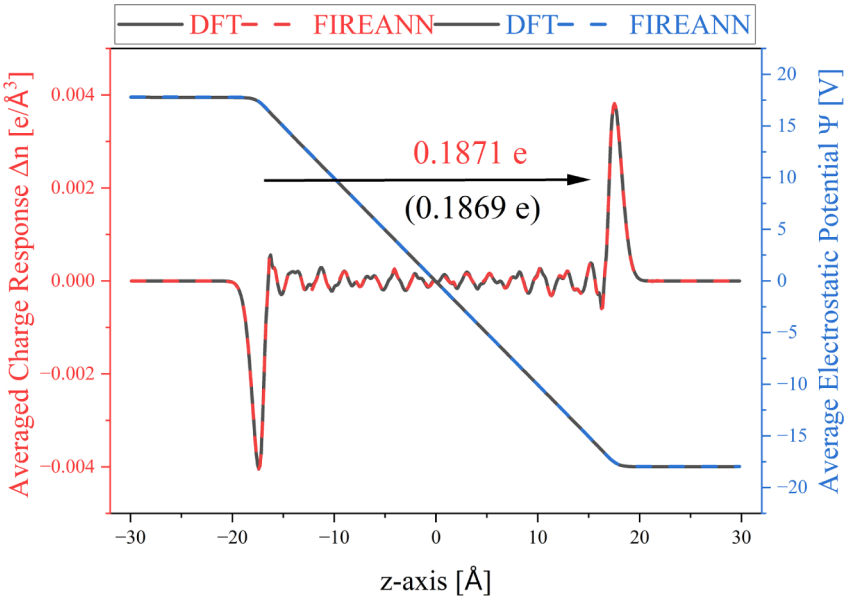


Figure 6. Response of charge density and electrostatic potential averaged over the *xy* plane of an Au (100) electrode in the test set under an external electric field of E_z = −1 V/Å.

the results of QM9, the RMSE for the test set (0.0026 e Å^{-3}) is slightly smaller than that in both the training and validation sets, as low-density data have a relatively larger proportion in the former. Figure 5 shows a perfect agreement between the DFT and FIREANN predicted charge density distributions of a representative configuration of H_2O/Pt(111) in the test set, where the maximum prediction error is less than 0.1 e·Å^{-2}, ∼0.5% of the maximum integrated density along the *X*-axis.

3.2. Predicting the Electron Density Response. To validate the capability of the TS + LIFS strategy and the FIREANN model in predicting the electron density response, we trained the MLEDR model for an Au (100) electrode. It is worth noting that the electron density response of a conducting system to an external field exhibits remarkable nonlocal behavior, as exemplified by the Au electrode system, leading to an accumulation of opposite charges on the two sides of the metal electrode. This renders the use of six message-passing iterations in the training process, as listed in Table 1.

Figure 6 compares the trained FIREANN model for the electron density response of the electrode with the corresponding DFT profile and the corresponding electrostatic potential (ESP) as a function of the vertical position of the Au(100) slab. The charge transfer between the two sides of the metal electrode is obtained by integrating the charge response from the central layer of the slab to the left or right vacuum regions. Our FIREANN model precisely reproduces the charge transfer of 0.1871*e* between the two sides of the electrode, in agreement with the DFT value of 0.1869*e*. The variation of the ESP is computed directly from the predicted charge density response. We find that the slope of the averaged ESP is approximately −1.0 V/Å inside the metal electrode, corre-

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

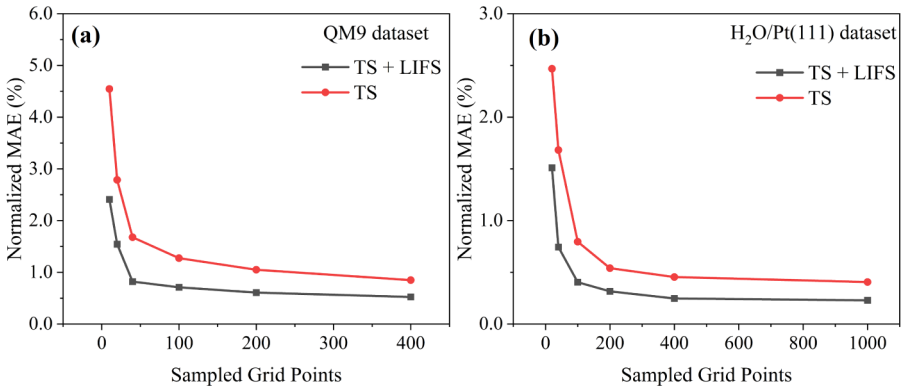


Figure 7. Learning curves of the FIREANN model for electron density in the QM9 (a) and $H_2O/Pt(111)$ data sets (b) using different grid point sampling strategies, i.e., the normalized MAE as a function of the average number of grid points sampled for each molecule (or configuration).

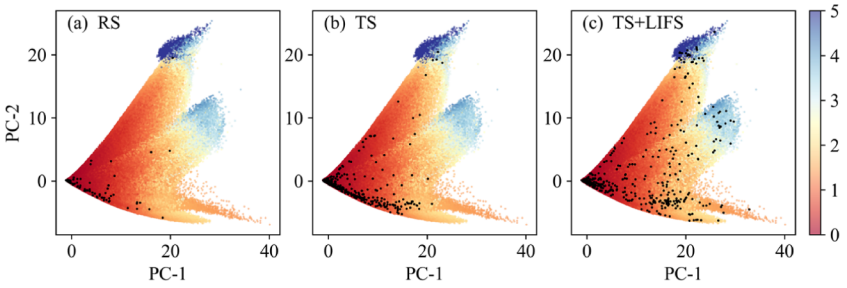


Figure 8. Comparison of the distribution of data points sampled by the RS (a), TS (b), and TS + LIFS (c) strategies in the reduced feature space for the QM9 test set (1000 randomly chosen molecules). Dimension reduction of the EAD features is performed by the principal component analysis (PCA) algorithm. For visual clarity, the electron density value is truncated between 0 and 5 e $Å^{-3}$, and only a random subset (1%) of the grid points is depicted.

sponding exactly to the field strength. This indicates that the internal electric field generated by the response perfectly screens the opposing external field. This result is very similar to that predicted by the SA-GPR model plus a long-distance equivariant descriptor for the same system.[59]

The charge response of a metal slab under an external field is relatively straightforward to capture. To further evaluate the model's capability in describing the long-range charge response of metals under external fields, we applied it to a simple Au nanoparticle data set. The validation result is presented in Figure S2. These results demonstrate that, even for nanoparticle systems with more complex surface structures, the nonlocal charge response under electric fields can still be accurately predicted by our FIREANN-response model. We note that in any extreme cases where the charge response distance is essentially infinite, our model can always serve as a fundamental model, which can incorporate long-range interactions by incorporating global descriptors, such as the long-distance equivariant (LODE) representations in the SALTED model for charge density response.[59,94]

3.3. Comparison of Sampling Strategies. We further examined to what extent the LIFS strategy enhances the data efficiency. In Figure 7, we compare the learning curves (the normalized MAE versus the size of the training set) derived from the TS and TS + LIFS strategies. The learning curve for the RS strategy was not shown here, as previous work has already proven that the RS strategy exhibits much worse efficiency and accuracy.[60] For the QM9 data set, the normalized MAE is again evaluated over the entire grid for outputting $\mathbf{n}(\mathbf{r})$ of randomly chosen 1,000 molecules. As shown in Figure 7a, the MAE decreases quickly with more grid points added in the first few data selection cycles. Impressively, the same level of accuracy is achieved by the TS + LIFS strategy using less than 40 grid points from each molecule, compared with that of the TS strategy alone using 400 points. This result highlights the superior data efficiency of the former method. A similar numerical experiment for the $H_2O/Pt(111)$ data set shown in Figure 7b also suggests that the MAE of the TS + LIFS strategy is always lower than that of the TS strategy for a given number of grid points. The former quickly converges with 100 grid points, while the latter slowly decreases with 200 points. In this case, the TS + LIFS strategy selecting 40 points can reach an MAE comparable to that by the TS strategy alone selecting 1,000 points.

To further understand the effectiveness of the LIFS strategy, we visualize the distribution of grid points sampled by different strategies in the feature space in Figure 8. Specifically, using the test set of QM9 as an example, 30,000 representative grid points were sampled using the RS, TS, and TS + LIFS strategies, respectively. The high-dimensional feature space is reduced by principal component analysis (PCA), and the two most important components are chosen for the visualization. As shown in Figure 8a, the RS strategy tends to sample more points located in the low-density region, as the probability of sampling each point in the real space is equal, and the low-density region is more broadly distributed. Consequently, it is necessary to collect an overwhelming number of points when applying the RS strategy in order to cover the chemical space more evenly and ensure the model accuracy, typically 10^3 to 10^4 grid points per molecule. In comparison, the large-density region is partially covered by the TS strategy, as shown in the upper-right blue region of panel (b), since it offers a higher

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

probability of sampling grid points with large electron densities. One step further, the TS + LIFS strategy allows us to sample grid points with linearly independent features only or, in other words, remove those with similar local environments, thus yielding a much more unbiased distribution of the sampled grid points in the feature space. Therefore, the data set size is significantly reduced to train a balanced MLED model using this sampling scheme.

3.4. Atomic Charge Predictions. A well-trained MLED model is supposed to be used to determine atomic charges, which are often useful to analyze electron transfer processes in chemical reactions. One common charge partitioning scheme is the so-called Bader charge analysis, which attributes partial charges to different atoms separated by zero-flux surfaces of electron density.[95−98] To validate our MLED model, Bader charge analysis has been performed with DFT-based and our FIREANN-based electron density distributions for a series of H_2O/Pt(111) configurations involving a proton transfer from one water molecule to another, which are used as the test set.

Figure 9 compares the Bader charges of the hydrogen atom moving from the donor to the acceptor water molecule. For

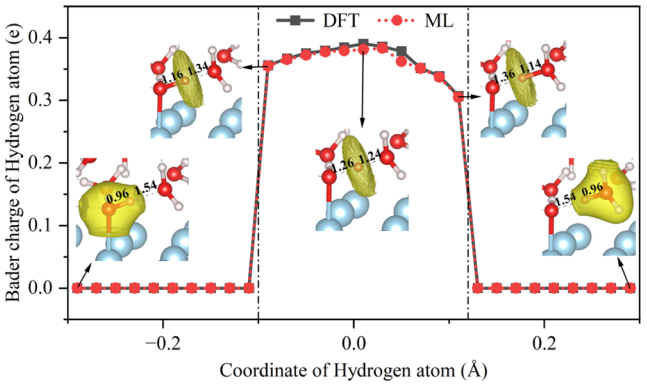


Figure 9. Variation of Bader charges of the hydrogen atom during a proton transfer process in the H_2O/Pt(111) system. Bader charge analysis is performed using charge densities obtained from DFT (black) and the FIREANN model (red). Five representative structures in the proton transfer process are shown with H (white), O (red), and Pt (blue) atoms. The two most relevant O−H distances in these configurations are marked in Å. The horizontal axis represents the displacement of the hydrogen atom relative to the midpoint between the two oxygen atoms.

simplicity, the distance between the oxygen atoms in the two relevant water molecules (O1−O2) is fixed at 2.50 Å, where the shorter O_1−H bond in the donor is initially ∼0.96 Å, and the longer hydrogen bond length (O2−H) is ∼1.54 Å. Interestingly, the Bader charge of the transferring hydrogen atom is constantly zero as the O1−H bond moderately elongates. This is because the electron density of the donor molecule is mostly assigned to the oxygen atom given its high electronegativity, making the hydrogen atom a virtually like a proton. However, as the O1−H distance exceeds ∼1.15 Å, roughly the sum of the covalent radii of isolated oxygen and hydrogen atoms, a zero-flux surface of the electron density appears between O1 and H. Consequently, a partial charge starts to be assigned to the hydrogen atom (0.35e), which slightly increases to a maximum (∼0.39e) when the hydrogen atom lies in the middle of the two oxygen atoms and then slowly decreases as the hydrogen atom further moves to the acceptor water molecule. Symmetrically, when the O2−H distance becomes shorter than ∼1.15 Å, the zero-flux surface disappears suddenly, so that the electron density of the acceptor is now assigned to O2, making the atomic charge of the transferred H atom zero again at the end of this proton transfer process. Since the zero-flux surface is sensitive to the actual electron density distribution, this result clearly demonstrates that the FIREANN model precisely captures the variation of $\mathbf{n}(\mathbf{r})$ upon the breaking and forming of the hydrogen bond.

Note that the integration of the charge density over the space predicted by the ML model differs from the number of valence electrons by less than 0.1%, although the charge balance is not enforced in our FIREANN model. This slight charge imbalance has little impact on the Bader charge analysis results. However, in some physical models of periodic systems, it is problematic if the total charge is not exactly balanced.[58] In such cases, the charge balance can be reinforced by scaling the predicted charge density.

4. CONCLUSIONS

In this work, we propose a very efficient hybrid sampling of grid points to train machine learning electron density and response models in real space. This strategy starts from a value-based targeted sampling, which favors grid points in the high-density region, followed by removing these points associated with linearly dependent atomic features that are regarded as having similar local chemical environments. Combining the proposed strategy and our field-induced recursively embedded atom neural network method allows us to obtain an accurate electron density or response representation using merely about 0.005−0.015% of the entire grid of charge density generated by density functional theory codes, in both the QM9 molecular data set and a periodic H_2O/Pt(111) interfacial system. Moreover, our model is also able to accurately predict the nonlocal charge transfer in an Au(100) electrode under an applied electric field. The resulting machine learning electron density model was used to perform Bader charge analysis and investigate the electron transfer involved in the proton transfer process of H_2O/Pt(111). This study suggests that the linearly independent feature selection is a general way to efficiently select not only atomic features but also data points to train machine learning models. Combined with machine-learned potentials, the proposed machine learning electron density model may be applied, for example, to electronic friction-based nonadiabatic molecular dynamics simulations on metal surfaces,[99] to predict the charge density response in electrochemical cells and perform constant-potential simulations of electrochemical systems with DFT accuracy.[59] Further applications in these scenarios are underway in our group.

■ ASSOCIATED CONTENT

Data Availability Statement

The original FIREANN package is openly available in the GitHub repository at https://github.com/bjiangch/FIREANN. The FIREANN model for the electron density and response module has been uploaded as a branch of the FIREANN package at https://github.com/bjiangch/FIREANN/tree/FIREANN-for-Density-and-Response. Additionally, the VASP (CP2K) input parameters necessary for generating the charge density (response) for the H_2O/Pt(111) data set and the independent test data set (Au electrode data set and Au nanoparticles data set) are provided. The charge densities of QM9 data set are taken from previous work,

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702

available at QM9 Charge Densities and Energies Calculated with VASP (dtu.dk).

Supporting Information

The supporting information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c01355.

Additional information for the details of five data sets including the data set size, element composition, and data division scheme; charge density difference across different metrics; charge response of an Au nanoparticle under an external electric field (PDF)

AUTHOR INFORMATION

Corresponding Author

Bin Jiang − *Key Laboratory of Precision and Intelligent Chemistry, Department of Chemical Physics, University of Science and Technology of China, Hefei, Anhui 230026, China; Hefei National Laboratory, University of Science and Technology of China, Hefei 230088, China;* orcid.org/0000-0003-2696-5436; Email: bjiangch@ustc.edu.cn

Authors

Chaoqiang Feng − *Hefei National Research Center for Physical Sciences at the Microscale, Department of Chemical Physics, University of Science and Technology of China, Hefei, Anhui 230026, China*

Yaolong Zhang − *Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States;* orcid.org/0000-0002-4601-0461

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c01355

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0450101), the Innovation Program for Quantum Science and Technology (2021ZD0303301), the CAS Project for Young Scientists in Basic Research (YSBR-005), and the National Natural Science Foundation of China (22325304, 22221003, and 22033007). We acknowledge the Supercomputing Center of USTC, Hefei Advanced Computing Center, and Beijing PARATERA Tech Co., Ltd. for providing high-performance computing services.

REFERENCES

(1) Veszprémi, T.; Fehér, M. Properties Related to Electron Density. *Quantum Chemistry: Fundamentals to Applications*. 1999, Springer US: Boston, MA pp. 249−257.

(2) Ballesteros, F.; Lao, K. U. Accelerating the Convergence of Self-Consistent Field Calculations Using the Many-Body Expansion. *J. Chem. Theory Comput.* **2022**, *18* (1), 179−191.

(3) Kang, P.-L.; Shang, C.; Liu, Z.-P. Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration. *Acc. Chem. Res.* **2020**, *53* (10), 2119−2129.

(4) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11* (6), 2336−2347.

(5) Wang, X.; Ye, S.; Hu, W.; Sharman, E.; Liu, R.; Liu, Y.; Luo, Y.; Jiang, J. Electric Dipole Descriptor for Machine Learning Prediction of Catalyst Surface−Molecular Adsorbate Interactions. *J. Am. Chem. Soc.* **2020**, *142* (17), 7737−7743.

(6) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71* (1), 361−390.

(7) Wu, A.; Ye, Q.; Zhuang, X.; Chen, Q.; Zhang, J.; Wu, J.; Xu, X. Elucidating Structures of Complex Organic Compounds Using a Machine Learning Model Based on the 13C NMR Chemical Shifts. *Precis. Chem.* **2023**, *1* (1), 57−68.

(8) Luo, Y. Chemistry in the Era of Artificial Intelligence. *Precis. Chem.* **2023**, *1* (2), 127−128.

(9) Li, W.; Wang, G.; Ma, J. Deep learning for complex chemical systems. *Natl. Sci. Rev.* **2023**, *10* (12), nwad335.

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

(11) 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.

(12) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting potential energy surfaces with fundamental invariant neural network. *J. Chem. Phys.* **2016**, *145* (7), 071101.

(13) Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14* (3), 1153−1173.

(14) Qu, C.; Yu, Q.; Bowman, J. M. Permutationally invariant potential energy surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69* (1), 151−175.

(15) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120* (14), 143001.

(16) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9* (1), 3887.

(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.

(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.

(19) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **2019**, *99* (1), 014104.

(20) Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant Molecular Neural Networks. In *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp 14537−14546.

(21) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152* (4), 044107.

(22) Park, C. W.; Kornbluth, M.; Vandermause, J.; Wolverton, C.; Kozinsky, B.; Mailoa, J. P. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *Npj Comput. Mater.* **2021**, *7* (1), 73.

(23) Wang, X.; Xu, Y.; Zheng, H.; Yu, K. A Scalable Graph Neural Network Method for Developing an Accurate Force Field of Large Flexible Organic Molecules. *J. Phys. Chem. Lett.* **2021**, *12* (33), 7982−7987.

(24) Guan, Y.; Yarkony, D. R.; Zhang, D. H. Permutation invariant polynomial neural network based diabatic ansatz for the (E + A) × (e + a) Jahn−Teller and Pseudo-Jahn−Teller systems. *J. Chem. Phys.* **2022**, *157* (1), 014110.

(25) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13* (1), 2453.

(26) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc, 2022, Vol. *35*, pp. 11423−11436.

(27) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning local equivariant representa-

tions for large-scale atomistic dynamics. *Nat. Commun.* **2023**, *14* (1), 579.
(28) Fu, B.; Zhang, D. H. Accurate fundamental invariant-neural network representation of ab initio potential energy surfaces. *Natl. Sci. Rev.* **2023**, *10* (12), nwad321.
(29) Kang, P.-L.; Yang, Z.-X.; Shang, C.; Liu, Z.-P. Global Neural Network Potential with Explicit Many-Body Functions for Improved Descriptions of Complex Potential Energy Surface. *J. Chem. Theory Comput.* **2023**, *19* (21), 7972−7981.
(30) Xie, X.-T.; Yang, Z.-X.; Chen, D.; Shi, Y.-F.; Kang, P.-L.; Ma, S.; Li, Y.-F.; Shang, C.; Liu, Z.-P. LASP to the Future of Atomic Simulation: Intelligence and Automation. *Precis. Chem.* **2024**, *2* (12), 612−627.
(31) Cheng, X.; Wu, C.; Xu, J.; Han, Y.; Xie, W.; Hu, P. Leveraging Machine Learning Potentials for In-Situ Searching of Active sites in Heterogeneous Catalysis. *Precis. Chem.* **2024**, *2* (11), 570−586.
(32) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121* (16), 10037−10072.
(33) Zhang, Y.; Lin, Q.; Jiang, B. Atomistic neural network representations for chemical dynamics simulations of molecular, condensed phase, and interfacial systems: Efficiency, representability, and generalization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2023**, *13* (3), No. e1645.
(34) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017**, *8* (1), 872.
(35) Alred, J. M.; Bets, K. V.; Xie, Y.; Yakobson, B. I. Machine learning electron density in sulfur crosslinked carbon nanotubes. *Compos. Sci. Technol.* **2018**, *166*, 3−9.
(36) Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **2019**, *10* (41), 9424−9432.
(37) Mills, K.; Ryczko, K.; Luchak, I.; Domurad, A.; Beeler, C.; Tamblyn, I. Extensive deep neural networks for transferring small scale learning to large scale systems. *Chem. Sci.* **2019**, *10* (15), 4129−4140.
(38) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **2019**, *5* (1), 57−64.
(39) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *Npj Comput. Mater.* **2019**, *5* (1), 22.
(40) Gong, S.; Xie, T.; Zhu, T.; Wang, S.; Fadel, E. R.; Li, Y.; Grossman, J. C. Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Phys. Rev. B* **2019**, *100* (18), 184103.
(41) Kamal, D.; Chandrasekaran, A.; Batra, R.; Ramprasad, R. A charge density prediction model for hydrocarbons using deep neural networks. *Mach. Learn.: Sci. Technol.* **2020**, *1* (2), 025003.
(42) Cuevas-Zuviría, B.; Pacios, L. F. Analytical model of electron density and its machine learning inference. *J. Chem. Inf. Model.* **2020**, *60* (8), 3831−3842.
(43) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11* (1), 5223.
(44) Tsubaki, M.; Mizoguchi, T. Quantum deep field: data-driven wave function, electron density generation, and atomization energy prediction and extrapolation with machine learning. *Phys. Rev. Lett.* **2020**, *125* (20), 206401.
(45) Del Rio, B. G.; Kuenneth, C.; Tran, H. D.; Ramprasad, R. An efficient deep learning scheme to predict the electronic structure of materials and molecules: The example of graphene-derived allotropes. *J. Phys. Chem. A* **2020**, *124* (45), 9496−9502.
(46) Cuevas-Zuviría, B.; Pacios, L. F. Machine learning of analytical electron density in large molecules through message-passing. *J. Chem. Inf. Model.* **2021**, *61* (6), 2658−2666.
(47) Ellis, J. A.; Fiedler, L.; Popoola, G. A.; Modine, N. A.; Stephens, J. A.; Thompson, A. P.; Cangi, A.; Rajamanickam, S. Accelerating finite-temperature Kohn-Sham density functional theory with deep neural networks. *Phys. Rev. B* **2021**, *104* (3), 035120.
(48) Zepeda-Núñez, L.; Chen, Y.; Zhang, J.; Jia, W.; Zhang, L.; Lin, L. Deep Density: circumventing the Kohn-Sham equations via symmetry preserving neural networks. *J. Comput. Phys.* **2021**, *443*, 110523.
(49) Lewis, A. M.; Grisafi, A.; Ceriotti, M.; Rossi, M. Learning electron densities in the condensed phase. *J. Chem. Theory Comput.* **2021**, *17* (11), 7203−7214.
(50) Qiao, Z.; Christensen, A. S.; Welborn, M.; Manby, F. R.; Anandkumar, A.; Miller Iii, T. F. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (31), No. e2205221119.
(51) Jørgensen, P. B.; Bhowmik, A. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *Npj Comput. Mater.* **2022**, *8* (1), 183.
(52) Lee, A. J.; Rackers, J. A.; Bricker, W. P. Predicting accurate ab initio DNA electron densities with equivariant neural networks. *Biophys. J.* **2022**, *121* (20), 3883−3895.
(53) Rackers, J. A.; Tecot, L.; Geiger, M.; Smidt, T. E. A recipe for cracking the quantum scaling limit with machine learned electron densities. *Mach. Learn.: sci. Technol.* **2023**, *4* (1), 015027.
(54) Achar, S. K.; Bernasconi, L.; Johnson, J. K. Machine Learning Electron Density Prediction Using Weighted Smooth Overlap of Atomic Positions. *Nanomaterials* **2023**, *13* (12), 1853.
(55) Fiedler, L.; Modine, N. A.; Schmerler, S.; Vogel, D. J.; Popoola, G. A.; Thompson, A. P.; Rajamanickam, S.; Cangi, A. Predicting electronic structures at any length scale with machine learning. *Npj Comput. Mater.* **2023**, *9* (1), 115.
(56) Grisafi, A.; Lewis, A. M.; Rossi, M.; Ceriotti, M. Electronic-structure properties from atom-centered predictions of the electron density. *J. Chem. Theory Comput.* **2023**, *19* (14), 4451−4460.
(57) Del Rio, B. G.; Phan, B.; Ramprasad, R. A deep learning framework to emulate density functional theory. *Npj Comput. Mater.* **2023**, *9* (1), 158.
(58) Sunshine, E. M.; Shuaibi, M.; Ulissi, Z. W.; Kitchin, J. R. Chemical Properties from Graph Neural Network-Predicted Electron Densities. *J. Phys. Chem. C* **2023**, *127* (48), 23459−23466.
(59) Grisafi, A.; Bussy, A.; Salanne, M.; Vuilleumier, R. Predicting the charge density response in metal electrodes. *Phys. Rev. Mater.* **2023**, *7* (12), 125403.
(60) Focassio, B.; Domina, M.; Patil, U.; Fazzio, A.; Sanvito, S. Linear Jacobi-Legendre expansion of the charge density for machine learning-accelerated electronic structure calculations. *Npj Comput. Mater.* **2023**, *9* (1), 87.
(61) Koker, T.; Quigley, K.; Taw, E.; Tibbetts, K.; Li, L. Higher-order equivariant neural networks for charge density prediction in materials. *Npj Comput. Mater.* **2024**, *10* (1), 161.
(62) Margraf, J. T. Science-Driven Atomistic Machine Learning. *Angew. Chem., Int. Ed.* **2023**, *135* (26), No. e202219170.
(63) Bai, Y.; Vogt-Maranto, L.; Tuckerman, M. E.; Glover, W. J. Machine learning the Hohenberg-Kohn map for molecular excited states. *Nat. Commun.* **2022**, *13* (1), 7044.
(64) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Phys. Rev. Lett.* **2018**, *120* (3), 036002.
(65) Hafner, J. Ab-initio simulations of materials using VASP: Density-functional theory and beyond. *J. Comput. Chem.* **2008**, *29* (13), 2044−2078.
(66) Kühne, T. D.; Iannuzzi, M.; Ben, M. D.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; et al. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152* (19), 194103.
(67) Yin, R.; Zhang, Y.; Jiang, B. Strong Vibrational Relaxation of NO Scattered from Au(111): Importance of the Adiabatic Potential Energy Surface. *J. Phys. Chem. Lett.* **2019**, *10* (19), 5969−5974.
(68) Zhou, X.; Zhang, Y.; Yin, R.; Hu, C.; Jiang, B. Neural Network Representations for Studying Gas-Surface Reaction Dynamics:

Beyond the Born-Oppenheimer Static Surface Approximation†. *Chin. J. Chem.* **2021**, *39* (10), 2917−2930.
(69) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **2021**, *11* (10), 6059−6072.
(70) Lewis, A. M.; Lazzaroni, P.; Rossi, M. Predicting the electronic density response of condensed-phase systems to electric field perturbations. *J. Chem. Phys.* **2023**, *159* (1), 014103.
(71) Zhang, Y.; Hu, C.; Jiang, B. Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation. *J. Phys. Chem. Lett.* **2019**, *10* (17), 4962−4967.
(72) Zhang, Y.; Xia, J.; Jiang, B. Physically motivated recursively embedded atom neural networks: Incorporating local completeness and nonlocality. *Phys. Rev. Lett.* **2021**, *127* (15), 156002.
(73) Zhang, Y.; Xia, J.; Jiang, B. REANN: A PyTorch-based end-to-end multi-functional deep neural network package for molecular, reactive, and periodic systems. *J. Chem. Phys.* **2022**, *156* (11), 114801.
(74) Zhang, Y.; Jiang, B. Universal machine learning for the response of atomistic systems to external fields. *Nat. Commun.* **2023**, *14* (1), 6424.
(75) Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys* **2017**, *146* (24), 244108.
(76) Xia, J.; Zhang, Y.; Jiang, B. Efficient Selection of Linearly Independent Atomic Features for Accurate Machine Learning Potentials. *Chin. J. Chem. Phys.* **2021**, *34* (6), 695−703.
(77) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.
(78) 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.
(79) Jørgensen, P. B.; Bhowmik, A. *QM9 Charge Densities and Energies Calculated with VASP*; Technical University of Denmark, 2022.
(80) Kresse, G.; Furthmuller, J. Efficient iterative schemes for ab initio total-energy calculations using plane wave basis set. *Phys. Rev. B* **1996**, *54*, 11169−11186.
(81) Kresse, G.; Furthmuller, J. Efficiency of ab initio total energy calculations for metals and semiconductors using plane wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(82) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(83) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953−17979.
(84) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110* (13), 6158−6170.
(85) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(86) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comput. Phys. Commun.* **2005**, *167* (2), 103−128.
(87) Goedecker, S.; Teter, M.; Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* **1996**, *54* (3), 1703−1710.
(88) Hartwigsen, C.; Goedecker, S.; Hutter, J. Relativistic separable dual-space Gaussian Pseudopotentials from H to Rn. *Phys. Rev. B* **1998**, *58* (7), 3641−3662.
(89) VandeVondele, J.; Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.* **2007**, *127* (11), 114105.
(90) Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13* (14), 6670−6688.
(91) Loshchilov, I.; Hutter, F. *Decoupled Weight Decay Regularization*; International Conference on Learning Representations, 2017.
(92) Vuckovic, S.; Song, S.; Kozlowski, J.; Sim, E.; Burke, K. Density Functional Analysis: The Theory of Density-Corrected DFT. *J. Chem. Theory Comput.* **2019**, *15* (12), 6636−6646.
(93) Gubler, M.; Schäfer, M. R.; Behler, J.; Goedecker, S. Accuracy of Charge Densities in Electronic Structure Calculations 2024. *arXiv*, 2024.
(94) Grisafi, A.; Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **2019**, *151* (20), 204105.
(95) Henkelman, G.; Arnaldsson, A.; Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Comput. Mater. Sci.* **2006**, *36* (3), 354−360.
(96) Sanville, E.; Kenny, S. D.; Smith, R.; Henkelman, G. Improved grid-based algorithm for Bader charge allocation. *J. Comput. Chem.* **2007**, *28* (5), 899−908.
(97) Tang, W.; Sanville, E.; Henkelman, G. A grid-based Bader analysis algorithm without lattice bias. *J. Phys.: condens. Matter* **2009**, *21* (8), 084204.
(98) Yu, M.; Trinkle, D. R. Accurate and efficient algorithm for Bader charge integration. *J. Chem. Phys.* **2011**, *134* (6), 064111.
(99) Žugec, I.; Tetenoire, A.; Muzas, A. S.; Zhang, Y.; Jiang, B.; Alducin, M.; Juaristi, J. I. Understanding the Photoinduced Desorption and Oxidation of CO on Ru(0001) Using a Neural Network Potential Energy Surface. *JACS Au* **2024**, *4* (5), 1997−2004.

https://doi.org/10.1021/acs.jctc.4c01355
J. Chem. Theory Comput. 2025, 21, 691−702