Machine Learning Identification of Active Sites in Graphite-Conjugated Catalysts

Kunal Lodaya, ¹ Nathan D. Ricke, ¹ Kelly Chen, ¹ Troy Van Voorhis¹*

¹Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

*Corresponding Author. Email: tvan@mit.edu. Phone: 617-253-1488
Abstract
Graphite-conjugated catalysts (GCCs) are a promising class of materials that combine many of the advantages of heterogenous and homogeneous catalysts. In particular, GCCs containing an aryl-pyridinium active site appear to be effective nonmetal catalysts for the oxygen reduction reaction (ORR). In this study, we analyzed both structural and electronic properties of a dataset of molecules containing nitrogen atoms embedded in aromatic molecules in order to understand which properties enable a particular site to bind O₂, which is a necessary step for the initiation of ORR. We found that carbon atoms ortho or para to nitrogen and at the edge of aromatic systems are especially likely to be active. Using both structural and electronic features to describe the individual atoms in each catalyst, we trained machine learning models capable of identifying catalyst active sites. Although permutation importance of the features used to train these models indicates that several key electronic features have the greatest impact on model performance, the model trained only on structural features still proved effective in simulated catalyst discovery scenarios where the objective is affected more by false positives than false negatives.
1. Introduction

The discovery of new catalysts that drive essential chemical reactions is a perpetual objective in chemistry. In light of the burgeoning availability of low cost solar electricity, a particularly timely challenge is the generation of materials that competently catalyze fuel-forming small molecule transformations which enable the storage of electricity as chemical energy.\textsuperscript{1–6} This chemical energy subsequently requires a corresponding set of materials to catalyze the conversion of this chemical energy back into electricity in a device known as a fuel cell.\textsuperscript{7–9} One of these reactions, the reduction of oxygen to water, is essential to proton exchange membrane fuel cell technologies, but significant research efforts are still required to make these technologies a reality.\textsuperscript{10} Platinum remains the best-in-class catalyst for O\textsubscript{2} reduction, but its relatively high cost and low abundance are prohibitive to widespread adoption.\textsuperscript{10} In its place, various first-row transition metal and metal-free alternatives have been considered.\textsuperscript{11–13} Among these alternative catalyst paradigms, nitrogen-doped carbon materials stand out as durable, low-cost, high-activity and high-efficiency catalysts for oxygen reduction.\textsuperscript{14,15} These structures come in various forms, including nitrogen-doped fullerenes, carbon nanotubes, and graphene.\textsuperscript{16,17} Regardless of type, though, these materials are intrinsically heterogeneous, featuring mixtures of pyridine-like, pyrrole-like and graphitic/quaternary nitrogen sites.\textsuperscript{18} This has complicated the study of active sites in these catalysts, rendering the rational iteration and improvement of these materials more difficult. To address this gap, the development of “designer” nitrogen-doped materials with well-defined active sites has been pursued by multiple groups.\textsuperscript{19,20}

These studies support the critical conclusion that carbons adjacent to pyridinic moieties are responsible for the oxygen reduction activity in these catalysts but raised further questions about the source of this activity. In the case of a previous study by Ricke et al. which focused on one particular nitrogen-doped graphite conjugated catalyst (N\textsuperscript{+}-GCC), there was only a single carbon atom in the catalyst that would form a bond with O\textsubscript{2}.\textsuperscript{19} This activity could not be solely a result of the pyridinic nitrogen acting as an ortho-para activator to the aromatic system, as the alternate carbon ortho to the pyridinic nitrogen would not bind O\textsubscript{2}. This study noted that the active carbon had a significant accumulation of spin density, which the authors interpreted as indicating this site could most stably donate electron density to O\textsubscript{2} during bond formation.

In this work, we build on this understanding to choose a set of structural and electronic features which we hypothesize could predict oxygen binding and ORR activity in nitrogen-doped carbon materials. Interpretable molecular features can be used not only to guide the manual search for new catalyst structures by rational design,\textsuperscript{21–24} but also to train machine learning (ML) models that predict activity with very low computational cost.\textsuperscript{25,26} If sufficiently accurate and generalizable, these ML models can be applied to perform high-throughput virtual screening (HTVS) of larger libraries of existing and generated structures than DFT could screen alone,\textsuperscript{27} facilitating the discovery of more chemically diverse catalysts with higher activity.

The overall process of developing the dataset, generating features, and applying machine learning models to predict O\textsubscript{2}-binding is shown in Figure 1. This approach enables us to quantitatively analyze the effect of both structural and electronic properties of the catalysts, providing statistically motivated insights on which carbon atoms are likely to be catalytically
active for any given N+-GCCs. We find that carbon atoms that are likely active sites are an odd number of bonds from a nitrogen atom, are part of only a single aromatic ring, and accumulate spin density when an excess electron is present on the catalyst. Further, such carbons are most likely to be active if they are part of a catalyst that can be more easily ionized, and if the change in partial charge at that site changes significantly between the neutral and cationic state of the catalyst. Although the electronic features explored in this work are the most important predictors of activity, the HTVS scenarios we explore show that models trained on structural features alone can still perform well for certain goals.

Figure 1: The pipeline for identifying molecules with ORR active sites using DFT and ML. A. Molecules with pyridinic nitrogen atoms collected from the ChemSpider database. B. Optimized geometries for these molecules using either force fields or DFT, and subsequently generate a combination of structural and electronic features. C. DFT O₂ binding calculations at two different atomic sites in a catalyst, one that is active to binding O₂ (left) and one that does not bind O₂ (right). D. ML model trained to predict which atomic sites bind O₂.

This paper is organized as follows. Section 2 describes how the dataset was generated and used to train atom-centric ML models. Section 3 describes the performance of the ML models, the relative feature importance for electronic and structural features, and evaluates the relative performance of ML models for several HTVS catalyst discovery scenarios. Section 4 concludes by discussing how these results inform future catalyst discovery.
2. Computational methods

Compilation of molecules
We collected a dataset of 258 molecules via a simple structure search for pyridinic moieties (cationic nitrogen atoms encased in an aromatic system) in the ChemSpider database. The molecules in this database are derived directly from literature, and as such all have established synthetic procedures. We downloaded the molecules from the database in Simplified Molecular Input Line Entry System (SMILES) format. We then converted to Cartesian coordinates via 3D structure generation in OpenBabel software and parsed manually for irregularities or insensible chemical structures prior to further calculations.

Quantum chemical geometry optimization and data collection
We first optimized the structures with the Merck Molecular Force Field method (MMFF94), implemented using OpenBabel. To refine the geometries, we subsequently optimized the geometries using DFT. We ran single point DFT calculations on both the force field-optimized and DFT-optimized geometries to determine the quantum chemical characteristics used in our analysis. For all catalysts, these single-point energies were computed for both the neutral and cationic forms. All DFT calculations were performed in Q-Chem using the PBE0 functional and 6-31g* basis set.

Starting from the DFT-optimized geometries, we ran O_2-binding DFT calculations to enumerate all possible O_2-binding locations and energies. As the known active sites for nitrogen-doped carbon catalysts are all carbon atoms, we prepared each input geometry by placing one O_2 per calculation at a distance of 1.55 Å oriented with a C-O-O bond angle of 120° from each carbon in the dataset of 258 molecules. This totaled to 4039 O_2 binding calculations. We then calculated the optimized geometries for each of these systems using DFT and generated electronic features as before. These calculations provided the output sets for our supervised learning tasks: the binary binding/non-binding determination for the classification task, and the O_2-binding energy for the regression task. We calculated the O_2 binding energy using the formula, \( E_{\text{O}_2 \text{Binding}} = E_{\text{CatO}_2} - E_{\text{Cat}} - E_{\text{O}_2} \), momentarily neglecting entropy and solvent corrections. We then used the values of \( E_{\text{O}_2 \text{Binding}} \) and the optimized geometry of each calculation to label whether the site binds O_2; a full description of this labeling process is in section 1 of the SI, but in the majority of cases we labeled a site as active if the C-O bond length was less than 2 Å and the binding energy was less than -0.1 eV. We chose to use a threshold of -0.1 eV because there were a number of cases where O_2 did not form a bond, ending the optimization at over 3 Å from the catalyst, but \( E_{\text{O}_2 \text{Binding}} \) was between 0 and -0.1 eV due to fractional electron transfer to O_2. Furthermore, given that the success of the O_2-binding reaction at such weakly-binding sites can be sensitive to the initial geometry used, we determined that the most consistent way to label these sites was as non-binding for O_2 (see SI section 1).

Free Energy, ORR, and Overpotential
We assume in this work that the ORR mechanism for all catalysts proceeds through four proton-coupled electron transfer steps as follows:

\[
R + O_2 \rightleftharpoons RO_2
\]
\[
RO_2 + H^+ + e^- \rightleftharpoons RO_2H
\]
\[
RO_2H + H^+ + e^- \rightleftharpoons RO + H_2O
\]
\[ RO + H^+ + e^- \rightleftharpoons ROH \]
\[ ROH + H^+ + e^- \rightleftharpoons R + H_2O \]

For the purpose of calculating overpotential, we grouped the chemical step 1 with the following step 2 to form the electrochemical step \[ RO_2 + H^+ + e^- \rightleftharpoons RO_2H \]. For each carbon in each of the molecules in the dataset that bound \( O_2 \), we ran DFT geometry optimization calculations for every intermediate in the catalytic cycle. Rather than performing normal mode vibrational calculations for all of these molecules to calculate the free energy, we performed these calculations on a single representative molecular analogue. Using these free energy corrections, we applied the method used by Baran et. al.\(^{38}\) to compute the reaction free energy for each step in this catalytic cycle.

Feature generation and preparation

Various approaches have been reported for feature design in machine learning applications within the chemical literature, which can broadly be separated into two categories: generalized strategies encoding as many physical properties of molecules as possible, and hand-picked approaches that select relevant features that support a mechanistic hypothesis. In the former case, feature generation has seen extensive development, including the Coulomb Matrix, Bag of Bonds, crystal structure representation, and fingerprinting methods.\(^{39-42}\) While these approaches have been shown to be suitable for a range of broad machine learning problems in chemistry without clear or obvious structure-output relationships, we chose the second class of feature design due to our clear hypotheses for relevant features to use to solve our learning tasks. In particular, we sought out electronic and structural features to encode information about the charge density and bond structure of candidate carbon atom \( O_2 \)-binding sites. These include the atomic charge differences in the neutral and cationic states molecule, the unpaired spin density, and properties of related to aromaticity. For the full list of features prepared for the machine learning models developed in this work, see SI section 2.

Machine learning algorithms and cross validation

In this work, we evaluated six different ML methods for this application: logistic regression,\(^{43}\) decision trees,\(^{44}\) linear support vectors,\(^{45}\) radial basis support vectors,\(^{46}\) multi-layer perceptrons,\(^{47,48}\) and random forests.\(^{49}\) All ML algorithms were implemented using the Python library Scikit-learn.\(^{50}\) We chose 10-fold cross validation to evaluate and report model accuracy in this work. Although ML has been applied in a wide variety of studies, spanning organic synthesis, materials discovery, and catalysis,\(^{51-56}\) this study differs from many of these approaches in that the models we describe in this work are trained to make predictions for each atom rather than the entire molecule. As the features and targets are specific to the atoms in each catalyst, we split the full dataset into training and testing data by molecule during 10-fold cross validation. This avoids scenarios in which a single catalyst’s sites are split between the training and validation data, where information about the overall activity of a catalyst could leak from the training data into predictions on the validation data. This distinction is critical due to our aforementioned focus on atomic-level prediction, rather than molecular prediction.

3. Results and Discussion

Classifier decision functions performance

For the ORR reaction, one of the most relevant classification questions is to answer, “is this atom capable of binding \( O_2 \)”\? We trained several classifiers using different decision functions on
the features collected from the catalyst dataset. In each case, we trained the machine learning model to the DFT-predicted \( \text{O}_2 \)-binding capability for each of the 4039 sites in our data set, using only descriptors from the unbound structure. The accuracy of each model is shown in Table 1. To visually evaluate the classifiers, we projected the decision boundaries of the classifiers onto the two continuous features most strongly correlated with activity, which are spin density and charge difference (Figure 2). Figure 2 shows that a high spin density on a carbon combined with a large negative change in the CHELPG\textsuperscript{57} partial charge difference leads to a strong probability of activity.

The decision boundaries in Figure 3 show that each method is able to capture this broad trend, but the varying ability to capture the additional non-linear boundaries for these two features illustrates their broader success for the full set of features. For linear regression and linear support vectors, the decision boundary is weighted heavily toward deciding based on charge difference in order to capture the points that are active due to large charge differences despite low spin density, but purely linear methods cannot simultaneously do this and readily identify sites that are active due to sufficiently high spin density despite a smaller charge difference. In contrast, decision trees and radial basis support vectors are able to capture this nonlinearity, but they do not appear to have cleanly captured the general trend and achieve roughly the same performance. The random forest classifier (RFC) model demonstrates superior performance relative to the other models, so we elected to use RFC models for further analysis on active site classification.

![Figure 2: The \( \text{O}_2 \)-binding ability of each carbon of the catalysts in the dataset, projected on the Mulliken excess spin population (Spin Density), and the CHELPG charge difference upon ionization (Partial Charge Difference) for each atomic site.](image)

Figure 2: The \( \text{O}_2 \)-binding ability of each carbon of the catalysts in the dataset, projected on the Mulliken excess spin population (Spin Density), and the CHELPG charge difference upon ionization (Partial Charge Difference) for each atomic site.
Figure 3: The decision boundaries for binding O₂ at atomic sites of classifier methods: A. Logistic Regression, B. Decision Tree, C. Multilayer Perceptron Neural Network, D. Linear Support Vector, E. Radial Basis Support Vector, F. Random Forest.

Table 1: Classifier performance by type, determined by 10-fold cross validation

<table>
<thead>
<tr>
<th>Classifier Type</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>92.1% (±1.5%)</td>
</tr>
<tr>
<td>Decision tree</td>
<td>92.2% (±1.3%)</td>
</tr>
<tr>
<td>Multilayer perceptron NN</td>
<td>94.3% (±1.2%)</td>
</tr>
<tr>
<td>Linear support vector</td>
<td>92.5% (±1.7%)</td>
</tr>
<tr>
<td>Radial basis support vector</td>
<td>93.4% (±1.1%)</td>
</tr>
<tr>
<td>Random forest classifier</td>
<td>95.3% (±1.1%)</td>
</tr>
</tbody>
</table>

Distribution of classifier probabilities

Although generating quantum chemistry features can be useful for developing a conceptual understanding of ORR for these catalysts, it is important to address whether the compute time of DFT is justified for HTVS in this case, or whether one might be able to achieve similar results without performing any DFT simulations at all. We trained a separate random forest classifier using only features obtained from the structure of each catalyst (Table S2), which obtained a classification accuracy of 87.0% (±1.8%), corresponding to a 277% increase in the misclassified sites (13.0% compared to 4.7%) relative to the classifier’s performance with electronic features. Note that for the distribution of catalysts in this study, 20% of the candidate sites are active and 90% of the candidate molecules contain at least one active site.

Feature importance

In order to elucidate design principles for developing N⁺-GCC ORR catalysts, we analyzed the impact of the input data on the classifier accuracy using permutation feature importance (Figure 4). This analysis identified Partial Charge Difference, the change in CHELPG charge at a carbon atom when the molecule is ionized, and Spin Density, the Mulliken spin population at the active site, as the two most important features for predicting whether a particular carbon will bind O₂. The ionization energy also appears to have a significant contribution, and between the impact of these three features (none of which are derived directly from the structure) it is clear why electronic features improved the model accuracy from 87% to 95%. Each of these features also reflects some aspect of the process of drawing electrons off of the molecule. In this dataset, the carbons most likely to be active are the ones with a high spin density and a large change in charge density after ionization, and which are in molecules with a low ionization energy. Overall, these features identify which atoms can most stably donate electron density to O₂ during bond formation.
Figure 4: The permutation importance of the features in the random forest classifier. The box-and-whisker for each feature shows the distribution of that feature’s importance across 10 permutation trials. The three most important features, Partial Charge Difference, Spin Density, and Ionization Energy, are all derived from DFT calculations.

Although electronic features are the three most important features identified by this analysis, the next five most important features are all derived directly from the structure. The Ortho or Para feature, which indicates whether an active site is ortho or para to a nitrogen atom in an aromatic ring, and the Distance to Nitrogen feature both demonstrate that the pyridinic moieties have a significant local impact on catalytic activity. For the Nearest Nitrogen Distance feature, **Figure 5** shows that carbon atoms an odd number of bonds away from a nitrogen are more likely to be active. This can be thought of as ortho-para activation to O₂ binding generalized to polyaromatic systems. Carbons that are three bonds away from a nitrogen, even if they are technically not para because they are in an adjoining aromatic ring, are still activated.

The next most important feature derived directly from the structure is Ring Embedding Count, which quantifies the number of aromatic rings a particular carbon is embedded in. Specifically, the Ring Embedding Count feature is the number of aromatic rings sharing a particular carbon, which takes a value from 0 (if it is not in a ring) to 3. As seen in **Figure 6**, the carbon atoms
most likely to be active in this dataset are members of exactly one aromatic ring. The carbons that are not members of aromatic rings are generally inactive in this dataset as they are usually in alkyl groups. The 1.1% of sites not in rings that still bind O₂ are the terminal atoms of alkenes, which appear to bind oxygen quite strongly, whereas no alkanes bound oxygen. The relative activity of the sites at the edges of aromatic systems is likely due to the fact that the active carbon becomes sp³-hybridized as it binds oxygen. Active sites at the edge of an aromatic system have a non-aromatic neighbor that can shift out of the plane of the aromatic system to accommodate the new sp³ hybridization without disrupting the geometry of the aromatic system. In contrast, active sites embedded in two or three aromatic rings must flex out of the plane of the aromatic system when they bind O₂, which has more bond angle strain due to disrupting a larger region of the aromatic system than sp³ hybridization at the edge sites causes.

This analysis provides a number of principles to consider when designing N⁺-GCCs for ORR. Based on the features discussed so far, a candidate molecule containing a pyridinic moiety ought to contain at least one carbon at the edge of an aromatic system that is one or three bonds from the nitrogen and can readily donate charge during O₂ binding. In contrast, the size of the rings in the aromatic system and the total number of nitrogen atoms or other heteroatoms in the molecule do not significantly influence the predicted catalytic activity of the molecules in this dataset, beyond how they affect the important electronic properties mentioned previously.

Figure 5: Percentage of carbon atoms a specified number of bonds away from the nearest nitrogen that are active in O₂-binding. Carbons that are a smaller and/or odd number of bonds away from nitrogen are more likely to be active.
Regression

Although active site classification methods can accelerate HTVS on their own, the impact of these ML approaches is much greater when the most active sites can be identified with regression models that predict the strength of O$_2$ binding at each site. There are two diverging approaches to developing regression models on this dataset: either choose a value for the inactive sites, or train only on the data from active sites. We considered both approaches in this work (for the regression model results trained on active and inactive sites, see section 3 of the SI). Using the same set of features as for the classification problem, the multilayer linear perceptron (MLP) regression model trained on the active site data to predict O$_2$ binding energy attained a mean absolute error (MAE) of 0.077 ± 0.010 eV and $R^2$ of 0.70 (Figure 7). The MAE increased to 0.125 eV when the model was restricted to structural features alone. The reduced accuracy in the latter case could perhaps be acceptable for HTVS, indicating that among active sites connectivity may be enough to predict relative activity. However, given that accurate classification of active and inactive sites already requires the electronic features from DFT, including them in the MLP regression model as well comes at no additional cost and clearly does improve the accuracy of the model. We chose to use MLP in the regression scenario because although the random forest regressor (RFR) model we trained achieved nearly identical performance with an MAE of 0.077 ± 0.009 eV, RFR cannot extrapolate outside of the training data.$^{58}$ This means that if there were catalysts in a search space that were better than

---

Figure 6: Percent of carbon atoms that are active to binding O$_2$ categorized by the number of aromatic rings the carbon atom is a part of. The molecules on the right show example atoms for each value of this feature, with numbers corresponding to the number of aromatic rings the atoms highlighted in red are a part of.
the ones in the training data the RFR model would not be able to predict their higher performance, whereas the MLP model could.

Figure 7: DFT values compared to MLP predicted values of O\(_2\) binding energy. A model with no error would predict exactly along the black line. The points shown are predictions on the test set for all of the splits of 10-fold cross validation combined.

Using the full set of features, we conducted the same permutation importance analysis on the model features that we did for the classifier (Figure 8). This method once again identifies several electronic features as the most important. In particular, Ionization Energy and Neighbor Charge Difference have an even more pronounced contribution here than in the classification model. As before, this suggests that the carbons that can more readily bind O\(_2\) tend to be those that can more easily rehybridize and donate electron density to O\(_2\) during bond formation. However, structural features such as Hydrogen Count, Ring Embedding Count, and Ortho or Para still contribute to the model’s performance, and their diminished importance likely stems from survivorship bias in the distribution: only sites that actually bind O\(_2\) have a quantifiable O\(_2\)
binding energy, so only these sites can be considered and discrete features that can disable activity entirely will have a smaller apparent impact.

Figure 8: Permutation importance of the features in the MLP regression model for predicting O₂ binding energy.
High-Throughput Virtual Screening Strategies using Machine Learning

To evaluate the effectiveness of the machine learning models that we developed, we considered three different HTVS scenarios and various possible strategies available in each case. We consider different scenarios because the available resources and requirements for a catalyst influence the optimal HTVS strategy. In both scenarios, we use the model predictions on the data withheld as the test set in each part of 10-fold cross validation such that we can reuse the entire dataset as the search space. The metric for computational cost in each scenario is the number of DFT O$_2$-binding calculations performed, as this is by far the most expensive step of HTVS for these catalysts. Although it would be desirable to eliminate O$_2$-binding calculations entirely, we assume for all strategies (even those that use machine learning) that they are necessary to verify that a particular site is indeed active. This is how such a search would likely be done in practice, as DFT will always be trusted over a model trained on DFT and verifying the ML model with DFT calculations, increasing the probability that the desired criteria are met when the catalyst is tested experimentally. This is important because experimental confirmation of a catalyst’s activity is generally orders of magnitude more expensive than the DFT compute time for that catalyst.

![Figure 9](image-url)

*Figure 9: Number of DFT O$_2$-binding calculations to find 100 active catalysts using the ML classifier (left) and sampling sites randomly (right). The error bars show the standard deviation of each of these strategies over 10 trials.*

In the first scenario, the goal was to find 100 catalysts each with at least one confirmed active site. This scenario applies whenever the relative activity ordering of the catalysts is less
important than the possibility of catalytic activity, for example when identifying structures that
can be used as starting points for further modification, or which will be filtered using additional
models or criteria. One strategy, which requires only DFT, is to try binding $\text{O}_2$ to potential active
sites at random until an active one is found. We compare this to a different strategy, which uses
our classifier to predict which sites are most likely to be active and then runs $\text{O}_2$-binding
calculations at only those sites. In this scenario, it is not necessary to find all of the active sites
for a particular catalyst, only to confirm that at least one is present, so in both methods all of a
catalyst's sites were removed from the search space as soon as the first active site on that
catalyst was found. The number of DFT $\text{O}_2$ binding calculations necessary to find 100 active
catalysts using each of these strategies is shown in **Figure 9**. The number of calculations over
10 trials using the classifier trained with all features averaged to $101.3 \pm 0.4$, the classifier
trained using only structural features averaged to $115.2 \pm 2.6$ calculations, and the randomly
testing sites averaged to $576.6 \pm 58.1$ calculations. The performance of the classifier trained on
structural features overall is not unexpected, as its accuracy in this trial was similar to its
accuracy of 88.3% during cross-validation. In contrast, the performance of the classifier trained
on all features is much better than a classification accuracy of 95.3% suggested, indicating this
classifier had a much greater accuracy for the sites it was most confident were active. This
principle is illustrated in **Figure 10**, which depicts the RFC class probability for the active and
inactive sites. Notably, there are only a handful of inactive sites that the model is incorrectly
confident are active, entirely analogous to the RFC's performance in this HTVS scenario. As at
minimum 100 $\text{O}_2$-binding calculations are necessary in this scenario to confirm the active sites
in this scenario, there is little room for further improvement on this classifier.

**Figure 10**: The binary cross entropy predictions of site activity by the random forest classifier. The Mulliken excess
spin population is plotted on the y-axis to aid in visualizing the active site density. The active sites are shown as blue
circles, and the inactive sites are shown as red x's. On the far left and the far right of the plot are the points where the
classifier is most confident in its prediction, and in turn this plot shows that the model's accuracy is much higher in these regions of high confidence.

Figure 11: The ratio of the number of DFT O\(_2\) binding calculations necessary to find varying quantities of the set of 15 catalyst active sites that bind O\(_2\) most strongly using two strategies compared to random sampling. The regression strategy uses a regression ML model to predict O\(_2\) binding strength alone to choose sites to test, whereas the mixed approach uses a classifier ML model to select a subset of sites that are then chosen based on a regression ML model. The error bars show the standard deviation of this speedup ratio over 10 trials. The molecule in the upper right of the figure has the highest DFT predicted O\(_2\) binding energies of all molecules in the dataset.

In the second scenario we developed, the goal was to find active sites within the set of active sites that bind O\(_2\) most strongly. In contrast to the previous scenario, the goal here is focused on finding particular active sites rather than catalysts, so catalysts are not removed from the search.
space after one active site on them is found as a catalyst could have multiple active sites. This scenario could apply, for example, if the activity of a set of catalysts were found to be limited by the binding strength of oxygen intermediates and the goal were to increase the binding strength. This goal is similar to the scenario where there is a specific known target value for the binding strength of all oxygen intermediates, as is common for many classes of ORR catalysts with strongly correlated intermediate energies frequently depicted in what are referred to as volcano plots.\textsuperscript{37,59} As before, the baseline we will compare against is the strategy that chooses active sites randomly. Specifically, this randomized approach selects catalysts randomly and binds O\textsubscript{2} to all possible active sites. We designated the 15 strongest binding sites (out of the total 4039) as the set of “best” sites that each method is trying to find.

Beyond the baseline strategy, we considered three approaches for this scenario. In the first, we set the O\textsubscript{2} binding energy for inactive sites to 0.0 eV and trained the regression model on the full set of sites. We then ordered the active sites to search based on the predicted strongest O\textsubscript{2} binding energy for all sites and conducted the simulated search of active sites in that order. The second and third approaches restrict the search to sites a classifier predicts to be active, and then searches those sites based on the order predicted by the regression model that was only trained on active site O\textsubscript{2} binding energies. The distinction between the second and third approaches was in the training features used: one used all structural and electronic features available, and the other used only the structural features. The speedup ratios of each method relative to randomly sampling are shown in Figure 11. Despite notable differences in the model architectures, features, and accuracies on validation datasets, the performances of the models are remarkably similar. In particular, despite a significant difference in their overall accuracy, the classification step for both of the sequenced classification-regression models identified the same 14 search target active sites as active, in each case misclassifying a different one of the 15 as inactive. This indicates that the properties cause very strong O\textsubscript{2} binding for these molecules can be predicted largely from the structure alone, and in such cases can be predicted with a good degree of certainty, diminishing the impact of electronic features and classification accuracy. In addition to model performance, Figure 11 also depicts the molecule with the strongest O\textsubscript{2} binding strength among all molecules considered in this study. This molecule possesses a rather small aromatic system around the active site where O\textsubscript{2} bound, which is a feature that was common to the other molecules that bound O\textsubscript{2} most strongly. If these molecules were converted into GCCs through addition of a diamine that was conjugated to graphite, the O\textsubscript{2} binding strength for this subset of molecules would likely change, and this effect should be considered in future studies.

Up to this point, we have used classification and regression to focus on only one step in the ORR cycle (O\textsubscript{2} binding) as previous studies have indicated binding oxygen to the catalyst is involved in limiting the rate.\textsuperscript{60} In order to connect these methods more directly to the process of discovering new N\textsuperscript{+}-GCC molecules that are ORR active, we applied this approach to search directly for the molecules in this dataset with the smallest ORR overpotential predicted by DFT (see Methods section for details). The MLP regressor using the same architecture as before achieved an MAE of 0.085 ± 0.021 on predicting overpotential for each catalyst when trained with all features, and 0.123 ± 0.014 when trained with only structural features. Analogous to the previous scenario, we conducted a simulated search within our dataset, with two differences: the target property is overpotential rather than O\textsubscript{2} binding, and we adjusted the target search space to the top 50 catalysts (Figure 12).
As before, in this simulated search the approach guided by ML predictions is many times faster than searching randomly, despite the added complexity of predicting a property dependent on multiple intermediates. The speedup ratio is smaller to the degree expected given the increase from 15 to 50 catalysts in the search space. We made this change because when we initially considered a search space of the top 15 catalysts, the classification method trained only on structural features did not correctly identify a single active site from that set, indicating that the catalysts with the highest predicted overpotential, which are not the same as the ones that bind O₂ most strongly, are much harder to classify as active based on structural features alone. Expanding the set to the top 50 catalysts, the structure only classification method was able to find eight sites, whereas the classification method trained on all features found 43. The speedup ratio for the sequenced classification/regression approach trained with all features and applied to the top 15 sites in this scenario is quite similar to the results from scenario two (Figure S2).

![Figure 12](image1.png)

*Figure 12: The ratio of the number of DFT overpotential calculations necessary to find varying quantities of the set of 50 catalyst active sites with the lowest overpotential using the sequenced classification-regression ML models compared to random sampling. The error bars show the standard deviation of this speedup ratio over 10 trials. The molecule depicted in the upper right is the one with the lowest DFT predicted overpotential, with the active site highlighted in the red circle.*
The results in Figure 12 show relatively similar performance for both methods relative to random up until the sixth site, at which point the structure feature trained method drops significantly in relative performance. This drop occurs because while the structure feature trained classifier predicted these sites are active, the regressor does not correctly predict all these sites will be very active. The regressor trained on all features has a similar behavior which is not visible here because the classifier trained on all features identified 43 active sites, so the drop in performance would be visible only if the total number of sites in the goal approached 43. As seen in Figure 10, the false positive rates of these models are low on the sites where they are most confident; the effect of including the electronic structure features in model training in this scenario is that it expands the confidence range in which the models can make accurate predictions. This behavior highlights the nuance of when additional accuracy is necessary for HTVS: if the goal is to find only a few active catalysts among many, a lower accuracy model that incurs lower computational cost may be completely sufficient as long as the false positive rate is not too high. If the goal is instead to find a more complete fraction of all active catalysts or to have a high likelihood of finding the very best catalysts, choosing a more accurate yet more expensive model may be worthwhile.

4. Conclusion

In this work, we use ML as a tool to provide statistical insight on the properties that cause some carbon atoms in N⁺-GCCs to bind O₂. For the distribution of 258 N⁺-GCCs explored in this work, over 90% contain at least one carbon that DFT predicts to bind O₂, but in turn these active sites comprise only 20% of the carbon atoms in the dataset. There is no single structural feature we found that distinguishes the active sites from the inactive sites, making this a suitable problem for analysis with ML models that can draw inferences from many contributing chemical features. Of the ML methods we evaluate to classify carbons as active or inactive with respect to binding O₂, our most effective method (which used both electronic and structural features of the catalysts) achieves 95.3% accuracy. We also train a regression model to predict O₂ binding energy, which achieves an MAE of 0.077 eV with an \( R^2 \) value of 0.70. Permutation importance indicates that the two most important predictors of O₂-binding activity are the site’s spin density and its change in charge at that site when the catalyst is oxidized. These features are likely important predictors because during the O₂ binding process, electron density moves from the catalyst onto the terminal oxygen atom. Several structural features are also important to a lesser degree; these features broadly reflect the energy required to convert the carbon from sp² to sp³ hybridization.

To demonstrate how these ML models could be used for HTVS, this study explores three scenarios for searching for catalysts and active sites. In each scenario, the strategies that use our ML methods significantly reduce the number of necessary DFT calculations, thus greatly decreasing the computational cost and time required to screen for active catalysts. Of note, despite the loss of accuracy, the models trained only on structural features are effective for HTVS as long as the goal is to only identify a limited number of active sites. Although the structure-only classifier is less accurate, its accuracy suffers more from false negatives than false positives, and thus it remains useful because HTVS suffers more from false positives when the distribution contains sufficient active sites that at least some are still found. In future work, the strategies applied in these scenarios could be used to screen larger molecular libraries for promising GCCs. As the synthetic framework for creating these catalysts works for a range of chemical structures that contain diamines, a good starting point would be collecting a
much larger library of diamines from such chemical libraries as ChemSpider. This library could be converted into virtual GCC structures, then screened by ML model trained on structural features, such as the one in this work, to select for promising candidates that are further screened with models that include DFT features. This process could readily be extended to other catalytic reactions that are suited for GCCs.

Beyond the discovery of novel GCCs, these atom-centric models are broadly relevant for generating insights on classes of catalysts with multiple possible active sites. The framework developed in this work is broadly applicable to heterogeneous catalysts, especially ones with a degree of disorder that creates a distribution of possible active sites. Rather than analyzing only a handful of example active sites, this approach can provide more general quantitative insights for a broader distribution of possible active sites for these materials.

Supporting Information
Section 1 describes the automated process for determining whether the output of DFT O₂ binding calculations successfully bound O₂ to an active site. Section 2 describes each ML model feature. Section 3 describes the performance of a different approach to training ML regression models on the data. The machine learning model data as well as the xyz files for the catalysts in all reaction intermediate states are provided at https://doi.org/10.6084/m9.figshare.21350607.v1.

Acknowledgments
This work was supported by a grant from the ACS Petroleum Research Fund (Grant No. 60503-ND6).

Notes
The authors have no conflicts of interest to disclose.

References


Ricke, N. D.; Murray, A. T.; Shepherd, J. J.; Welborn, M. G.; Fukushima, T.; Van Voorhis, T.; Surendranath, Y. Molecular-Level Insights into Oxygen Reduction Catalysis by
https://doi.org/10.1021/acscatal.7b03086.
Table of Contents Image

Successfully bound (active)  No bond formed (inactive)

$O_2$ binding calculations run for each candidate binding site in training set

Trained models applied to test catalyst set to predict $O_2$ binding with 95% accuracy