

Quantifying Structural and Elemental Contributions to Local Electric Fields on Metal Nanoclusters with Interpretable Machine Learning

Runze Zhao

Worcester Polytechnic Institute

Jiaqi Yang

Worcester Polytechnic Institute

Qiang Li

Worcester Polytechnic Institute

Fanglin Che

fche@wpi.edu

Worcester Polytechnic Institute

Article

Keywords: Local electric field, field-assisted catalysis, metal nanoclusters, interpretable machine learning, SHAP, symbolic regression

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11061082/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantifying Structural and Elemental Contributions to Local Electric Fields on Metal Nanoclusters with Interpretable Machine Learning

Runze Zhao, Jiaqi Yang, Qiang Li, Fanglin Che*

Department of Chemical Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA

* Corresponding author: fche@wpi.edu

Abstract

Local electric fields (LEFs) govern field-assisted catalysis, but distinguishing the contributions of local geometry and metal identity remains challenging because changing the metal inevitably alters the equilibrium structure of the catalyst. Here we combine density functional theory (DFT) calculations with interpretable machine learning to quantify the structural and elemental contributions to the LEF on monometallic nanoclusters spanning nine transition metals, cluster sizes from 20 to 90 atoms, and external electric fields from -0.5 to $+0.5$ V/Å. A matched-site Δ LEF analysis shows that metal-substitution-associated LEF variations are below 0.08 V/Å and are explained predominantly by accompanying changes in local geometry. Consistently, SHAP analysis of the absolute LEF identifies geometric descriptors as the dominant predictors across the full dataset, whereas electronic contributions are secondary and are better captured by higher-order band-shape descriptors than by the conventional d -band center. An XGBoost model predicts the LEF with a test mean absolute error of 0.013 V/Å, and symbolic regression further reduces the relationship to a four-variable analytical equation with a test error of 0.031 V/Å. These results identify local geometry as the primary determinant of the LEF on transition-metal nanoclusters and provide an interpretable framework for predicting local electric fields in field-assisted catalysis.

Keywords

Local electric field; field-assisted catalysis; metal nanoclusters; interpretable machine learning; SHAP; symbolic regression

1. Introduction

Electric fields (EFs) can alter chemical reactivity by interacting with molecular charge distributions and dipole moments, thereby changing adsorption energies, activation barriers, reaction pathways, and selectivity.^{1,2} Such effects have been demonstrated across a wide range of chemical and catalytic systems, including electrostatic catalysis in single-molecule junctions,^{3,4} electrochemical CO₂ reduction,⁵ N₂/NH₃ interconversion,⁶ methane steam reforming,⁷ and enzyme catalysis.⁸ For example, electric fields on the order of 0.6 V/Å can alter the rate-limiting step of ammonia synthesis,⁶ while field-assisted methane steam reforming can lower the required reaction temperature and suppress coke formation.⁷ These studies demonstrate the considerable potential of electric fields for catalyst design.

Among heterogeneous catalysts, transition-metal nanoparticles constitute one of the most important classes of catalytic materials because they provide abundant active surface sites, tunable electronic structures, and catalytic performance across a broad range of industrially relevant reactions. Their catalytic behavior is highly sensitive to both metal identity and nanoparticle morphology, making them an ideal model system for investigating electric-field effects.

For metal catalysts, however, the externally applied electric field (EEF) is not the quantity that directly governs surface chemistry. Instead, reactions respond to the local electric field (LEF) experienced by adsorbates at individual active sites. The LEF arises from the interaction between the applied field and the local catalyst environment, where charge redistribution causes field enhancement at low-coordinated atoms, edges, corners, and sharp tips.^{5,9} For example, finite-element modelling by Guo et al.¹⁰ demonstrates that the local field at a metal-oxide nanograin increases sixfold as its tip angle narrows from 90° to 15°, and the field near a gold nanostructure increases by an order of magnitude when a rounded tip is reshaped into a sharp one.¹¹ Consequently, nanoparticle morphology can amplify the local field by an order of magnitude, allowing relatively modest external fields to produce substantial changes in reaction energetics. Quantitative prediction of the LEF is therefore essential for the rational design of field-assisted catalysts.

Although both catalyst structure and metal identity influence the LEF, their individual contributions remain unresolved because these two factors are intrinsically coupled. Changing the metal inevitably alters the equilibrium structure of the nanoparticle through differences in lattice constants, bonding, and surface energetics. As a result, comparisons between different metals simultaneously change both the electronic structure and the local geometry, making it difficult to determine whether variations in the LEF originate from intrinsic elemental electronic properties or from geometry induced by structural relaxation. This coupling therefore raises a central mechanistic question: does changing the metal modify the LEF primarily through intrinsic electronic differences, or indirectly through the accompanying changes in relaxed local

geometry? Resolving this question has important implications for catalyst design. If metal identity directly controls the LEF, catalyst optimization should primarily focus on composition. Conversely, if geometry dominates, engineering nanoparticle morphology becomes the more effective strategy for tuning local electric fields.

Our previous density functional theory (DFT) study mapped the LEF across Ni nanoparticles¹² and demonstrated that the local field is strongly influenced by the local geometric environment and nanoparticle size.¹³⁻¹⁵ However, because only a single metal was considered, the role of metal identity could not be determined.

To distinguish structural and elemental effects, descriptors are needed that independently characterize the geometric environment and the electronic structure of the catalyst. Geometric environments in heterogeneous catalysis are commonly characterized using coordination-based descriptors such as generalized coordination number (GCN),¹³⁻¹⁵ whereas elemental electronic effects are often represented by *d*-band descriptors, particularly the *d*-band center.^{16,17} More recent studies have shown that higher-order characteristics of the *d*-band, including its width, skewness, and kurtosis, can capture additional information beyond the *d*-band center alone.^{18,19} Whether these geometric and electronic descriptors can isolate structural and elemental contributions to the LEF, however, remains unknown.

Here, we combine DFT calculations with interpretable machine learning to determine how structural geometry and metal-dependent electronic properties contribute to the LEF on transition-metal nanoclusters. We first construct a site-resolved dataset spanning nine transition metals, twelve cluster sizes, and a broad range of externally applied electric fields. We then introduce a matched-site Δ LEF analysis to quantify the LEF variation associated with metal substitution and determine whether this variation is explained primarily by the accompanying structural relaxation or by electronic descriptors. We subsequently test whether the resulting physical hierarchy persists in absolute LEF prediction across the complete dataset using XGBoost models and SHapley Additive exPlanations (SHAP)²⁰ analysis. Finally, symbolic regression²¹ is used to determine whether the learned relationship can be reduced to a compact analytical expression.

2. Results and Discussion

2.1 Dataset, descriptors, and workflow

Figure 1 summarizes the dataset, descriptor framework, and overall workflow of this study. Figure 1a illustrates a representative Ni₉₀ nanocluster under a uniform external electric field (EEF). The EEF is applied along the *z*-direction (black arrows), and the nanocluster responds by generating local electric fields

(LEFs, blue arrows) at its surface. Because the local atomic environment varies from site to site, the LEF is non-uniform across the nanocluster surface, even under a uniform applied field.

This work focuses on small metal nanoclusters because their catalytic surfaces consist of irregular, low-symmetry atomic environments, where each adsorption site experiences a distinct local geometry and therefore a distinct local electric field. Unlike extended surfaces and larger nanoparticles, which contain a greater proportion of well-defined facet-like environments,²² nanoclusters in the 20-90 atom range exhibit a rich variety of under-coordinated sites, edges, corners, and locally distorted environments. Quantifying the LEF across these heterogeneous sites is therefore essential for understanding field effects at the nanocluster scale.

The dataset comprises nine transition metals (Co, Cu, Fe, Ir, Ni, Pd, Pt, Rh, and Ru), each represented by twelve nanoclusters containing 20-90 atoms. Every relaxed cluster was evaluated under eleven externally applied electric fields ranging from -0.5 to +0.5 V/Å in increments of 0.1 V/Å. Only top adsorption sites on the upper hemisphere of each cluster were considered so that the LEF is aligned with the applied field (**Figure 1a**). The dataset is therefore organized at the level of individual adsorption sites rather than whole clusters.

The Ir-aligned modeling dataset is summarized in **Section S1 and Table S1**. Structural-similarity filtering is described in **Section S2, Table S2, and Figure S1**. **Figure 1b** shows the number of unique surface sites retained after filtering. The number of distinct sites generally increases with cluster size, from only a few sites for the 20-atom clusters to more than thirty sites for the 90-atom clusters, demonstrating that the selected size range spans progressively richer local geometric environments while remaining within the nanocluster regime. After base-cluster selection, structural-similarity filtering, descriptor-completeness filtering, and Ir alignment (**Sections S1 and S2**), the final dataset contains 15,851 site-field records.

Each LEF record is represented by three physically distinct descriptor families (**Figure 1c**): global cluster morphology, local site geometry, and elemental electronic structure. Global descriptors characterize the overall morphology of the nanocluster and include quantities such as atom count, cluster diameter, tip height, and surface-to-volume ratio. Local descriptors describe the adsorption-site environment, including the site height above the center of mass, GCN, and distance from the cluster center. Elemental descriptors characterize the electronic structure of each metal and include the *s*-, *p*-, and *d*-band centers, band fillings, band edges, and higher-order band-shape descriptors such as skewness and kurtosis. In addition to the commonly used *d*-band descriptors, we include *s*- and *p*-band descriptors because the electronic structure of nanoclusters smaller than 2 nm differs substantially from that of extended slab surfaces, motivating a broader description of their electronic structure (**definitions in Section S3 and Tables S3 and S4; site-**

resolved *d*-band center trends and representative *s*- and *p*-projected densities of states in Section S4 and Figures S2 and S3).²³

The overall workflow is summarized in **Figure 1d**. DFT calculations first generate the site-resolved LEF dataset across the nine transition metals, twelve cluster sizes, and eleven applied electric fields. The resulting data are then used to train supervised machine-learning models that predict the LEF from the geometric and electronic descriptors. SHapley Additive exPlanations (SHAP)²⁰ are subsequently employed to quantify the LEF variation associated with metal substitution and determine the extent to which this variation is explained by accompanying structural relaxation versus electronic descriptors. Finally, symbolic regression with PySR²¹ reduces the learned relationship to a compact analytical equation for rapid LEF prediction. The model-training and evaluation protocol is detailed in **Section S5**. With this site-resolved dataset and descriptor framework established, we first address the central mechanistic question of whether changing metal identity produces substantial LEF variation and, if so, whether that variation originates predominantly from metal-dependent structural relaxation or electronic structure.

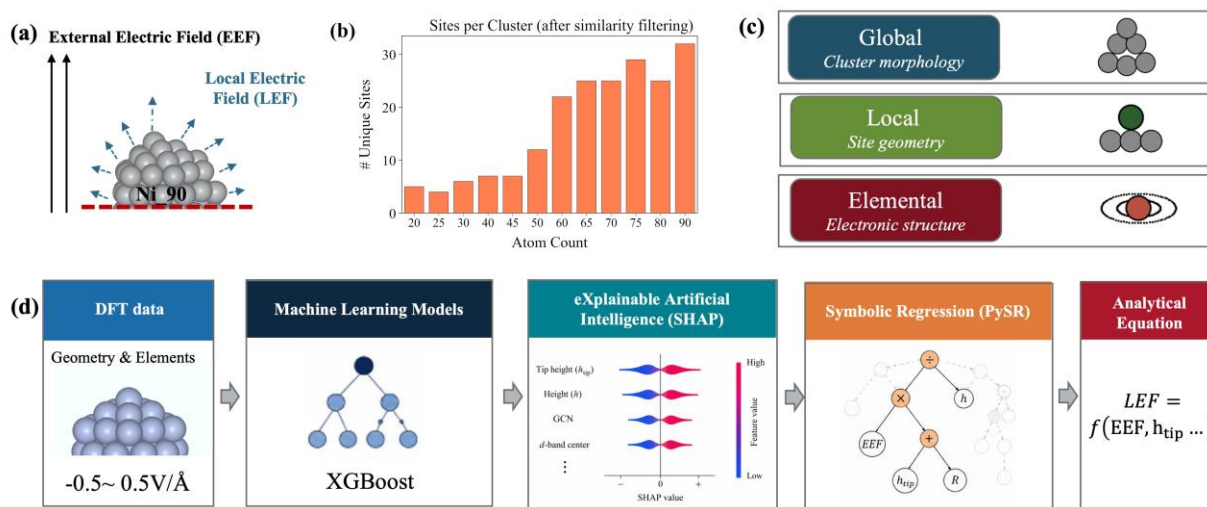


Figure 1. Overview of the dataset, descriptor framework, and machine-learning workflow. (a) Schematic of a representative Ni_{90} nanocluster under a uniform external electric field (EEF, black arrows), illustrating the site-dependent local electric fields (LEFs, blue arrows) generated on the cluster surface. The red dashed line indicates the plane through the center of mass, and only top adsorption sites on the upper hemisphere are included in the dataset. (b) Number of unique surface adsorption sites retained after structural-similarity filtering for nanoclusters containing 20-90 atoms. (c) Three descriptor families used to characterize each adsorption site: global cluster morphology, local site geometry, and elemental electronic structure. (d) Overall workflow. Site-resolved LEFs obtained from DFT calculations are used to train machine-learning models, followed by SHAP analysis for descriptor interpretation and symbolic regression (PySR) to derive a compact analytical equation for LEF prediction.

2.2 Quantifying the LEF variation associated with metal substitution via ΔLEF

To quantify the LEF variation associated with metal substitution, we introduce a ΔLEF formulation referenced to Ir at corresponding adsorption sites. Direct comparison of LEFs across different metals is complicated because changing the metal also changes the relaxed nanocluster geometry. We therefore compare corresponding sites originating from a common structural topology at identical cluster size and EEF. Metal identity is changed, while the resulting metal-dependent structural relaxation is retained and quantified through the geometric descriptors. We define the matched-site LEF difference as:

$$\Delta\text{LEF}(\text{M}) = \text{LEF}(\text{M}) - \text{LEF}(\text{Ir}) \quad (1)$$

where Ir serves as the reference metal. Ir was selected because its DFT calculations converged reliably with a magnetization of 0 μB . The choice of reference changes the numerical values of ΔLEF but does not alter the purpose of the matched-site comparison. By comparing corresponding adsorption sites under identical cluster size and applied-field conditions, the ΔLEF formulation removes variations arising from differences in site identity, particle size, and external field, leaving the LEF variation associated with changing the metal.

The ΔLEF model was constructed using the 90-atom nanoclusters at an applied field of +0.5 V/Å, comprising 180 site samples across the eight non-reference metals. The 90-atom clusters provide the largest number of common adsorption sites among the metals (**Figure 1b**), enabling the broadest comparison of equivalent local environments, while the strongest EEF maximizes the LEF response and improves the resolution of the comparatively small differences associated with metal substitution. For this analysis, each structural descriptor was expressed as its difference relative to Ir at the corresponding site, whereas the elemental descriptors entered unchanged. An XGBoost regressor²⁴ was trained on this subset. Because all thirteen regression models exhibited comparable performance on this smaller dataset (**Section S6 and Figure S4**), XGBoost was selected for consistency with the absolute-LEF analysis presented in **Section 2.3**. The 180 samples were evaluated using an 80:20 random sample-level holdout split under the common cross-validation protocol described in **Section S5**.

Figure 2a compares the predicted and DFT-computed ΔLEF . Both the training and test data follow the parity line, with a train MAE of 0.0070 V/Å and a test MAE of 0.0133 V/Å, respectively. More importantly, The DFT-computed ΔLEF spans only approximately -0.08 to +0.08 V/Å, substantially narrower than the approximately -0.5 to +0.5 V/Å range of absolute LEFs represented in the complete dataset. This comparison indicates that metal substitution contributes a comparatively limited fraction of the overall LEF variation.

To determine what controls the remaining LEF variation associated with metal substitution, we next examine the SHAP attribution of the Δ LEF model (**Figure 2b**). The five highest-ranked descriptors are exclusively geometric: radial distance to the center of mass (R), the mean coordination number of the first-shell neighbors ($\langle CN \rangle_1$), the mean first-neighbor distance ($\bar{r}_1$), the tip height (h_{tip}), and the site height (h). Although corresponding adsorption sites are compared at fixed cluster size and applied field, substituting one metal for another changes the equilibrium bond lengths and local atomic arrangement, causing each cluster to relax into a slightly different geometry. These element-induced structural relaxations are therefore captured primarily by geometric descriptors, which dominate the Δ LEF prediction.

Electronic descriptors first appear below the dominant geometric descriptors and are represented mainly by d -band shape quantities, including the d -band skewness (S_d), lower and upper band edges (ϵ_d^l and ϵ_d^u), and down-spin kurtosis ($K_d^\downarrow$). One local geometric descriptor, the standard deviation of the first-neighbor distances (σ_{r_1}), also appears among these lower-ranked features. Notably, the d -band center, which is widely used as a single-parameter electronic descriptor in heterogeneous catalysis, does not appear among the ten most influential descriptors. The same qualitative descriptor ranking is obtained for the 60-atom clusters (**Figure S5**), where the radial distance from the center of mass again ranks first and electronic descriptors appear only below the leading geometric descriptors. This agreement indicates that the geometry-dominated elemental effect is not unique to the 90-atom clusters.

These results indicate that metal identity affects the LEF predominantly indirectly, by modifying the relaxed local geometry. Consequently, morphology and local site geometry emerge as the primary candidate design variables for controlling the LEF, while composition remains important insofar as it determines the relaxed structure and secondary electronic characteristics.

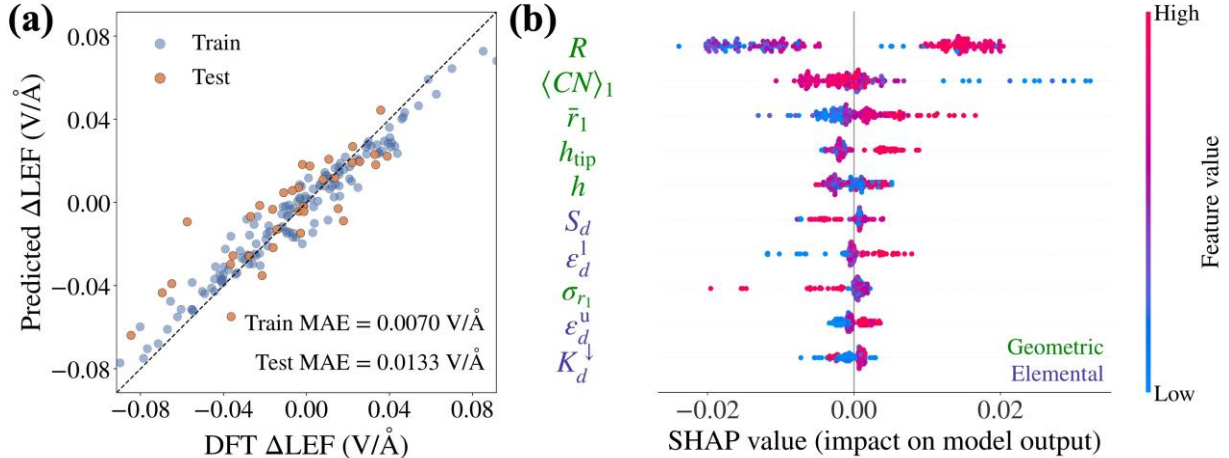


Figure 2. Quantifying the LEF variation associated with metal substitution using the ΔLEF formulation. (a) Predicted versus DFT-computed ΔLEF for the 90-atom nanoclusters at an external electric field of +0.5 V/Å. The dashed line denotes perfect agreement between the model predictions and DFT calculations. (b) SHAP attribution for the ΔLEF model, with descriptors colored by their feature values (blue: low; pink: high). Geometric descriptors are shown in green and elemental electronic descriptors in purple. R is the radial distance to the center of mass; $\langle CN \rangle_1$ is the mean coordination number of the first-shell neighbors; $\bar{r}_1$ and σ_{r_1} are the mean and standard deviation of the first-neighbor distances; h_{tip} is tip height; h is site height; S_d is d -band skewness; ε_d^l and ε_d^u are lower and upper d -band edges, and K_d is d -band kurtosis. The superscript arrow denotes the spin-resolved component.

2.3 LEF prediction with interpretable machine learning

Having established from the matched-site analysis that metal-substitution-associated LEF variations are comparatively small and predominantly associated with structural relaxation, we next ask whether this geometry-dominated hierarchy persists when predicting the absolute LEF across the complete chemical, structural, and field space. The curated dataset contains 12,680 training records and 3,171 test records in an 80:20 split, spanning nine transition metals, twelve cluster sizes, and eleven external-field values. The feature set combines the global, local, and elemental descriptors with the external electric field (**Table S3**), allowing a single model to describe the LEF across all metals, cluster sizes, and applied-field conditions.

Thirteen regression models were benchmarked, spanning linear, neighbor-based, kernel-based, neural-network, and tree-ensemble methods. Their predictive performance is summarized in **Section S7 (Tables S7 and S8; Figure S6)**. XGBoost achieved the lowest seed-averaged test MAE and was therefore selected for the subsequent analysis. **Figure 3a** compares the XGBoost-predicted LEF with the DFT-computed values. The predictions closely follow the parity line over the full LEF range of -0.8 to +0.8 V/Å, with

training and test MAEs of 0.0078 and 0.0131 V/Å, respectively. The model therefore accurately reproduces LEFs within the structural, compositional, and field space represented by the curated dataset.

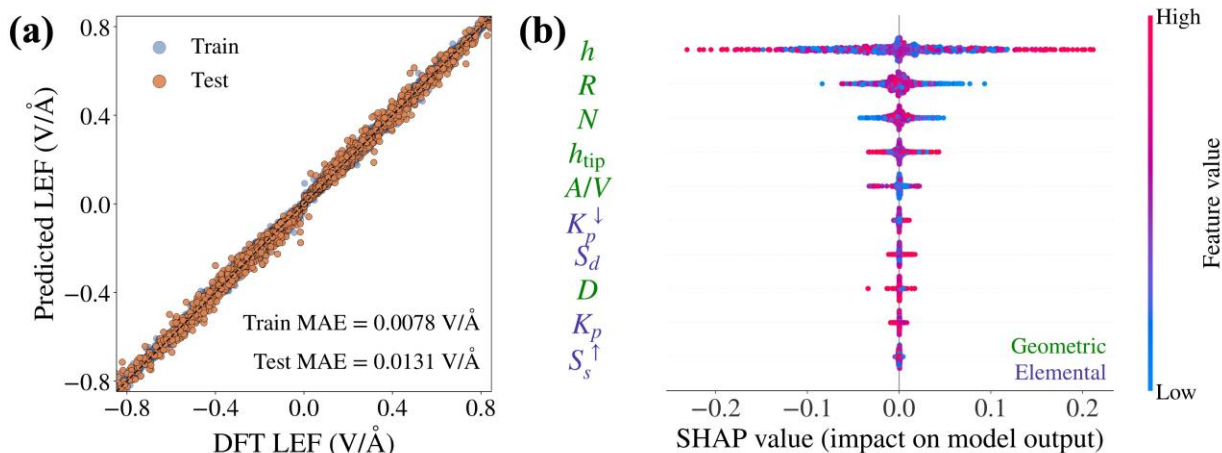


Figure 3. Prediction and interpretation of the local electric field using XGBoost. (a) Predicted versus DFT-computed LEF for the training and test sets. The dashed line denotes perfect agreement between the model predictions and DFT calculations; the training and test MAEs are 0.0078 and 0.0131 V/Å, respectively. (b) SHAP attribution for the LEF model. Descriptors are ranked by their mean absolute SHAP values, and individual points are colored by feature value (blue: low; pink: high). Geometric descriptors are shown in green and elemental electronic descriptors in purple. h is the site height; R is the radial distance to the center of mass; N is the atom count; h_{tip} is the tip height; A/V is the surface-to-volume ratio; K_p is the p -band kurtosis; S_d is the d -band skewness, D is the cluster diameter at the center of mass, and S_s is the s -band skewness. Superscript arrows denote spin-resolved components.

Figure 3b shows the SHAP attribution for predicting the LEF. The five highest-ranked descriptors are all geometric, including the site height (h), the radial distance to the center of mass (R), the atom count (N), the tip height (h_{tip}), and the surface-to-volume ratio (A/V). These descriptors include both local and global geometric information, indicating that the LEF is governed by the position of the adsorption site together with the overall nanocluster morphology.

Electronic descriptors appear below the leading geometric features. Ranks six through ten include the down-spin and the total p -band kurtosis ($K_p^\downarrow$ and K_p), the d -band skewness (S_d), and the up-spin s -band skewness ($S_s^\uparrow$), together with one global geometric descriptor, the cluster diameter at the center of mass (D). Like the Δ LEF analysis, the d -band center does not appear among the ten most influential descriptors. This suggests that the electronic information relevant to LEF prediction is captured more effectively by higher-order

band-shape characteristics than by the d -band center. The electronic contribution is also not confined to the d -band. In these small nanoclusters, the s - and p -band structures differ appreciably from their extended-surface counterparts, allowing their higher-order moments to contribute alongside the d -band descriptors. The projected s -, p -, and d -band densities of states for the nine metals are presented in **Section S4**. Supplementary SHAP analyses are provided in **Section S8**, and **Figure S7**.

Overall, the absolute-LEF and Δ LEF analyses provide two complementary tests of the same physical hierarchy. The matched-site Δ LEF analysis shows that the comparatively small LEF variation accompanying metal substitution is explained predominantly by structural relaxation, whereas the full-dataset analysis shows that geometric descriptors also dominate absolute LEF prediction across different cluster sizes, metals, and applied fields. The consistent absence of the d -band center among the leading descriptors further indicates that the electronic contribution to LEF is not well represented by this conventional d -band center descriptor.

2.4 Closed-form analytical equation from symbolic regression

The consistency between the Δ LEF and absolute-LEF analyses suggests that the dominant LEF physics may be expressible using only the applied field and a small number of geometric variables. We therefore ask whether the high-dimensional descriptor relationship learned by XGBoost can be reduced to a compact analytical expression. We applied symbolic regression using the PySR package²¹ with a loop-based feature-selection procedure to the descriptors retained from the SHAP ranking. Four variables were selected: the external electric field F , the site height h , the tip height h_{ip} , and the radial distance to the center of mass R . These variables form the smallest subset that preserved the predictive performance of the full descriptor set according to the feature-selection criterion described in **Section S9**. Symbolic regression was then performed using these four variables to identify expressions that balance predictive accuracy and equation complexity.

Figure 4a shows the resulting Pareto front of loss as a function of equation complexity. The loss decreases rapidly at low complexity and then changes more gradually beyond approximately complexity eight. Four representative expressions along the Pareto front are listed in **Table 1**. The simplest expression depends only on the external electric field F , while increasing complexity progressively introduces the site height h , the tip height h_{ip} , and the radial distance to the center of mass R . This sequence is consistent with the SHAP analysis in **Section 2.3**, where these geometric quantities rank among the most influential descriptors for predicting the LEF.

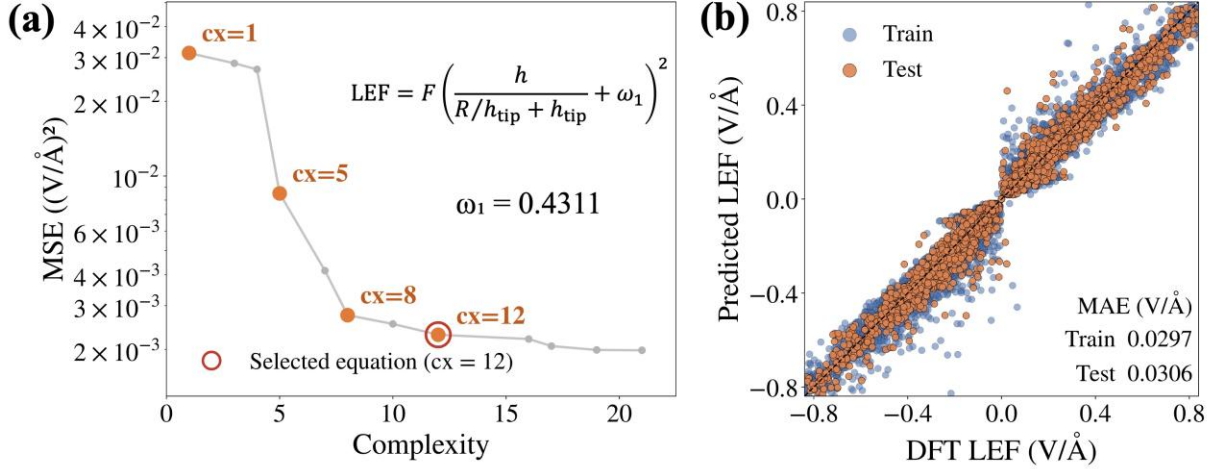


Figure 4. Closed-form analytical description of the local electric field obtained by symbolic regression. (a) Pareto front of prediction loss as a function of equation complexity, with representative expressions selected at complexities 1, 5, 8, and 12. The variables include the external electric field F , the site height h , the tip height h_{tip} , and the radial distance to the center of mass R . (b) Predicted versus DFT-computed LEF using the complexity-12 equation for the training and test sets.

Table 1. Closed-form equations selected at four points along the Pareto front of **Figure 4a**.

Complexity	Equation	Fitted constant
1	$\text{LEF} = F$	—
5	$\text{LEF} = F h \omega_1$	$\omega_1 = 0.3159$
8	$\text{LEF} = F \frac{(h + h_{\text{tip}} \omega_1)^2}{h_{\text{tip}}^2}$	$\omega_1 = 0.3567$
12	$\text{LEF} = F \left(\frac{h}{R/h_{\text{tip}} + h_{\text{tip}}} + \omega_1 \right)^2$	$\omega_1 = 0.4311$

The complexity-12 expression was selected as the final analytical equation because it is the lowest-complexity expression containing all four retained variables and lies near the region where further increases in complexity provide progressively smaller reductions in loss. **Figure 4b** compares this equation with the DFT-computed LEF. The predicted values follow the parity line over the full LEF range of -0.8 to +0.8 V/Å, with training and test MAEs of 0.0297 and 0.0306 V/Å, respectively. The equation reproduces the LEF across the metals and cluster sizes represented in the dataset. No electronic descriptors appear in the selected expression, consistent with the comparatively small LEF variation associated with metal substitution identified in **Section 2.2** and the dominant geometric contributions revealed by SHAP. The fact that the compact expression retains the applied field together with three geometric variables while requiring

no explicit electronic descriptors provides an independent mathematical distillation of the physical hierarchy identified by the SHAP analyses. This does not imply that electronic structure has no influence on the LEF; rather, within the chemical and structural space examined here, its contribution is sufficiently secondary that the dominant LEF variation can be represented without explicit electronic descriptors.

The XGBoost model achieves a lower test MAE of 0.0131 V/Å, whereas the analytical expression gives a test MAE of 0.0306 V/Å. The increase in error is accompanied by a substantial reduction in model complexity, since the LEF can now be evaluated from a single equation using only the applied external electric field and three geometric descriptors. The resulting expression is differentiable with respect to the external field and can therefore be incorporated directly into kinetic models, field-dependent energy corrections, and catalyst-screening workflows without evaluating the trained machine-learning model.

3. Conclusions

We combined DFT calculations with interpretable machine learning to determine whether local electric fields on transition-metal nanoclusters are controlled primarily by metal-dependent electronic properties or catalyst geometry. Matched-site Δ LEF analysis shows that changing the metal produces comparatively small LEF variations and that these variations are explained predominantly by the accompanying structural relaxation. Across the complete dataset, SHAP analysis independently identifies local site position and global cluster morphology as the dominant descriptors of absolute LEF, whereas electronic contributions are secondary and are better represented by higher-order band-shape characteristics than by the conventional d-band center. Symbolic regression further distills this hierarchy into a four-variable expression containing the applied field and three geometric descriptors. In summary, these results identify morphology and local site geometry as the primary design variables for controlling LEFs on the transition-metal nanoclusters examined here, while metal composition influences LEF substantially through its effect on the relaxed structure.

4. Methods

4.1 DFT calculations

All density functional theory (DFT) calculations were performed using the Vienna Ab initio Simulation Package (VASP).²⁵ Core-valence interactions were described using the projector augmented-wave method,²⁶ and exchange-correlation effects were treated with the Perdew-Burke-Ernzerhof (PBE) generalized-gradient approximation.²⁷ A plane-wave kinetic-energy cutoff of 400 eV was used for geometry optimization and the subsequent electric-field and density-of-states calculations. Because the nanoclusters were modeled in large supercells, Brillouin-zone sampling was restricted to the Γ point. Each nanocluster

was separated from its periodic boundary by at least 10 Å of vacuum in all directions, and the structures were relaxed until the residual atomic forces were below 0.03 eV/Å.

The external electric field (EEF) was applied using the Neugebauer-Scheffler²⁸ dipole-correction method, with a uniform field along the z direction ranging from -0.5 to +0.5 V/Å in increments of 0.1 V/Å. For each top adsorption site, the electrostatic potential was sampled along the local vertical line passing through the adsorption-site coordinates. The LEF was obtained from the gradient of this site-specific potential profile along the z direction and evaluated at the center-of-mass position of a CO probe molecule. The same procedure was applied to all adsorption sites, nanocluster sizes, metals, and external-field strengths.

For each of the twelve atom counts, the lowest-formation-energy Ni cluster structure was used as the common starting template for all nine metals and was rescaled using the corresponding relaxed FCC lattice constant before geometry relaxation. This common structural template enables corresponding adsorption sites to be compared across different metals without introducing differences in the initial crystal topology. Co, Ru, and Fe do not adopt FCC as their bulk equilibrium structures; therefore, the FCC configurations used for these metals represent controlled reference structures rather than their bulk ground-state phases. The use of a common lattice type avoids introducing differences in bulk crystal packing into the cross-metal comparison and allows the effects of metal-dependent electronic structure and structural relaxation to be examined within a consistent structural framework. Relaxed lattice constants are reported in **Table S12**, and additional computational details are provided in **Section S10**.

4.2 Dataset cleaning and curation

The initial dataset contained LEF values for top adsorption sites on the upper half of each relaxed nanocluster under eleven external electric fields. Curation retained only base-cluster entries, reduced structural redundancy, required complete electronic descriptors, and aligned site-size pairs to Ir.

Structural redundancy was then reduced using a similarity filter based on fourteen geometric descriptors. Within each metal and atom-count group, the descriptors were standardized and pairwise Euclidean distances were calculated between adsorption sites. Sites separated by less than a distance cutoff, τ , were grouped as structurally similar, and one representative site was retained from each group. The size-specific cutoffs, reported in **Table S2**, were selected in preliminary Δ LEF cross-validation and are therefore target-informed preprocessing parameters.

The dataset was then restricted to site and cluster-size combinations for which corresponding Ir data were available, enabling matched-site Δ LEF comparisons. The 1,477 retained physical adsorption sites

contribute 15,851 site-field records after descriptor-completeness filtering and Ir alignment (**Sections S1 and S2**).

4.3 Descriptor calculation

The descriptors used for machine learning were grouped into global geometric, local geometric, and elemental electronic descriptors, as introduced in **Section 2.1**. Complete definitions and the full descriptor list are provided in **Section S3 and Tables S3 and S4**.

Global and local geometric descriptors were calculated directly from the relaxed nanocluster coordinates. Global descriptors characterize the overall cluster size and morphology, whereas local descriptors describe the geometric environment of the adsorption site, including its position relative to the cluster center and the coordination of neighboring atoms.

Elemental electronic descriptors were obtained from the site-resolved projected density of states of the adsorption-site atom in each relaxed nanocluster. For the *s*, *p*, and *d* states, we calculated quantities describing the band position, occupation, width, shape, and band edges, including their spin-resolved components. These descriptors were evaluated separately for each physical adsorption site and cluster size; external-field replicas of a given site share the same site-specific electronic descriptors rather than using a single bulk or slab reference. Further details of the electronic descriptors and projected density-of-states calculations are provided in **Sections S4**.

4.4 Machine-learning model training and evaluation

Thirteen regression algorithms were evaluated for LEF prediction: ordinary least-squares, ridge, LASSO, and elastic-net regression; k-nearest neighbors; support vector regression; a multilayer perceptron; decision tree; random forest; extra trees; AdaBoost; gradient boosting; and XGBoost.²⁴ All models were implemented using scikit-learn²⁹ and evaluated using the same preprocessing and model-selection procedure. Missing descriptor values were replaced by the mean of the corresponding feature in the training data, followed by feature standardization and model fitting.

The structural-similarity filtering described in **Section 4.2** was performed before train-test partitioning so that structurally redundant sites were reduced to a single representative before model development. For each of ten random seeds, the curated dataset was divided into 80% training and 20% test records, corresponding to 12,680 training and 3,171 test records. Hyperparameters were optimized once by five-fold grid-search cross-validation on the complete curated modeling matrix and then fixed for all ten random record-level splits. Model performance was evaluated on the corresponding test subsets, and the reported model-comparison metrics were averaged over the ten splits. Because tuning preceded the repeated splits

and the same physical site can occur in both subsets, these test metrics quantify interpolation within represented sites and fields rather than fully independent unseen-site generalization. The model comparison, selected XGBoost hyperparameters, and prediction errors are reported in **Section S7, Tables S7 and S8, and Figure S6**.

4.5 SHAP analysis

Feature contributions were analyzed using SHAP (SHapley Additive exPlanations),²⁰ which assigns a contribution from each input descriptor to an individual model prediction. SHAP values were calculated using TreeExplainer for the trained XGBoost models of the absolute LEF and Δ LEF. Descriptors were ranked according to their mean absolute SHAP values, while the signed SHAP values were used to determine the direction of each descriptor contribution.

For the Δ LEF analysis, SHAP values were evaluated for the records included in the 90-atom nanocluster dataset at an external field of +0.5 V/Å. For the absolute-LEF analysis, SHAP values were evaluated for a random subset of 1,000 records from the curated dataset to reduce computational cost while sampling the descriptor range represented in the complete dataset.

Two groups of inputs were excluded from the rankings shown in **Figures 2b and 3b**. Site-label one-hot variables were excluded because their individual SHAP values are not physically interpretable. For the absolute-LEF model, the external electric field was also excluded because it is the dominant predictor and would otherwise compress the contributions of the remaining descriptors (**Sections 2.2 and 2.3, Section S8**).

4.6 Symbolic regression and feature selection

Closed-form expressions for the LEF were obtained using the PySR symbolic-regression package.²¹ Candidate descriptors were ranked according to their mean absolute SHAP values from the trained XGBoost model (**Section 4.5**), after excluding the site-label dummy variables and placing the external electric field at the beginning of the candidate list. Descriptors were then added sequentially following the SHAP ranking, and symbolic regression was performed after each addition. The predictive performance of each descriptor subset was evaluated on the held-out test set, and the smallest subset that preserved the performance of the full descriptor set was retained. This procedure identified four variables: the external electric field F , site height h , tip height h_{tip} , and radial distance R . Symbolic regression was subsequently performed using these four variables to generate a Pareto front relating prediction error to equation complexity. The analytical equation reported in **Section 2.4** was selected from this Pareto front based on

the balance between predictive accuracy and model simplicity. Complete feature-selection criteria, symbolic-regression settings, and the full Pareto front are provided in **Section S9**.

Supporting Information

Additional computational details; nanocluster structural parameters; dataset-cleaning and structural-similarity criteria; descriptor definitions; machine-learning model comparison and hyperparameter optimization; supplementary SHAP analyses; projected density-of-states analysis; and symbolic-regression feature selection and Pareto-front results (PDF).

Data Availability

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files. Should any raw data files be needed in another format they are available from the corresponding author upon reasonable request.

Code Availability

DFT calculations were performed using VASP, which is available from its developers under license. The custom scripts used for data curation, model training and evaluation, SHAP analysis, symbolic regression, and figure generation are available from the corresponding author upon reasonable request.

Author Contributions

F.C. conceived the project, supervised the study, contributed to theory development, assisted in theoretical data analysis, and oversaw project management. R.Z. conducted the DFT calculations and developed the interpretable machine learning model. Q.L. assisted with DFT calculations. J.Y. assisted with the machine learning model development. All authors contributed to the writing and editing of the manuscript and Supporting Information.

Notes

The authors declare no competing financial interests.

Acknowledgement

F.C., R.Z., J.Y., and Q.L. acknowledge the support provided by the Department of Energy, Basic Energy Science, Catalysis Science, Early Career Research Program under the award number of DE-SC SC0026399. F.C., R.Z., J.Y., and Q.L. also extend their thanks for the computational resources provided by the San Diego Supercomputer Center at University of California San Diego, and the Texas A&M High Performance

Research Computing FASTER cluster, through allocation no. CHM220074 and no. CHE200076 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program. Additionally, this research used the resources of the National Energy Research Scientific Computing Center (NERSC), a US Department of Energy Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC awards BES-ERCAP0027465, BES-ERCAP0028368 and BES-ERCAP0036444.

References

- (1) Zhao, R.; Singh, P.; Li, Q.; Yang, J.; Deshlahra, P.; Liu, H.; Che, F. Field-Enhanced catalysis: Integrating experiment, theory, and machine learning for catalytic innovation. *Applied Catalysis B: Environment and Energy* **2026**, 382, 125901. DOI: 10.1016/j.apcatb.2025.125901.
- (2) Yang, C.; Guo, Y.; Zhang, H.; Guo, X. Utilization of Electric Fields to Modulate Molecular Activities on the Nanoscale: From Physical Properties to Chemical Reactions. *Chemical Reviews* **2025**, 125 (1), 223-293. DOI: 10.1021/acs.chemrev.4c00327.
- (3) Aragonès, A. C.; Haworth, N. L.; Darwish, N.; Ciampi, S.; Bloomfield, N. J.; Wallace, G. G.; Diez-Perez, I.; Coote, M. L. Electrostatic catalysis of a Diels–Alder reaction. *Nature* **2016**, 531 (7592), 88-91. DOI: 10.1038/nature16989.
- (4) Shaik, S.; Mandal, D.; Ramanan, R. Oriented electric fields as future smart reagents in chemistry. *Nature Chemistry* **2016**, 8 (12), 1091-1098. DOI: 10.1038/nchem.2651.
- (5) Che, F.; Gray, J. T.; Ha, S.; Kruse, N.; Scott, S. L.; McEwen, J.-S. Elucidating the Roles of Electric Fields in Catalysis: A Perspective. *ACS Catalysis* **2018**, 8 (6), 5153-5174. DOI: 10.1021/acscatal.7b02899.
- (6) Wan, M.; Yue, H.; Notarangelo, J.; Liu, H.; Che, F. Deep learning-assisted investigation of electric field–dipole effects on catalytic ammonia synthesis. *JACS Au* **2022**, 2 (6), 1338-1349. DOI: 10.1021/jacsau.2c00003.
- (7) Che, F.; Gray, J. T.; Ha, S.; McEwen, J.-S. Reducing reaction temperature, steam requirements, and coke formation during methane steam reforming using electric fields: A microkinetic modeling and experimental Study. *ACS Catalysis* **2017**, 7 (10), 6957-6968. DOI: 10.1021/acscatal.7b01587.
- (8) Léonard, N. G.; Dhaoui, R.; Chantarojsiri, T.; Yang, J. Y. Electric Fields in Catalysis: From Enzymes to Molecular Catalysts. *ACS Catalysis* **2021**, 11 (17), 10923-10932. DOI: 10.1021/acscatal.1c02084.
- (9) Welborn, V. V.; Ruiz Pestana, L.; Head-Gordon, T. Computational optimization of electric fields for better catalysis design. *Nature Catalysis* **2018**, 1 (9), 649-655. DOI: 10.1038/s41929-018-0109-2.
- (10) Guo, W.; Zhang, S.; Zhang, J.; Wu, H.; Ma, Y.; Song, Y.; Cheng, L.; Chang, L.; Li, G.; Liu, Y.; et al. Accelerating multielectron reduction at CuxO nanograins interfaces with controlled local electric field. *Nature Communications* **2023**, 14 (1), 7383. DOI: 10.1038/s41467-023-43303-1.
- (11) Chen, L.; Guo, Z.; Huang, H.-Z.; Tian, W.; Chang, X.; Hu, Q.; Liu, D.; Sun, C.; Gao, M.; Han, S.; et al. Electric-field-driven CO₂ polarization and bioinspired proton blocking unlock CO₂ reduction in strong acid without metal cations. *Nature Communications* **2026**, 17 (1), 1734. DOI: 10.1038/s41467-026-68435-y.
- (12) Zhao, R.; Li, Q.; Yang, J.; Zhu, C.; Che, F. Integrating Physical Principles with Machine Learning for Predicting Field-Enhanced Catalysis. *JACS Au* **2025**, 5 (3), 1121-1132. DOI: 10.1021/jacsau.4c00901.
- (13) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T.; Sautet, P. Introducing structural sensitivity into adsorption–energy scaling relations by means of coordination numbers. *Nature chemistry* **2015**, 7 (5), 403-410. DOI: 10.1038/nchem.2226.
- (14) Calle-Vallejo, F. The ABC of generalized coordination numbers and their use as a descriptor in electrocatalysis. *Advanced Science* **2023**, 10 (20). DOI: 10.1002/advs.202207644.
- (15) Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science* **2015**, 350 (6257), 185-189.
- (16) Hammer, B.; Nørskov, J. K. Theoretical surface science and catalysis—calculations and concepts. In *Advances in catalysis*, Vol. 45; Elsevier, 2000; pp 71-129.
- (17) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chemistry* **2009**, 1 (1), 37-46. DOI: 10.1038/nchem.121.
- (18) Xin, H.; Vojvodic, A.; Voss, J.; Nørskov, J. K.; Abild-Pedersen, F. Effects of d-band shape on the surface reactivity of transition-metal alloys. *Physical Review B* **2014**, 89 (11), 115114.
- (19) Vojvodic, A.; Nørskov, J.; Abild-Pedersen, F. Electronic structure effects in transition metal surface chemistry. *Topics in catalysis* **2014**, 57 (1), 25-32.

- (20) Feng, D.-C.; Wang, W.-J.; Mangalathu, S.; Taciroglu, E. Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. *Journal of Structural Engineering* **2021**, *147* (11), 04021173.
- (21) Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression. *jl. arXiv preprint arXiv:2305.01582* **2023**.
- (22) Lambeets, S. V.; Cardwell, N.; Onyango, I.; Wirth, M. G.; Vo, E.; Wang, Y.; Gaspard, P.; Ivory, C. F.; Perea, D. E.; Visart de Bocarmé, T.; et al. Elucidating the Role of Electric Fields in Fe Oxidation via an Environmental Atom Probe. *Angewandte Chemie* **2025**, *137* (18), e202423434. DOI: <https://doi.org/10.1002/ange.202423434> (accessed 2026/08/17).
- (23) Nørskov, J. K.; Studt, F.; Abild-Pedersen, F.; Bligaard, T. *Fundamental concepts in heterogeneous catalysis*; John Wiley & Sons, 2014.
- (24) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016; pp 785-794.
- (25) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996**, *54* (16), 11169-11186. DOI: 10.1103/PhysRevB.54.11169.
- (26) Blöchl, P. E. Projector augmented-wave method. *Physical Review B* **1994**, *50* (24), 17953-17979. DOI: 10.1103/PhysRevB.50.17953.
- (27) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **1996**, *77* (18), 3865-3868. DOI: 10.1103/PhysRevLett.77.3865.
- (28) Neugebauer, J.; Scheffler, M. Adsorbate-substrate and adsorbate-adsorbate interactions of Na and K adlayers on Al (111). *Physical Review B* **1992**, *46* (24), 16067. DOI: 10.1103/PhysRevB.46.16067.
- (29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825-2830.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)