Alchemical perturbation density functional theory (APDFT) has great promise for enabling rapid and accurate computational screening of hypothetical catalyst sites, but first order approximations are unsatisfactorily inaccurate when alchemical derivatives are large. In this work, we analyze errors in first order APDFT calculation schemes for binding energies of CHx, NHx, OHx, and OOH adsorbates over a range of different coverages on hypothetical alloys based on a Pt(111) reference system. We then construct feature vectors by fingerprinting the dopant locations in the alloy and then use a data set of about 11,100 data points to train three different support vector regression machine learning models that correct systematic APDFT prediction errors for each of the three classes of carbon, nitrogen, and oxygen based adsorbates. While uncorrected first order APDFT alone can approximate reasonably accurate adsorbate binding energies on up to 36 hypothetical alloys based on a single Kohn-Sham DFT calculation on a 3 × 3 unit cell for Pt(111), the machine learning-corrected APDFT in principle extends this number to more than 20,000 and provides a recipe for developing other machine learning models to aid future high throughput screening studies.
Machine Learning Models Correct Systematic Errors in Alchemical Perturbation Density Functional Theory Applications to Catalysis

Charles D. Griego¹ | Lingyan Zhao¹ | Karthikeyan Saravanan¹ | John A. Keith¹

¹Department of Chemical and Petroleum Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania, 15261, United States

Correspondence
Department of Chemical and Petroleum Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania, 15261, United States
Email: jakeith@pitt.edu

Funding information
National Science Foundation, Graduate Research Fellowship Program, Grant/Award Number: 1747452; National Science Foundation, Grant/Award Number: CHE-1856460; American Chemical Society Petroleum Research Fund: 55595-DNI4

Alchemical perturbation density functional theory (APDFT) has great promise for enabling rapid and accurate computational screening of hypothetical catalyst sites, but first order approximations are unsatisfactorily inaccurate when alchemical derivatives are large. In this work, we analyze errors in first order APDFT calculation schemes for binding energies of CHₓ, NHₓ, OHₓ, and OOH adsorbates over a range of different coverages on hypothetical alloys based on a Pt(111) reference system. We then construct feature vectors by fingerprinting the dopant locations in the alloy and then use a data set of about 11,100 data points to train three different support vector regression machine learning models that correct systematic APDFT prediction errors for each of the three classes of carbon, nitrogen, and oxygen based adsorbates. While uncorrected first order APDFT alone can approximate reasonably accurate adsorbate binding energies on up to 36 hypothetical alloys based on a single Kohn-Sham DFT calculation on a 3 × 3 unit cell for Pt(111), the machine learning-corrected APDFT in principle extends this number to more than 20,000 and provides a recipe for developing other machine learning models to aid future high throughput screening studies.
1 | INTRODUCTION

Many efforts in computational catalysis are focused on screening materials for innovative catalysts that promote high chemical activity.[1][2][3] Descriptors for catalyst activity, e.g. an adsorbate binding energy (BE) on a catalyst surface, are frequently obtained using Kohn-Sham density functional theory (DFT) calculations because they are both reasonably accurate and computationally feasible for studying systems that might contain up to about 200 atoms. However, even relatively efficient methods such as DFT are too computationally costly to systematically model many different adsorbates at different coverages on different facets and compositions of hypothetical alloys. There have been efforts to address such points by developing approximate models such as cluster expansions,[4][5][6] coordination number models,[7][8][9] as well as an assortment of machine learning models[10][11][12][13][14][15][16] that all can be trained from quantum mechanics calculations but leveraged to make predictions on systems outside of the training set.

An alternative approach that technically requires no a priori model training is alchemical perturbation density functional theory (APDFT).[17][18][19][20] This approach is sometimes referred to as "computational alchemy", but it should not be confused with computational schemes where atomic or molecular interactions are gradually tuned out over the course of a molecular dynamics simulation to define more convenient paths for carrying out thermodynamic integration.[21] In contrast, APDFT is generally carried out with static quantum mechanics calculations and is a means to predict a hypothetical energy contribution based on a perturbation on a reference system. In APDFT, the hypothetical energy contribution is based on an approximate Taylor series expansion in terms of differentials in atomic electrostatic potentials (so called "alchemical derivatives"), see Eq. 1.

\[
\Delta E^0|_{\lambda=1} = \Delta E^0|_{\lambda=0} + \partial_\lambda \Delta E^0 \Delta \lambda + \frac{1}{2} \partial_\lambda^2 \Delta E^0 \Delta \lambda^2 + \ldots
\]  

Here, the energy of the perturbed system (\(\Delta E^0|_{\lambda=1}\)) is the sum of the energy of the reference system (\(\Delta E^0|_{\lambda=0}\)) and terms involving alchemical derivatives, where the nth derivative is represented by \(\partial_\lambda^n \Delta E^0\). APDFT has been used in several diverse applications such as for predicting molecular reaction energetics, deprotonation energies, and properties of BN-doped variants of benzene, coronene, fullerene, and graphene. APDFT has also been tested for predicting material properties of bulk transition metals, semiconductor band gaps and stability, and molecular adsorptions on nanoparticles and extended surfaces. The majority of these applications have used the Taylor series expression truncated to first order, which thus assigns a linear relationship between the energy perturbation and the alchemical derivatives of the reference shown in Eq. 2.

\[
\partial_\lambda \Delta E^0 = \sum_I \Delta \mu_{\text{n}_I} \partial_\lambda \text{n}_I - \sum_I \Delta F_I \partial_\lambda \text{R}_I + \Delta \mu_{\text{e}} \partial_\lambda \text{n}_e
\]  

The terms above, in order, represent the energy gradients of nuclear chemical potential (\(\Delta \mu_{\text{n}_I}\)) due to variation in nuclear charge (\(\text{n}_I\)), forces on atoms (\(\Delta F_I\)) resulting from changes in atomic positions (\(\text{R}_I\)), and the electronic chemical potential (\(\Delta \mu_{\text{e}}\)) due to variation in number of electrons (\(\text{n}_e\)). First order APDFT approximations require relatively simple arithmetic operations, and they bring effectively no additional computational cost once a reference calculation has been calculated with a quantum mechanics method such as DFT. Non-linear contributions in principle can also be
treated using higher order terms, but these require more quantum mechanics calculations. Note that throughout the rest of this work, ‘APDFT’ will refer to APDFT truncated to first order.

A tutorial on catalysis applications using APDFT was provided before, but we briefly summarize key points here. The objective in APDFT catalysis applications is to approximate an adsorbate BE or a barrier height on a hypothetical catalyst surface without running an explicit quantum mechanics calculation on that system. To do this, the hypothetical surface must be created by changing (or ‘transmuting’) atoms from a reference system that already has been explicitly calculated. The key parameters for generating a hypothetical surface are the number of atoms transmuted to create the hypothetical system \(N_T\), and the magnitude of the change in nuclear charge for that transmutation \(\Delta Z\). To simplify Eq. 2 as much as possible, nuclear coordinates are fixed, and so energy contributions due to relaxation energies that follow transmutations are presently not addressed. Furthermore, transmutations are done in pairs so that the total number of electrons in both the reference and hypothetical systems remain the same. In other words, if a Pt atom in the reference system is to be changed into an Au atom for the hypothetical system, another Pt atom in the reference system needs to be changed into an Ir atom somewhere in the hypothetical system. APDFT with the simple, first order Taylor series approximation of the energy contribution is given by Eq. 3

\[
\Delta E^0_{\lambda=1} = \Delta E^0_{\lambda=0} + \sum_T \Delta \mu_{nt} \delta_I N_I
\]  

In our previous work, we showed that the accuracy of APDFT approximations requires that \(N_T\) and \(\Delta Z\) should both be small. In general, this means that APDFT approximations are most accurate when the hypothetical system does not have more than a single atom transmuted by more than one group of the periodic table. Furthermore, alchemical derivatives in the reference system should also be relatively small. This means that APDFT approximations are most accurate when the hypothetical system does not have a transmuted atom directly under an adsorbate bound at an on-top site. Again, recall that first order APDFT assumes a linear relationship between alchemical derivatives and the change in an energy contribution. When \(N_T\), \(\Delta Z\), and/or the alchemical derivatives are large, that indicates that the linear approximation is no longer valid.

Even when the above considerations are accounted, there are still relatively large numbers of hypothetical systems for which APDFT can be used. For example, consider a model of a \(2 \times 2\) surface unit cell for Pt(111). With four atoms in each layer, there will be eight transmutable atoms in the top two layers since these would be closest to the binding site and therefore the most important for impacting catalysis. If the adsorbate does not bind to an on-top site, APDFT should be reasonably accurate for when transmuting any one of these eight atoms by \(\Delta Z = \pm 1\). To constrain the system to be isoelectronic, a second atom far from the binding site is transmuted (usually in the bottom layer of the slab) by \(\Delta Z = \mp 1\). In such a model, APDFT could be expected to predict reasonably accurate adsorbate binding energies for \(2 \cdot \binom{8}{1} = 16\) hypothetical systems with effectively no computational cost. Here the binomial notation is used for showing the combinatorics of one transmuted atom in any of the eight atomic sites in the top two layers of this system, and the prefactor of 2 refers to the fact that transmutations could be either +1 (into Au) or -1 (into Ir). Note that the utility of APDFT increases if there is a larger \(3 \times 3\) unit cell model for Pt(111) where there are now a total of 18 atoms within the top two layers. The number of hypothetical systems accurately modeled would then be \(2 \cdot \binom{18}{1} = 36\). APDFT could be even more useful if both \(N_T\) and \(\Delta Z\) could be further increased in magnitude since this would substantially increase the number of hypothetical combinations. For instance, if one or two transmutations were done in the top two layers, and each hypothetical alloy involved Pt being transmuted into Os (\(\Delta Z = -2\)), Ir (\(\Delta Z = -1\)), Au (\(\Delta Z = +1\)), or Hg (\(\Delta Z = +2\)) along with appropriate transmutations in the bottom layer for isoelectricity, a single DFT calculation on the \(2 \times 2\) model would grant up to \(2 \cdot \binom{2}{1} + 2 \cdot \binom{2}{2} = 6\) APDFT predictions, while a single DFT calculation on the \(3 \times 3\) model would grant up to \(2 \cdot (2 \cdot \binom{18}{1} + 2 \cdot \binom{18}{2}) = 684\) APDFT
predictions. Note these combinatorial estimates provide an upper estimate of possible adsorbate structures since they do not account for symmetries within each unit cell. More refined enumerations are possible with graph-theory based techniques.[32]

Here, we investigated the feasibility of correcting errors in first order APDFT using machine learning (ML) approaches. We hypothesized that errors in adsorbate binding energies in these systems reflect a change from a linear energy relationship with alchemical derivatives to an unknown non-linear relationship. Thus, ML models that effectively use statistical techniques to learn non-linear relationships from training sets of data should be a productive way to correct these errors.

Indeed, ML has wide applications in biology, chemistry and materials science.[33, 34, 35, 36, 37] Notable examples range from small molecule chemistry to protein interactions to materials screening.[38, 39, 40, 41, 42, 43, 44, 45] ML models have also been applied to catalyst screening studies.[46] Combined ML and descriptor-based kinetic analysis framework was found to rapidly screen bimetallic catalysts for CO$_2$ reduction.[11, 12] Fully automated screening methods that combined ML and optimization techniques facilitated screening through a massive chemical space of inter-metallic catalysts.[15] Catalyst activity predictions were also accelerated coupling ML with active site configurations,[47] scaling relations,[10] and volcano plots.[48] We assessed the relationship between accuracy (relative to DFT) and the properties of systems evaluated with first order APDFT, we chose features that most correlate to APDFT’s accuracy, and we trained and tested a machine learning model on a large data set of adsorbate BEs on hypothetical alloys.

2 | COMPUTATIONAL METHODS

All DFT calculations in this work were carried out using PBE[49] as the exchange-correlation functional and projector augmented wave pseudopotentials as implemented in VASP.[50] An energy cutoff of 350 eV and an 8x8x1 Monkhorst-Pack grid sampling of $k$-point were used for all models. Structures were relaxed using conjugate gradient iterative minimization until the difference in energies between consecutive geometry steps was less than 1 meV. Adsorbate binding energies were calculated using the equation $BE = E_{site} + E_{ads} - E_{ads-site}$. Where $E_{site}$ is the energy of the bare catalyst surface, $E_{ads}$ the energy of the catalyst with an adsorbate bound to the surface site, and $E_{ads-site}$ is the energy of the lone molecular adsorbate.

In this work, we benchmark BEs for different systems on a four-layer Pt(111) surface with the lower two layers fixed and the upper two layers relaxed. The adsorbates we chose are CH$_x$ and NH$_x$ where $x = 0 - 3$, as well as OH$_x$ where $x = 0 - 2$ and OOH. Surface coverage ($\theta$) is calculated as the number of adsorbates per atom in a surface layer. Here we model $\theta = 1, 1/4, 1/9$ by placing one adsorbate molecule on 1x1, 2x2 and 3x3 slab models. We transmute atoms in the top two layers by $\pm \Delta Z$ and maintain isoelectronicity by transmuting an atom in the bottom layer by $\pm \Delta Z$.

Once a reference BE was obtained from DFT calculations, we constructed a list of atom−centered electrostatic potential differences. Then, we constructed a second list that describes the $\Delta Z$ change to each atom in the slab model after we made transmutations. We approximated $\Delta BE_{APDFT}$ by taking the dot product of the two lists and benchmarking it to $\Delta BE_{DFT}$ values obtained from single point DFT calculations (again, relaxed with respect to adsorption of the reference system but not with respect to alchemical transmutations). Detailed descriptions of this method are described in our previous work.[31] For each reference system, we evaluated alloys made from $N_T = 1 - 4$ and with $\Delta Z = \pm 1, 2, 3$. All possible $N_T$ changes were made among all sites in the first layer and the second layer for the 1x1 and 2x2 surface models. For the 3x3 model, we chose a random subset of about 150 configurations to reduce the number of DFT calculations needed to benchmark since 24,282 configurations are combinatorially possi-
ble. The data production was facilitated using Phystone, an APDFT Python library developed by our team found at https://github.com/KeithLabPitt/comp_alchemy.

We first assessed the errors with APDFT predicted BE with respect to different characteristics of the alloys and the adsorbate binding. To analyze these variables individually, we have bucketed the APDFT BE errors from our entire dataset into categories. In Figure 1, we show bar graphs separated by coverage. Individual bars show the mean absolute error (MAE) with respect to both $\Delta Z$ and $N_T$. For $\theta = 1/9$, the MAE for all cases is roughly below 0.6 eV and generally increases when $\Delta Z$ or $N_T$ increase. For $\theta = 1/4$, the MAE for each case is below 1.6 eV and systematically increases with $\Delta Z$ and $N_T$. For $\theta = 1$, $N_T = 1$ or 2 because there are only two sites available in the first and second layers. We found that the MAE is below 2 eV and increases with both variables. Overall, MAE increased when the coverage increased. This is because with higher coverage, the concentration of sites with large alchemical derivatives, that overestimate BE, were greater.

In Figure 2, we compare the MAE with respect to the type of adsorbate in the reference system. Here, the adsorbate type is described by the amount of central atom hydrogenation (i.e., the value of $x$ in CH$_x$, NH$_x$, or OH$_x$ (or OOH)). There are two notable observations from this analysis. The APDFT predicted MAE decreases with the degree of increasing hydrogenation and when the central atom in the adsorbate is farther to the right of the periodic table. In Figures 1 and 2, we see strong dependence of MAE with $\Delta Z$, $N_T$, $\theta$, and the type of adsorbate, suggesting that these variables are viable features for making predictions.

3 | MACHINE LEARNING MODEL

Here we developed ML models that can predict the magnitude of errors of APDFT relative to DFT accuracy. Figure 3 shows the workflow we followed to obtain ML corrected APDFT BE predictions. With a hypothetical alloy surface as the input, we fingerprinted the hypothetical dopant configuration that resulted from transmutations, and we constructed a feature vector from these fingerprints and other descriptors unique to the adsorbate-alloy system. The feature vector was fed to the model and outputted a predicted value of the APDFT’s deviation from DFT. The error was then added to the original APDFT predictions as a correction.

In order for a machine learning model to correctly learn the underlying patterns in the training data, the most relevant features must be identified and processed. These features are variables that describe particular properties among all groups in a large dataset. We included in our model the variables that strongly correlate with BE predictions errors as identified in Figures 1 and 2: $\Delta Z$, $N_T$, $\theta$, and type of adsorbate. While these variables give a simple description of the kind of adsorption and the type of changes made in the reference system, a feature representation of the dopant location in the alloys would provide a much more unique description of the alloy to link to APDFT prediction accuracy. This was done by considering a feature vector with length equal to the number of atoms in the top two layers of our slab models. Each element of the vector represents one of the sites in the top two layers where a binary indication is used to describe the site being transmuted or not. For more details, our feature vector construction is explained in the supporting information. Once we construct the feature vector for the dopant configuration, we add additional elements with binary descriptions of the selected features.

The dataset we use to train our model includes all surface alloys outlined in the error analysis above. For each of these alloys, a feature vector was constructed as outlined above, and the error in BE between APDFT and DFT was calculated. To most accurately assess ML models, we partition the data into a training set and a test set and make validations using k-fold cross validation (CV) (more information in the supporting information).

We considered several machine learning algorithms to predict APDFT error corrections. We evaluated support
TABLE 1  Summary of machine learning model evaluation and selection based on the CV MAEs for each dataset.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MAE CH⁺ (eV)</th>
<th>MAE NH⁺ (eV)</th>
<th>MAE OH⁺ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR (RBF)</td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>SVR (Polynomial)</td>
<td>0.17</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.11</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>Linear</td>
<td>0.31</td>
<td>0.37</td>
<td>0.35</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.31</td>
<td>0.37</td>
<td>0.35</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.43</td>
<td>0.44</td>
<td>0.36</td>
</tr>
</tbody>
</table>

MAE, mean absolute error; SVR, support vector regression; RBF, radial basis function.

vector regression (SVR) models with radial basis function (RBF) and polynomial kernels, random forest, linear regression, ridge regression, and elastic net regression by measuring the performance with a 10-fold CV. The CV MAE for each model trained using the CH⁺, NH⁺, and OH⁺ subsets is listed in Table 1. The SVR model with the RBF kernel gave the best prediction accuracy, and the hyperparameters were optimized also with a 10-fold CV. We found that the default parameters provided through the scikit-learn library [51] gave the best performing version of this model. With the optimized hyperparameters, the model is retrained for each subset. When building the final models, we split the data of each subset into 80% training and 20% testing partitions. The assessment of this split, the learning performance, and how well the model is generalized is shown in Figure 4. We have plotted the MAE of model predictions using both the training (red curve) and test (green curve) set against the training sample size. With increasing training samples (BEs), the training score curve converges and becomes stable at a minimum MAE. The test score curve also exhibits this behavior and quickly approaches the training score curve with increasing training samples.

As shown in the final steps of Figure 5, our models output predicted errors that are added to the APDFT calculated BE as a correction. Three separate SVR models were trained using datasets containing CH⁺, NH⁺, OH⁺, and OOH species. Figure 5 shows pure APDFT (a) and ML-corrected APDFT (b) ΔBE predictions for alloys in the CH⁺ test set against respective DFT predictions of ΔBE. Similar plots for NH⁺ and OH⁺ adsorbate test sets are included in the supporting information. Upon ML corrections, the prediction accuracy notably increases for each case, and with many improved cases where APDFT systematically under and over-predicted the BE. For CH⁺, MAE decreased by more than an order of magnitude from 0.78 to 0.07 eV. MAE decreases from 0.65 to 0.07 eV for NH⁺, and for OH⁺ and OOH, the MAE decreased from 0.39 to 0.06 eV.

We then investigated which features were most important in the model training. Each feature was scored with a normalized negative log of the p-value. These values are recorded in the bar chart in Figure 6. A feature is most important if the value is close to 1 and least important when it is close to 0. We initially found it curious that the transmutation site is the least important feature because our previous work reports that APDFT’s largest prediction errors belong to systems where the transmutation is at the binding location. However, this result is sensible because there are up to 18 possible transmutations sites in our models, and only a relatively small number of these sites are close to the adsorbate. This means that it is less common to see in the dataset cases where large errors come from changing sites near the adsorbate. Conversely, this argument explains why ∆Z is the most important feature. It was also reported in our previous work [29, 30] (and shown in Figure 1) that for a larger nuclear charge perturbation, the errors are larger, regardless of the location of the change. Thus, for the part of the dataset where the change is large (ie. ∆Z = 3), almost all cases have large errors, and this feature plays a greater role in the model training.
Since $\Delta Z$ and $N_T$ were the top two most important features, we outlined how our models were especially effective in improving predictions for alloys made from large $\Delta Z$ and $N_T$ transmutations, where first order APDFT fails. Figure 7 illustrates the distribution and level of improvement of predictions with alloys in the test data set relative to $\Delta Z$ and $N_T$. The radius of each circle corresponds to the percentage of alloys present in the test data set made from that particular pair of $\Delta Z$ and $N_T$. Each circle is colored based on the difference in MAEs for the uncorrected APDFT and ML-corrected APDFT in that set of alloys. For example, dark blue represents that ML-corrections effectively resulted in zero improvement while dark red represents substantial improvement by more than 1 eV. Note that alloys with larger $N_T$ had more possible combinations than smaller $N_T$ and thus comprised greater percentages of the data set. With increasing $\Delta Z$, the relative percentages of different alloys that were tested does not change, but the magnitude of the corrected MAE significantly increases. This shows where ML corrections are most impactful while also showing which subsets of alloys that were tested are most at risk for being underfit. For advanced ML models that interface with APDFT, it is recommended to use of datasets that account for intrinsic symmetries of each adsorbate configuration as well as to focus on allocating even more training data on hypothetical alloys using the larger $|\Delta Z|$ and $N_T$ values.

4 | CONCLUSIONS

We have demonstrated how an ML-based treatment can significantly improve the accuracy of first order APDFT BE predictions of a range of adsorbates on hypothetical alloys of Pt(111). First, we identify the source of errors related to the system variables. Generally, the MAE among systems in our dataset increases with increasing $\Delta Z$, $N_T$, and coverage. There is also a correlation to the type of adsorbate in the reference system as the MAE is found to decrease with increasing levels of hydrogenation at the central atom. We used this insight to fingerprint our alloy systems and construct feature vectors that include these variables. We found through a $k$-fold CV analysis that an SVR model with a radial basis function kernel gave the most accurate prediction of the APDFT errors. Training separate SVR models on CH$_x$, NH$_x$, and OH$_x$ datasets, BE predictions for systems in our test set had accuracy improved by 0.71, 0.58, and 0.33 eV, respectively. Each model used at least 3,600 BE data points for training and were validated using at least 900 data points. While an appreciable amount of DFT calculations were needed to develop the machine learning models, they should be useful beyond the number of BE predictions in our test set. Note that the section of our dataset where $\theta = 1/9$ (using a 3×3 surface model) only includes a rather small subset of all possible combinations involving $N_T = 1, 2, 3, 4$ and $\Delta Z = \pm1, \pm2, \pm3$ transmutations among the 18 atoms in the top two layers. In this case, there are $2 \cdot (3 \cdot (\binom{18}{1}) + 3 \cdot (\binom{18}{2}) + 3 \cdot (\binom{18}{3}) + 3 \cdot (\binom{18}{4})) = 24,282$ hypothetical alloy configurations that should be modeled reasonably accurately given a single reference BE calculation and a trained ML model, and only about 600 of those were used for training.

Upon investigating the feature importance, we found that the $\Delta Z$ and $N_T$ variables were most important, while the choice in the transmutation site was least important in the model training. This is apparent because choices in $\Delta Z$ and $N_T$ that are higher in magnitude lead to more systems in the dataset with low prediction accuracy. These findings suggest that ML models provide a straightforward and practical remedy for first order APDFT predictions of BEs, and that ML models can in principle achieve the same accuracy as DFT calculations. Future work will still focus on developing schemes to more correct these errors more physically with higher order terms as well as ensure that ML training sets are not underfit. This is expected to allow more accurate predictions of catalysis through materials space.
acknowledgements

We acknowledge financial support from the U.S. National Science Foundation: GRFP-1747452 (for CDG), CHE-1856460, and support from the donors of the Petroleum Research Fund, administered by the American Chemical Society, ACS PRF Award 55595-DNI4. Calculations were run using resources provided by the University of Pittsburgh Center for Research Computing.

references


GRAPHICAL ABSTRACT
**FIGURE 1** Mean absolute error (MAE) with respect to change in nuclear charge at each transmuted site ($\Delta Z$) and the number of transmuted atoms ($N_T$) for binding energy calculations of all adsorbates at coverages ($\theta$) = 1/9, 1/4, 1.
**Figure 2** MAE with respect to hydrogenation of the central atom for CH$_x$, NH$_x$, and OH$_x$ adsorbate BE.

**Figure 3** Machine learning workflow for obtaining corrected alchemy BE predictions.
FIGURE 4  Learning curves of the model performance with respect to the number of systems used for model training and testing (% partitioned from the entire dataset). Model performance is measured by the MAE (eV) of predictions compared to DFT. The convergence of MAE with respect to training data show the models reaching the learning limit. The learning curve for test data also converges with increasing percentage of data and approaches the training learning curve. This behavior shows the model generalization, the ability to make accurate predictions with unseen data.
**FIGURE 5** Parity plot of the BE change ($\Delta BE$) of CH$_x$ adsorbates on alloys of transmuted Pt predicted with APDFT (a) and ML-corrected APDFT (b) compared to DFT. $\Delta BE$ are in units of eV. The MAE decreases from 0.78 (a) to 0.07 eV (b) upon ML-correction. See the supporting information for analogous plots for other adsorbates (NH$_x$ and OH$_x$).

**FIGURE 6** Feature importance plot showing the significance of different parameters for the machine learning model. Features are scored based on a normalized negative log of the p-value, where the most important feature ($\Delta Z$) ranks at a score of 1.0
FIGURE 7  Visualization of the distribution of the types of hypothetical alloys in the test data set and how well ML improves APDFT predictions for those types of alloys. Circle radii correspond to the percentage of alloys in the test data set that are made from $N_T$ transmutations using $\Delta Z$, and the filled color corresponds to the difference in the MAE after ML corrections have been applied. Numerical data that composes this figure are provided in the supporting information.
Supporting Information for: Machine Learning Models Correct Systematic Errors in Alchemical Perturbation Density Functional Theory Applications to Catalysis

Charles D. Griego¹ | Lingyan Zhao¹ | Karthikeyan Saravanan¹ | John A. Keith¹

Contents

- Numbers of systems evaluated S2
- Feature vector construction S3
- Model Validation S4
- NH₃ parity plots S5
- OH₃ and OOH parity plots S6
- Error analysis by alloy type S7
In this section, we explain how many unique binding energies (BEs) were calculated. We considered twelve adsorbates in total (\(CH_x, NH_x\), with \(x = 0 - 3\), \(OH_x\), with \(x = 0 - 2\), and OOH) and modelled multiple adsorbate coverages (\(\theta = 1, 1/4, 1/9\)) using 1\(\times\)1, 2\(\times\)2, and 3\(\times\)3 surface slab dimensions. Hypothetical alloys were formed from each slab by making 1-4 numbers of transmutations (\(N_T\)) with nuclear charge change of \(\Delta Z = -3, -2, -1, 1, 2, 3\) (six possibilities).

For the 1\(\times\)1 and 2\(\times\)2 slabs, the total number of hypothetical alloys we made was equal to the number of ways you can choose \(N_T\) changes of \(\Delta Z\) from all sites in the top two layers (two and eight, for 1\(\times\)1 and 2\(\times\)2 respectively). For 1\(\times\)1, we could only do \(N_T = 1\) or 2 transmutations, with six possible choices in \(\Delta Z\), and twelve adsorbates. The total number of systems evaluated was \(6 \cdot 12 \cdot \left(\binom{2}{1} + \binom{2}{2}\right) = 216\), where the binomial notation is used for showing the resulting combinations of \(N_T\) transmuted atom in the two atomic sites in the top two layers. For the 2\(\times\)2 slab, the total number of systems was \(6 \cdot 12 \cdot \left(\binom{8}{1} + \binom{8}{2} + \binom{8}{3} + \binom{8}{4}\right) = 11664\). For the 3\(\times\)3 slab, we could not evaluate all possible systems because this would require as many as 291,384 DFT calculations for benchmarking. Instead, we chose approximately 2,000 systems at random.
Feature vector construction

We represented the dopant location in the alloys by considering a feature vector with length equal to the number of atoms in each surface layers of our slab models, and each element of the vector represents an atomic site. This presents a problem where the length of the vector varies depending on the size of the surface unit cell used to model varying adsorbate coverage. We remediated this by normalizing the vector, making its length equal to the number of atoms in a layer of the largest unit cell included in the dataset. This is illustrated in Figure S1 where we show top views of the first and second layers of three varying sized unit cells of Pt with Au dopants. Since a 3x3 unit cell of fcc Pt(111) was the largest unit cell included, and we considered all atoms in the top two layers, the length of the feature vector 18 elements. Additionally, the positions of each atom in the vector are set according to their positions in the 3x3 unit cell. In Figure S1, the atoms in the first layer are indexed from 0 to 8, following the atom ordering from the unit cell origin at the bottom left to the top right. Since atoms indexed 0, 4, and 8 are Au dopants, we encode the vector with 1’s at these positions. If we follow this constraint when encoding changes to the first layer of a 2x2 unit cell, we would only allow the elements whose positions match those in the 3x3 unit cell to be changed to 1’s, and the remaining elements would be a 0. According to Figure S1, we are only able to encode atoms indexed by 0, 1, 3, 4, 9, 10, 12, and 13 in a 2x2 unit cell. Since atoms 0 and 4 are Au, those two positions are encoded as 1’s. The padded zeroes, for positions that can’t be encoded for a 2x2 unit cell, are colored in red. For a 1x1 unit cell, only atoms indexed by 0 and 9 can be encoded, and atom 0 is an Au dopant, so the zeroth position of the vector is changed to 1. The remaining positions that can’t be changed are the padded zeroes colored in red.

For features like type of adsorbate, coverage, and resulting atom, we add an element to our vector for each possibility in those groups. For example, we add three elements that represent possibilities of $\theta$, and the first, second, and third pertain to coverages of 1/9, 1/4, and 1, respectively. If our input alloy is a 3x3 unit cell, the first element would equal 1 and the second and third would be zero.
Model Validation

Our models were validated with a test set partitioned from the entire dataset. A test set is necessary because if you make predictions on data using a model that was trained with the same data, the conclusions are trivial. Typically, the data is split at a ratio where a majority of the data is used to train our model, and with the amount of remaining test data, a reasonable number of predictions are made to validate the model's performance. This tactic is useful when addressing overfitting, a common problem that is encountered when trying to make reliable and transferable machine learning models. A model is overfit when the fit is optimized so close to the training data that the performance is poor when making predictions from new data. Overfitting can be avoided with resampling techniques that employ the train-test split approach like k-fold cross validation (CV). In k-fold CV, the training set is split into k numbers of smaller datasets. For k iterations, a model is trained using k – 1 folds of training data from the k-fold split dataset, and the resulting model is validated on the remaining fold using the mean absolute error (MAE). For our model selection, we performed 10-fold CV on each, and the average MAE of all k iterations measured the performance of the model.
NH$_x$ parity plots

Figure S2 compares first order APDFT (a) and ML-corrected APDFT (b) $\Delta BE$ predictions for alloys in the NH$_x$ test set against respective DFT predictions of $\Delta BE$.

**FIGURE S2** Parity plot of the BE change ($\Delta BE$) of NH$_x$ adsorbates on alloys of transmuted Pt predicted with APDFT (a) and ML-corrected APDFT (b) compared to DFT. $\Delta BE$ are in units of eV. The MAE decreases from 0.65 (a) to 0.07 eV (b) upon ML-correction.
Figure S3 compares first order APDFT (a) and ML-corrected APDFT (b) $\Delta BE$ predictions for alloys in the OH$_x$ and OOH test set against respective DFT predictions of $\Delta BE$.

**FIGURE S3** Parity plot of the BE change ($\Delta BE$) of OH$_x$ adsorbates on alloys of transmuted Pt predicted with APDFT (a) and ML-corrected APDFT (b) compared to DFT. $\Delta BE$ are in units of eV. The MAE decreases from 0.39 (a) to 0.06 eV (b) upon ML-correction.
Error analysis by alloy type

We provide a numerical breakdown of errors across the full training set, where we specify the percentage of data and reduction of errors with respect to $\Delta Z$ and $N_T$.

**TABLE S1** Distribution of the types of hypothetical alloys found in the test data set and how well ML improves APDFT predictions for those types of alloys.

<table>
<thead>
<tr>
<th>$\Delta Z$</th>
<th>$N_T$</th>
<th>% of Test Data Set</th>
<th>$\Delta$ MAE (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3.1</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7.3</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9.8</td>
<td>0.16</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>12.9</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3.7</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>7.1</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>10.5</td>
<td>0.58</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>12.5</td>
<td>0.76</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2.5</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7.3</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>10.7</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>12.9</td>
<td>1.23</td>
</tr>
</tbody>
</table>

MAE, mean absolute error.