Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis

Pushkar G. Ghanekar[1*], Siddharth Deshpande[1*‡], and Jeffrey Greeley[1‡]

[1] Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907, USA

* Equal contribution

‡ Corresponding authors

Abstract

Heterogeneous catalytic reactions are influenced by a subtle interplay of atomic-scale factors, ranging from the catalysts' local morphology to the presence of high adsorbate coverages. Describing such phenomena via computational models requires generation and analysis of a large space of surface atomic configurations. To address this challenge, we present the Adsorbate Chemical Environment-based Graph Convolution Neural Network (ACE-GCN), a screening workflow that can account for atomistic configurations comprising diverse adsorbates, binding locations, coordination environments, and substrate morphologies. Using this workflow, we develop catalyst surface models for two illustrative systems: (i) NO adsorbed on a $Pt_3Sn(111)$ alloy surface, of interest for nitrate electroreduction processes, where high adsorbate coverages combine with the low symmetry of the alloy substrate to produce a large configurational space, and (ii) OH* adsorbed on a stepped Pt(221) facet, of relevance to the Oxygen Reduction Reaction, wherein the presence of irregular crystal surfaces, high adsorbate coverages, and directionally-dependent adsorbate-adsorbate interactions result in the configurational complexity. In both cases, the ACE-GCN model, having trained on a fraction (~10%) of the total DFT-relaxed configurations, successfully ranks the relative stabilities of *unrelaxed* atomic configurations sampled from a large configurational space. This approach is expected to accelerate development of rigorous descriptions of catalyst surfaces under *in-situ* conditions.

Tags: Heterogeneous catalysis, density-functional theory, machine-learning, graph networks

Introduction

Theoretical computational models have become indispensable in elucidating the intricate molecular-level details of heterogeneous catalysts. High-throughput material screening strategies, combined with descriptor-based correlations such as scaling and Brønsted-Evan-Polanyi relationships,[1–4] have played a central role in identifying promising candidates for important oxygen, nitrogen, and carbon-based chemistries. These approaches have been augmented by the recent emergence of improved computational modeling algorithms, some based on machine learning, which have made screening of diverse materials classes, such as oxides, perovskites, zeolites, and metal-organic frameworks (MOFs), possible through the facile generation of diverse materials-specific motifs, [5–10] and accelerated predictions of binding energies of reaction intermediates have further contributed to the descriptor-based catalyst screening paradigm. [6,7,11–15] These computational strategies, which iteratively improve through experience, have enabled the (re)discovery of exciting catalytic materials and chemical insights.

In spite of these advances, it remains challenging to obtain atomic-scale understanding of catalyst properties under realistic reaction conditions, as heterogeneous catalytic reactions are sensitive to the atomic-scale complexities arising from adsorbate-adsorbate interactions at high adsorbate coverages, the local morphology of the catalysts, and variations in the catalysts' surface composition induced by adsorption, among other factors. [16–22] To successfully overcome these difficulties, efficient generation and analysis of atomistic models is critical and requires development of methods that can efficiently sample the large configurational space of surface atomic configurations for diverse catalyst compositions and surface structures. [23,24]

Herein, we present a generalized screening workflow that seeks to address these challenges. The approach involves systematic enumeration of atomic configurations using graph-based representations. [23] The relevant chemical and geometric properties of the generated motifs are learned and mapped to the target property of choice using a machine learning model based on a graph neural network architecture, [25,26] which is termed the Adsorbate Chemical Environment-based Graph Convolution Neural Network (ACE-GCN). ACE-GCN serves as a surrogate model for expensive electronic structure optimization routines and efficiently provides estimates for the target properties of catalyst surfaces, thereby facilitating high throughput evaluation of a large space of complex active site models.

The proposed workflow can systematically describe a variety of atomistic configurations comprised of diverse adsorbates, binding locations, coordination environments, and catalyst morphologies. This flexibility is demonstrated in the context of two catalytic systems that are relevant to practical electrocatalytic applications and that represent the typical complexities encountered when developing computational models of heterogeneous catalysts. The first case treats high coverage configurations of the adsorbate NO* on a $Pt_3Sn(111)$ terrace surface, wherein a vast surface configurational space resulting from both the reduction in the catalyst surface symmetry due to alloying [27–30] and the strong binding nature of NO* yields rich catalytic behavior. This chemistry is of interest in electrocatalytic water treatment strategies, and similar complexities arise in chemistries such as Fischer-Tropsch synthesis and water-gas shift.

[17,31] With our proposed workflow, all high coverage NO* configurations (~3400) are analyzed by performing only a small fraction of explicit DFT calculations (~350). In the second case, the challenge of modeling irregular or defected crystal surfaces, together with strong, directionally-dependent adsorbate-adsorbate interactions, is addressed. High coverage configurations of OH*, known to be stabilized through intermolecular hydrogen bonds (H-bonding), are analyzed on the Pt(221) stepped and Pt(100) square surfaces. These types of interactions can strongly impact the energetics of electrocatalytic reactions such as hydrogen evolution, oxygen reduction, and CO electro-oxidation. [32–35] An approach inspired by transfer learning is employed, wherein explicit DFT calculations of high coverage OH* configurations on Pt(100) terraces (~200) are combined with selected calculations of OH* on Pt(221) (~400). Using the ACE-GCN approach, and subsequently including a modest number of additional high coverage geometries (~ 800) for incremental model improvement, a comprehensive set of high coverage OH* configurations on the Pt(221) surface (~11500) is explored to identify low energy adsorbate structures. This generalized approach shows how multiple datasets may be used to incorporate information from diverse catalyst morphologies to efficiently describe complex, low symmetry surfaces with vast configurational spaces in the ACE-GCN framework. [36–38] Finally, we briefly illustrate the utility of these approaches for determining *in-situ* catalyst structures under realistic reaction conditions by analyzing the state of Pt(221) surface via an ab-initio Pourbaix analysis.

Results and Discussion

As mentioned above, prediction of catalyst structures under realistic reaction conditions requires addressing two primary sources of complexity: (i) the structural intricacies of the catalyst, stemming from variations in compositional and morphological properties, and (ii) adsorbate structures, which may involve multiple adsorbed species and directionally-dependent adsorbate-adsorbate interactions such as hydrogen bonding. Such chemical complexities yield a large phase space of possible atomic configurations, motivating development of a systematic computational framework to screen configurations with less expense than is required by exhaustive first principles analysis.

Workflow and ACE-GCN Framework

Figure 1(A) summarizes the proposed screening framework. The cyclic workflow is divided into four parts: (i) systematic enumeration of unique atomic configurations, (ii) (re)training the surrogate model with data of incremental complexity, (iii) accelerated screening using the surrogate model to identify the most relevant configurations amongst possible geometries, and (iv) electronic structure relaxation of selected structures, which can be used for in-depth mechanistic analysis, or to improve the surrogate model.

First, adsorbate configurations are generated by enumerating adsorbate binding locations on the catalyst surface using the SurfGraph algorithm. [23] This algorithm utilizes graph-based representations to identify and create unique surface adsorbate configurations, systematically accelerating the task of generating complex catalytic model motifs. [23,24] Next, ACE-GCN is utilized as a surrogate model for screening the generated motifs. The algorithm captures the geometric and chemical properties of a given surface adsorbate's local environment and maps them to a target property of choice. In this work, ACE-GCN is initially trained on a small subset of relaxed adsorbate configurations, and then utilized as a surrogate model to systematically rank the energies of a much larger number of *unrelaxed* adsorbate configurations. The approach thus provides a framework to efficiently identify a subset of highly promising candidate structures, as generated by SurfGraph, for subsequent electronic structure relaxation, therefore bypassing the computationally expensive step of DFT-optimizing all possible atomistic configurations. After electronic-structure optimization of the most promising structures, the selected candidate configurations are used to further improve the prediction capabilities of the ACE-GCN model by including them in an expanded training pool, as well as to perform an in-depth analysis of the reaction mechanism. Below, additional descriptions of the ACE-GCN framework, as well as two examples of its application are provided.

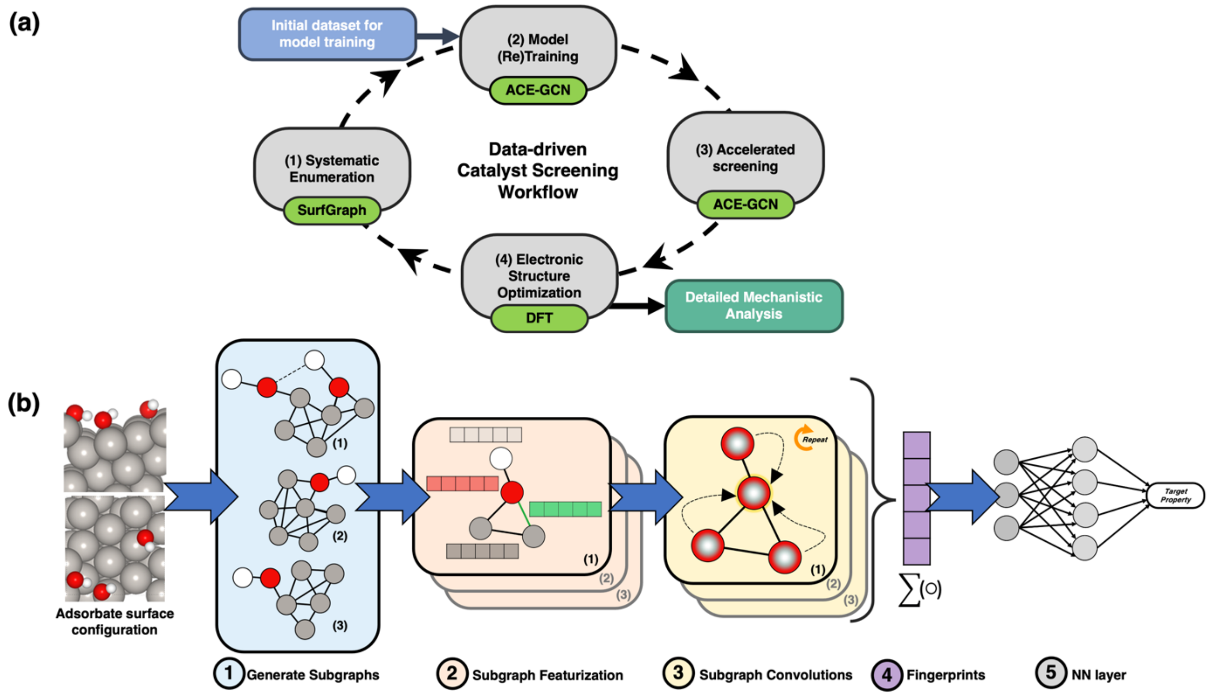
(a)
Initial dataset for model training
(2) Model (Re)Training
ACE-GCN
(1) Systematic Enumeration
SurfGraph
Data-driven Catalyst Screening Workflow
(3) Accelerated screening
ACE-GCN
(4) Electronic Structure Optimization
DFT
Detailed Mechanistic Analysis
(b)
Adsorbate surface configuration
(1)
(2)
(3)
Repeat
Target Property
∑(○)
1 Generate Subgraphs
2 Subgraph Featurization
3 Subgraph Convolutions
4 Fingerprints
5 NN layer

Figure 1: (a) Screening workflow for identifying stable surface adsorbate configurations. The workflow demonstrates an incremental training approach to predict thermodynamically stable catalytic configurations. The cyclic workflow includes the following steps (1) Systematic Enumeration: all possible and unique high coverage surface adsorbate representations are generated using the SurfGraph algorithm, (2) Model Training: ACE-GCN model is (re)trained on selected structures utilizing the relevant surface representations identified in the previous steps. (3) Accelerated Screening: The unrelaxed surface configurations generated in step 1 are ranked using the ACE-GCN model, which is pre-trained on smaller subset of relevant DFT-relaxed cases. (4) Electronic Structure Optimization: selected *unrelaxed* configurations ranked by ACE-GCN are optimized using electronic structure optimization code of choice and then utilized either for subsequent analysis or to re-train and improve the ACE-GCN model.

(b) ACE-GCN algorithm to encode and train high coverage adsorbate configurations. (1) Generate sub-graphs: each configuration is split into multiple subgraphs, as identified by the SurfGraph algorithm. A distinct ego-graph is generated for each adsorbate to encode local geometric and chemical properties around the adsorbate in a subgraph representation. (2) Subgraph Featurization: each atom and its corresponding bond attribute in the subgraph is expressed as a vector representation according to the chemical identity (elemental properties) and spatial bond distance, termed as node and edge features, respectively. (3) Subgraph Convolutions: every node vector in the subgraph is iteratively updated through multiple rounds of graph convolution operations, which account for the atom's geometric and chemical neighborhood using node and edge vectors of the neighboring atoms. (4) Fingerprints: a hierarchical pooling operation condenses all subgraphs for every adsorbate into one fingerprint vector. (5) NN Layer: the fingerprint vector is passed to a feed-forward neural network (NN) which maps it to the target property of choice, such as the average adsorption energy.

Adsorbate chemical environment-based graph neural networks

The ACE-GCN framework is based on a graph neural networks (GNN) architecture. [25,39] Graph-based learning, wherein small molecules or crystals are presented as undirected graphs with atoms described as nodes and edges representing the connections between the atoms, has been used to accurately account for the underlying structural and chemical properties for a diverse class of materials including small molecules, [39] periodic materials, [25,40] metal-organic frameworks, [8] and selected surfaces. [6] However, a successful implementation of such graph-based representations, or any surrogate model framework, for complex surface models incorporating a combination of multiple adsorbates, high-coverage ensembles, and complex surface geometries (steps, kinks, and other defects), remains highly challenging. The ACE-GCN model constitutes a simple strategy for treating these sources of complexity.

The schematic in Figure 1B shows the steps involved in predicting a target property using ACE-GCN. Each adsorbate surface configuration is initially split into subgraphs (Figure 1b(1)), which are in turn undirected 'ego-graphs' centered around a particular adsorbate generated using the SurfGraph algorithm. These subgraphs explicitly account for the local chemical and structural environment of the adsorbate and can accurately represent the complexities arising from the presence of local co-adsorbates, defect sites, and compositional variations, enabling a systematic description of the surface-adsorbate and adsorbate-adsorbate interactions. Next, every node and edge attribute of the subgraph is expanded as a vector representation of the user-defined chemical and geometric features (Figure 1b(2)). To systematically capture the geometric and chemical environment features surrounding every node, the node feature vector for each node in a subgraph is iteratively updated based on the neighboring environment through multiple rounds of graph convolution (message-passing) steps (Figure 1b(3)). Next, hierarchical pooling-like operations are performed to condense multiple arbitrary-sized subgraphs into a fixed-length vector fingerprint (Figure 1b(4)). This strategy allows ACE-GCN to successfully operate on cases containing arbitrary numbers of adsorbates and associated neighbors. Finally, the fingerprint vector is used as an input to a fully-connected neural network to predict the property of interest, such as the average adsorption energy (Figure 1b(5)). Additional information regarding the attributes considered for chemical and geometric encoding, the graph convolution equation, supplemental indexing, and hierarchical pooling operations is provided in the Methods section.

Modeling complex heterogeneous catalytic systems using the ACE-GCN scheme

We consider two representative heterogenous catalytic reactions to illustrate the application of ACE-CGN. First, we analyze the stability of high coverage configurations of NO* ('*' represents an adsorbed moiety) adsorbed on a $Pt_3Sn(111)$ surface, and second, we determine the most energetically favorable high coverage configurations of OH* adsorbed on Pt(221) and Pt(100) surfaces. Below, we briefly describe the features of the ACE-CGN algorithm that are highlighted in each example, and in subsequent sections, we provide details of the results.

The first example demonstrates how the concepts of crystal graph generation and neural network analysis can accelerate analysis of the large configurational spaces arising from the presence of high coverages of adsorbates (in this case, NO*) on multi-elemental alloy surfaces. Both surface and bulk alloying introduce a plethora of surface adsorption sites, thereby decreasing the symmetry of the surface and increasing the number of distinct adsorption configurations. As shown in Figure 2a, even for a single NO* adsorbate, twice as many distinct adsorption configurations exist on $Pt_3Sn(111)$ as on a pure Pt(111) surface. This configurational space increases exponentially as the coverage of surface adsorbates increases (Figure 2(b)(i)). Considering between 1 and 6 NO* molecules, corresponding to surface coverages between 1/12 and 1/2 ML (monolayers), and neglecting active sites that incorporate 'Sn' atoms, there are approximately 3400 unique adsorbate configurations with 2500 configurations for the 5 and 6 NO* cases alone. A recent publication explored this $NO/Pt_3Sn(111)$ phase space using an evolutionary algorithm-based scheme, and the present work leverages this prior experience to test and validate the ACE-GCN workflow. [17,41]

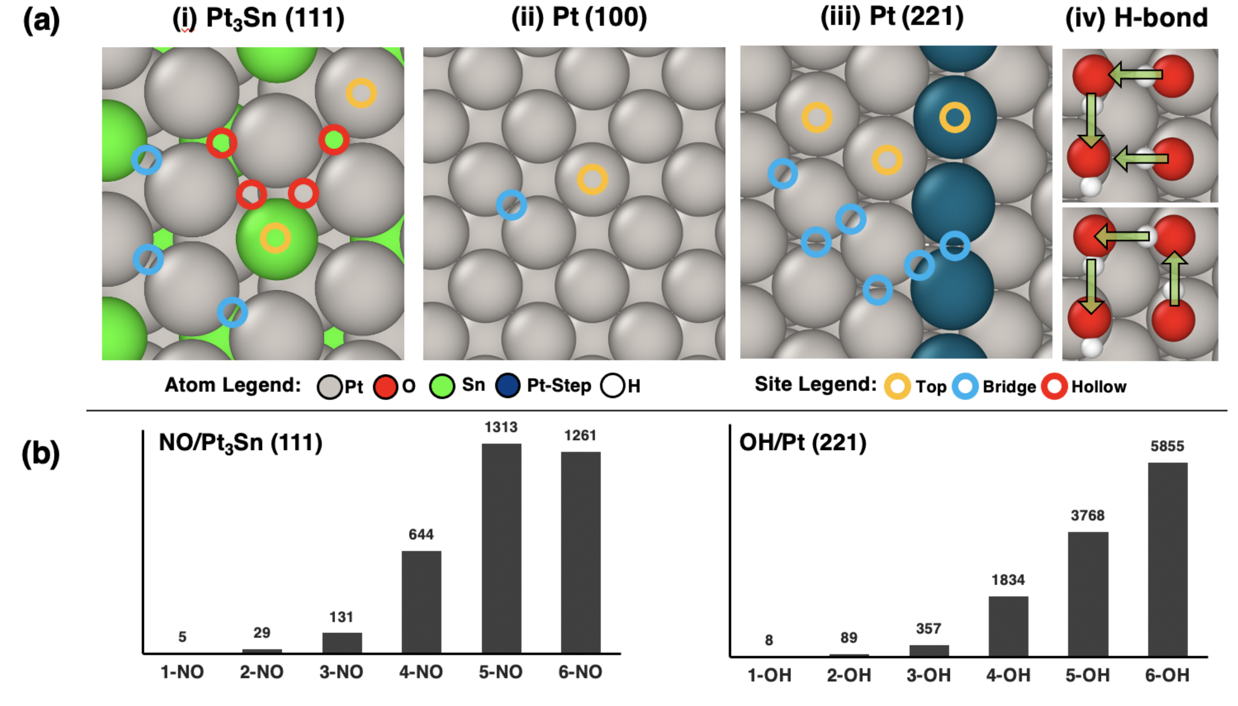

Figure 2: Catalyst configurations analyzed with ACE-CGN. (a) Structural motifs considered in the catalyst models: (i) alloying ($Pt_3Sn(111)$), (ii) diversity of binding sites on Pt(100) and (iii) Pt(221) (terrace in gray, step in blue) surfaces, and (iv) directionally-dependent intermolecular interactions between adsorbates, such as OH*. Green arrows show the direction of H-bonding for each hydroxyl group. (b) The total number of unique surface

configurations, as a function of adsorbate coverage, for $Pt_3Sn(111)$ and Pt(221). All configurations are generated using the SurfGraph algorithm.

The second example demonstrates how high coverage configurations of adsorbates may be enumerated on surfaces with defects such as steps and non-hexagonal geometries. This case, which focuses on OH*, explicitly considers the effect of adsorbate directionality, stemming from intermolecular hydrogen bonding, on the configurational space. Figure 2a(iii) shows a top view of the Pt(221) step surface, which has a three-atom wide terrace resembling the Pt(111) surface. The number of possible OH* configurations on Pt(221) is significantly larger than that on terrace models such as Pt(100) (Figure 2a(ii)) or Pt(111), since each row of Pt atoms in Pt(221) has a unique coordination environment, necessitating separate consideration of adsorption sites on each row of Pt atoms parallel to the step edge. Additionally, for given OH* positions on the surface, several hydrogen bonding networks are possible, and since each may have a very different energy, [42] it is important to explicitly enumerate all such networks (Figure 2a(iv)). Directed graphs, in turn, are an efficient means of incorporating adsorbate directionality into graph-based representations. Initially, all possible O-O pairs that can form hydrogen bonds are determined, following which all unique hydrogen bonded networks amongst the different pairs are estimated (see Methods section for more information). Every hydrogen bond is explicitly encoded as an additional edge attribute in the subgraph generation in ACE-GCN. An illustrative example is presented in Figure 2a(iv), wherein two possible H-bonding configurations for 4-OH* on Pt(221) are shown. Figure 2b(ii), in turn, shows the histogram of the number of configurations as a function of OH* coverage, which were generated by considering both top and bridge sites till 3

OH* (coverage of 1/4 ML), and subsequently for the cases of 4,5,6 OH* (1/3, 5/12 and 1/2 ML respectively), only top sites were added. The total configurations are ~ 12000, while 1834, 3768, and 5855 configurations are found for the 4, 5, and 6 OH* cases (1/3, 5/12 and 1/2 ML coverages), respectively. As described further below, we use ACE-GCN to efficiently probe these complex configurational spaces, and we additionally illustrate how the approach can be used to combine insights from diverse datasets, in a strategy reminiscent of transfer learning, [36–38] by including OH* adsorption on the geometrically distinct Pt(100) surface, to yield improved predictions.

Estimating most relevant high coverage configurations of NO* on a $Pt_3Sn(111)$ alloy catalyst

As shown in Figure 2b(i), the total number of unique initial configurations for 1-6 NO* adsorbed on a $\sqrt{12} \times \sqrt{12}$ $Pt_3Sn(111)$ unit cell (coverage range of 1/12 – 1/2 ML) are on the order of ~ 3400, with roughly 2500 configurations for the 5 and 6 NO* cases (5/12 and ½ ML) alone. The goal of the proposed screening strategy (Figure 1), utilizing ACE-GCN, is to systematically develop a surrogate model, which describes the important interactions governing the stability of low coverage NO* models (1/2/3/4 NO* or 1/12 – 4/12 ML coverage), and to use the resulting insights to efficiently screen the vast number of high coverage configurations (5/6 NO*, 5/12 and ½ ML coverage) with minimal additional computational effort. First, an ACE-GCN model is trained on the average NO* binding energies of all of the low coverage (1, 2, and 3 NO*, 1/12 to 1/4 ML) DFT-relaxed structures (see the detail for the 1-3 NO* model fit in the Supplemental Information S4), and next, the model is used to predict binding energetics for the 4-NO* (1/3 ML) case. Based on these ACE-GCN predictions, 100 energetically stable and 100 unstable candidates (200 total) of the 644 possible 4-NO* configurations, are then selected. These configurations are relaxed using DFT and added to the incremental model training. Figure 3 (a-b) shows the parity plots for the training and validation sets for the new 1/2/3/4-NO* dataset (1/12 to 1/3 ML coverages). The model fits the target property, average NO* binding energy, with a mean absolute error of 0.02 eV for training and validation sets, demonstrating that the ACE-GCN architecture can distinguish amongst different coverages through representations consisting of subgraph-based graph convolutions and hierarchical pooling. Next, the modified ACE-GCN model, trained on the exhaustive 1/2/3-NO* ensemble and some 4-NO* data points, is used to rank the unrelaxed 5-NO* and 6-NO* configurations (5/12 and ½ ML coverages), generated through SurfGraph, as shown in Figure 3 (c). This dataset is comprised of 1314 and 1261 configurations for 5 and 6 NO*, respectively. In the figure, the x-axis represents the ACE-GCN predicted average binding energy of the initial unrelaxed 5/6-NO* configurations, and the y-axis gives the corresponding DFT relaxed energy (for clarity, only those NO* configurations whose binding locations did not change post-DFT relaxation are plotted; additional discussion is provided in the Supplemental Information S4). Importantly, the top 10% lowest energy unrelaxed configurations identified by ACE-GCN include the most stable DFT relaxed atomistic configurations for both the 5 and 6 NO* cases, and, no additional stable configurations were found after DFT relaxation that were not already identified by SurfGraph (see Supplemental Information for additional details). These results, taken together, strongly suggest that the combination of SurfGraph and ACE-GCN is capable of efficiently identifying all stable high coverage configurations for NO* adsorption.

Additionally, the ACE-GCN model captures important information regarding the governing interactions dictating the adsorption geometries of NO* on $Pt_3Sn(111)$. From our recent analysis, [23] it is known that higher coverages of NO* are stable in mixed top and bridge configurations on this surface, while combinations of bridge and threefold sites are unstable. The ACE-GCN model captures this insight, without any explicit user input, using only the low coverage (1/2/3/4 NO*, 1/12 to 1/3 ML) data, and, as described above, efficiently identifies the energetically most stable 5-NO* and 6-NO* (5/12 and ½ ML) configurations. The low energy configurations, in turn (shown in their final configurations post-DFT relaxation in region (i) and (ii) in Figure 3(c)), consist of NO occupying the top and bridge sites on Pt_3Sn. In contrast, higher energy configurations, also shown in region (ii) in Figure 3(c), consist of NO* occupying a mixture of bridge and hollow sites, and are also accurately identified by the ACE-GCN surrogate model. Finally, it is interesting to note that the degree of restructuring of the adsorbate site after DFT relaxation is directly correlated with the stability of a given configuration as predicted using ACE-GCN. The sites predicted to be the most unstable by ACE-GCN underwent the largest change in the adsorbate position after relaxation. Additional details on the model's prediction capabilities as a function of different training data sets, and further discussion on reconstructed NO* configurations, are included in Supplemental Information S4.

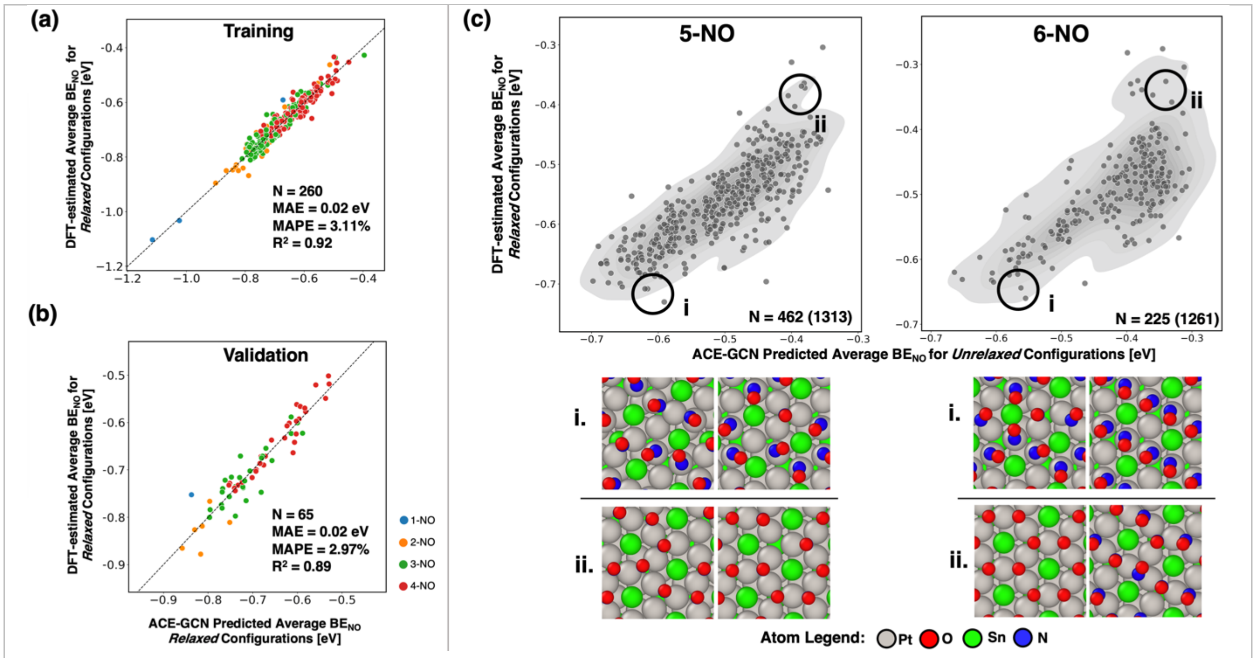

Figure 3: Configurational analysis of NO* adsorption on $Pt_3Sn(111)$, where ACE-GCN is used to predict energetics of the unrelaxed configurations generated using SurfGraph. (a) and (b) correspond to training and validation parity plots for an ACE-GCN model with NO* configurations consisting of 1-4 NO molecules. (c) gives predictions of the ACE-GCN model, trained on configurations of 1-4 NO* molecules, for stability of *unrelaxed* 5 and 6 NO configurations generated with SurfGraph. The predicted average BE of the unrelaxed configurations is plotted on the x-axis, while the final energy of the same configurations after DFT relaxation is plotted on the y-axis. Only configurations where the binding location of the NO* did not change after DFT relaxation are included. The ACE-GCN algorithm successfully predicts the trends in adsorption energies based solely on the unrelaxed configurations generated by SurfGraph. Selected relaxed low and high energy configurations are shown in insets (i) and (ii), respectively.

These results strongly suggest that, through selective incorporation of a small subset of data points of increasingly higher coverages, the ACE-GCN model, trained only on low coverage configurations (1-4 NO*, 1/12-1/3 ML), successfully identifies stable high coverage configurations (5/12-1/2 ML) based solely on the unrelaxed geometries generated from SurfGraph. In comparison to the evolutionary algorithm (EA) scheme used in our previous work, the ACE-GCN model (i) required fewer DFT calculations, 350 versus over 500 data points, compared to the EA, [23] and (ii) independently captured the underlying chemical and geometric intuition affecting the adsorption energetics. This is an important advantage that becomes even more significant for larger chemical spaces, where careful analysis of individual configurations and development of chemical intuition could become infeasible.

Identifying stable high coverage configuration of interacting hydroxyl adsorbates on defected Pt surfaces

This case study illustrates the application of our proposed workflow to adsorbates with directionally-dependent hydrogen bonding on non-hexagonally close-packed single crystal surfaces, Pt(100) and Pt(221). The former is chosen as the simplest possible non-hexagonal surface, while the latter represents model step defects that have been shown to exert a significant influence on electrochemical oxygen reduction rates on polycrystalline Pt electrocatalysts. [33,43] Along with the comprehensive training/testing/extrapolation strategy for the Pt(100) and Pt(221) surfaces, similar to that described for the NO/$Pt_3Sn(111)$ case study, we additionally explore the ability of the ACE-GCN framework to synergistically combine insights from training datasets from these two surface morphologies (the benefit of considering such a mixed training dataset on model prediction is further discussed in the Supplemental Information S5). Such strategies will ultimately be key to understanding adsorption configurations on highly complex catalysts, such as polycrystalline nanoparticles, which encompass a variety of different catalyst morphologies. [44,45]

The overall workflow is summarized here and described in more detail in subsequent paragraphs. First, a comprehensive training dataset, consisting of configurations with between 1 and 5 OH* molecules per 8 Pt atoms on the Pt(100) surface (coverages of between 1/8 and 5/8 ML), is generated, while a second training set of between 1 and 3 OH* adsorbed per 12 Pt atoms on Pt(221) (coverages of 1/12 to

1/4 ML) is also created. Although the coverages considered on the stepped surface are much lower than those analyzed on Pt(100), the total number of training data points are very similar in each case. These datasets, through ACE-GCN, are then combined to efficiently identify low energy adsorption configurations of OH* on Pt(221) at much higher coverages (4-6 OH*/12 Pt, coverages of 1/3-1/2 ML), where the total number of configurations is exponentially larger (Figure 2c) than the number of configurations associated with similar coverages on Pt(100).

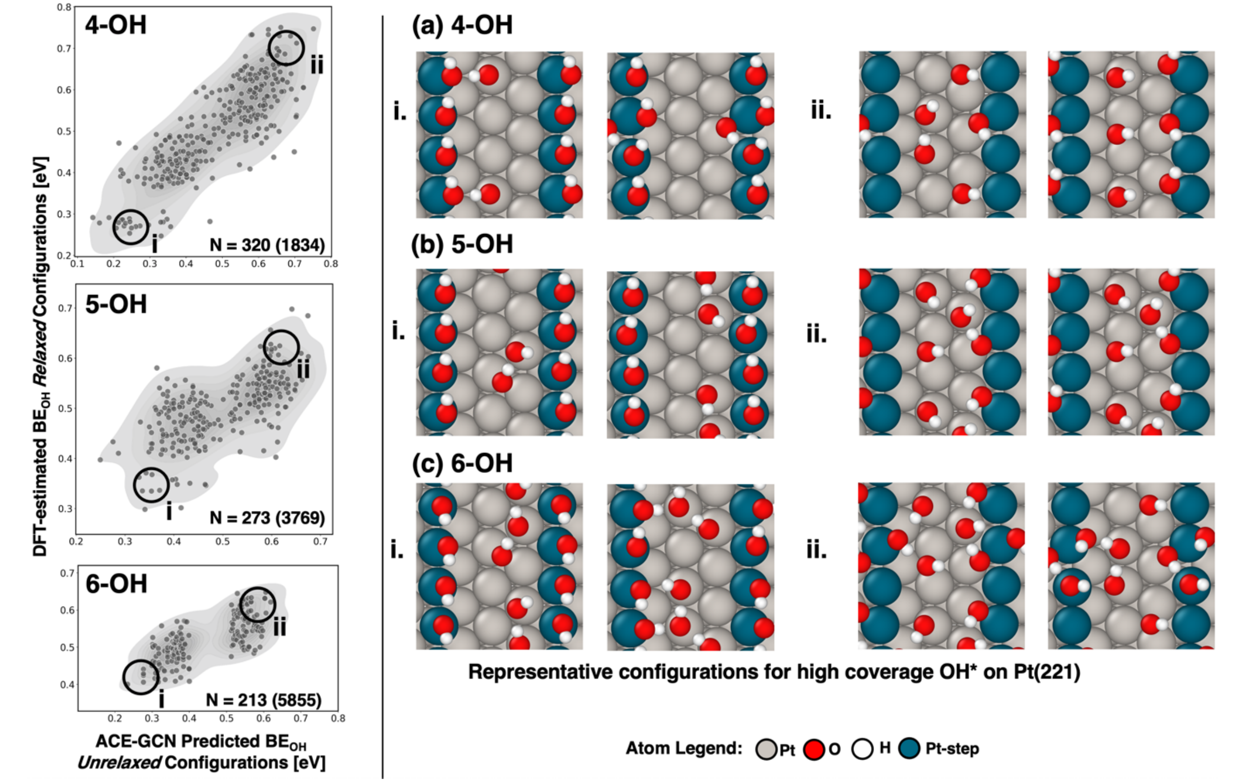

Figure 4: Screening high coverage OH* configurations on Pt(221). Scatter plots showing average OH* binding energies of unrelaxed configurations, as predicted by ACE-GCN (x-axis), with DFT-relaxed energies of the corresponding structures (y-axis). A few relaxed configurations showing OH* species dissociation after DFT relaxation were not included in the plots or model retraining (analysis of dissociated configurations is discussed in the Supplemental Information S5). Numbers in the inset show the total DFT relaxed configurations compared to the total possible structures enumerated by SurfGraph. The ACE-GCN model for each succeeding coverage (4/5/6 OH*, 1/3 to ½ ML) is trained on configurations with lower coverages (see text for details).

The OH* configurations are generated using a modified SurfGraph code that accounts for directional hydrogen bonds among different OH* species (see Figure 2(a) for an example). As mentioned above, the ACE-GCN model is initially trained on the dataset comprising of configurations between 1-3 OH* adsorbates on Pt(221) and 1-5 OH* on Pt(100). Next, the ACE-GCN model is used to rank the unrelaxed 4OH*/ Pt(221) configurations (1834 in total) (1/3 ML coverage), from which 400 configurations, representing a range of energy values and adsorbate binding configurations, are chosen for full DFT relaxation. Figure 4 (top) shows a comparison of the ACE-GCN predicted average binding energies of the unrelaxed 4-OH* configurations and the corresponding DFT relaxed energies. There is a robust correlation between these two quantities, demonstrating that configurations predicted to be low (or high) in energy based on the ACE-GCN predictions of initial unrelaxed geometries track well with post-DFT relaxation results. Shown on the left of the scatter plot are some of the key 4-OH* configurations post DFT-relaxation belonging to the low/high energy 4-OH* arrangements. The most stable structures, represented by region (i) in the plot, have the Pt-step edge (marked in dark blue) completely occupied, and any additional OH* moieties have clustered around the Pt-edge to increase the level of hydrogen bonding. In contrast, the high energy structures, as shown in region (ii), are comprised of separated OH* species, most of which are not directly adsorbed on the Pt step edge, and with relatively few hydrogen bonds. These results indicate that the ACE-GCN model, trained on the diverse data from Pt(100) and Pt(221), accurately learns the underlying features that stabilize the 4-OH* on Pt steps.

Following the scheme laid out in Figure 1(A), higher coverage (5-OH*, 5/12 ML coverage) configurations are generated by using Surfgraph to systematically add an additional OH* moiety to the exhaustive set of unrelaxed 4-OH* configurations. These configurations are then ranked using a retrained ACE-GCN model incorporating the previously DFT-relaxed 4-OH* configurations in the training set. A few of the identified configurations resulted in dissociated OH* species after relaxation, and these cases have not been included in the analysis or model retraining (see Supplemental Information S5). Analogous to the 4-OH* case, a total of 400 unrelaxed configurations, 200 each chosen from high and low energy zones as identified by the ACE-GCN predictions, are selected for DFT relaxation. Finally, a similar strategy is applied when searching for 6-OH* configurations (1/2 ML of coverage), where the emphasis is again placed on high and low energy structures. 3769 and 5855 possible OH* configurations exist for the 5 and 6 OH* cases, respectively, of which only about 400 configurations each for 5 and 6 OH* cases are evaluated using DFT and about 273 and 213 cases remain undissociated post DFT relaxation. The correlation between the stability of structures predicted via ACE-GCN and those after DFT optimization is again quite reasonable (Figure 4); the quasi-bimodal nature of the 5 and 6-OH* plots is simply the result of our choice to sample high and low energy structures, as predicted by ACE-GCN, for DFT optimization. Further, in line with the chemical intuition developed with lower coverages, for both 5 and 6 OH* cases (5/12 and ½ ML coverages), the most stable configurations are comprised of clustered OH* species on the Pt-step edge, whereas unstable cases involve spatially separated OH* with few OH* moieties adsorbed on the step edge. We note, however, that despite the reasonable energetic and chemically intuitive predictions from the ACE-

GCN analysis, there can be non-trivial relaxations of the unrelaxed structures, especially for the high coverage cases of 5 and 6 OH* on the surface (5/12 and ½ ML). We attribute these relaxations to the observation that multiple highly clustered OH* representations may have similar average OH* interaction energies but, may undergo substantially different relaxation during DFT optimization.

The Pt(221) and Pt(100) analyses demonstrate the capability of ACE-GCN to (i) learn important underlying interactions governing the stability of adsorbates with directionally-dependent interactions, such as OH*, on irregular catalyst models by simulating only about 5-6% of the total number of possible configurations, and (ii) combine data having different catalytic morphologies, in a transfer learning-inspired approach, to train surrogate models with high efficiency. Such an analysis can aid in developing chemical intuition regarding the underlying interactions that are crucial for stabilizing the adsorbates and understanding the state of the system in realistic reaction environments.

Mechanistic implications of high OH* coverages for electrochemical reactions on Pt

Based on the identified OH* configurations on the irregular Pt surfaces, a detailed thermodynamic and mechanistic analysis to investigate the state of the catalyst surface under electrochemical reaction conditions, such as those relevant to oxygen reduction reaction (ORR), can now be undertaken. Previous reports have demonstrated that (111) terraces on Pt catalysts are among the most active facets for ORR, and recent investigations on irregular crystal facets of Pt, having variable step sizes ((221), (331) and

(211)), suggest high ORR activity on these surfaces, as well. [33,43,46] A mechanistic analysis incorporating the effects of catalyst morphology and OH* coverages is, in turn, needed to understand these experimentally observed trends. However, the large phase space of possible atomic configurations, especially for the case of stepped catalyst surfaces, makes the analysis challenging.

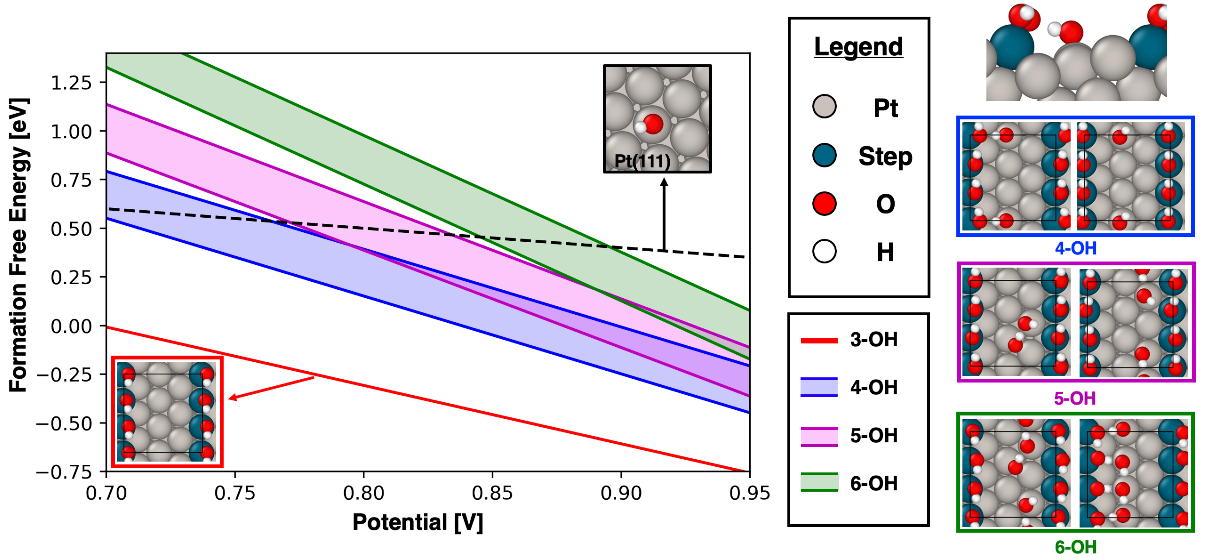

Figure 5: Ab-initio Pourbaix diagram based on binding energies of various OH* configurations on Pt(111) and Pt(211) surfaces. In the potential region of interest for ORR, 0.8-0.9 V, competition between ensembles having 4, 5, and 6 OH* adsorbates on Pt(221) (1/4 to ½ ML OH* coverage) is predicted, and is shown by the overlapping free energy bands. On the rightmost side of the figure, a side view of the most stable 4-OH* configuration in our analysis is shown (the corresponding top view is in the 4-OH inset on the left side). It is observed that the OH* present on the terrace site immediately below the Pt step edge (termed "back-terrace" in our discussion) has a favorable H-bond with the OH* absorbed on the edge. Such an arrangement of OH* moieties is possible due to the particular geometry of the step and edge sites. This arrangement results in appreciable stabilization compared to the scenario where no such H-bond exists (shown in the 4-OH inset, right side). Representative surface configurations for 5 and 6-OH* are also indicated on the insets to the right of the figure, with the most stable configurations on the left of the insets.

Utilizing the results generated in the previous section, an *ab-initio* surface Pourbaix diagram is generated (Figure 5) to explain the state of the Pt(221) surface under ORR-relevant conditions. For simulations reported in Figure 5, larger unit cells along with higher energy cutoffs and k-points are utilized, with additional details reported in the methods section. The formation free energies of the identified high coverage structures

(4-6 OH* on Pt(221)) are plotted as a function of the applied external potential vs. the Standard Hydrogen Electrode (SHE). The formation free energy for each OH* coverage is presented as an energy band, which is 0.25 eV wide, starting from the energy of the most stable configuration identified using the workflow shown in Figure 1. The schematics on the right side of the Pourbaix diagram show the most stable and selected metastable (~ 0.25 eV higher in energy) configurations. In addition, the free energy of the most stable 3 OH* configuration on the Pt(221) facet, together with that of a single OH* moiety on Pt(111), is plotted for reference. The 3 OH* ensemble on Pt(221), where the OH* species occupy the Pt-step edge, is identified as the most stable OH* configuration. This result suggests that the Pt edge might be completely poisoned at ORR-relevant conditions (red inset and line in Figure 5). Additional population of OH* on the surface of the catalyst (4, 5, and 6 OH*) shows competition amongst different configurations, especially above applied potentials of 0.8 V vs. SHE. An interesting feature of the identified high coverage configurations on Pt(221) is the presence of the OH* adsorbed on the terrace sites that lie adjacent to and below the Pt step edge. Such a binding configuration is a result of the unique spatial arrangement of Pt(221) step sites (a representative configuration is shown in Figure 5, right side, top inset). Discovering such a unique OH* binding arrangement, which, to the best of our knowledge, has not yet been reported elsewhere, speaks to the value that data-driven screening workflows such as ACE-GCN can add in helping to identify interesting regions in the chemical phase space, which can then be further explored rigorously to better understand the complex reaction systems.

Furthermore, we observe that multiple possible H-bonding arrangements can possess comparable energies. The most stable OH* arrangements often exhibit hydrogen bonding between the OH* moiety on the lower terrace with the OH* adsorbed on the Pt edge (Figure 5, inset for 4OH* case), or they possess a combination of OH* adsorbed on both bridge and top sites in chain-like structures near the step on the upper terrace (Figure 5, inset for 5 and 6 OH*).

It is important to note that, while the identified structural motifs for high coverage adsorbed OH* may be relevant to practical ORR catalysis, these configurations only consider stabilization due to adsorbate-adsorbate and adsorbate-substrate interactions and do not explicitly account for interactions between adsorbed hydroxyl groups and ambient water solvent molecules, which can have energies on the order of 0.5-0.6 eV per OH*. [42,47,48] To illustrate the effect of such corrections, a black dashed line, representing the OH* adsorption energy on Pt(111), is plotted in Figure 5. At an applied potential of 0.8 V vs SHE, the formation free energy for 1 OH* adsorbed on a top site of Pt(111) is 0.55 eV, excluding any solvent corrections, which is consistent with previous reports. It is only the solvent stabilization that reduces the energy of OH* to near zero on Pt(111) (at 0.8 V vs. SHE) and hence promotes its reactivity. Since the energy of 1-OH* on Pt(111), devoid of any solvent correction, is comparable to the uncorrected energy of the 4/5/6-OH* ensembles on Pt(221), one might expect that some of these ensembles on Pt(221) would be stabilized under ORR condition and contribute to the ORR activity. Further, it is possible that the solvation correction for the high coverage 4/5/6-OH* cases (1/4 to ½ ML of OH*) on Pt(221) could be different compared to the correction for the low coverage OH* ensembles on Pt(111). To fully capture the impact of solvent-adsorbate interactions on ORR chemistry, further analysis, rigorously incorporating explicit solvent molecules (H_2O), along with ab-initio molecular dynamics analysis to understand the electrode-electrolyte double-layer structure, would be necessary. The identified 4/5/6 OH* high coverage configurations provide a strong foundation for undertaking such an analysis, and it is likely that many of the key qualitative conclusions from the analysis, such as the favorable adsorption of OH* on the step edges and the preference for OH* on the lower terrace to interact with the step-adsorbed OH* groups, will not be altered by the presence of additional water molecules.

Conclusions and Outlook

We present a machine learning-based hierarchical screening workflow to systematically estimate active site morphology for complex heterogeneous surface catalytic reactions. The proposed workflow utilizes the graph theory-based SurfGraph algorithm for systematic enumeration and generation of surface adsorbate representations with variable coverages. The generated models are screened using Adsorbate Chemical Environment-based Graph Convolution Neural Network (ACE-GCN), a graph neural network-based framework, which utilizes the chemical and structural environment of a given adsorbate as the input and maps these features to the target property of choice. Using this workflow, we demonstrate the identification of relevant active site models for heterogeneous catalytic systems relevant to strong binding adsorbates on low symmetry alloyed surfaces and to directionally-dependent adsorption on defect surface structures. In both the cases, our model successfully ranks the relative stability of different atomic configurations at a fraction of the computational cost (~10%) of exhaustive DFT calculations, thereby providing a framework to identify relevant atomic configurations for surface environments with large and complex configurational spaces. In addition to reducing the overall computational cost, this automated approach reduces the possibility of systematic bias resulting from use of chemical intuition alone to identify structures with target properties. This approach can therefore serve as a starting point for developing detailed atomic description of complex catalyst surfaces under *in-situ* conditions, help identify interesting regions of the chemical solution space to be investigated with rigorous state-of-the-art methods, ultimately leading to fundamental insights into factors that govern heterogeneous catalysis in structurally and chemically complex environments.

Methods

Dataset

The dataset used for model training and prediction is a collection of a diverse set of calculations corresponding to 1) NO*, varying from 1-6 adsorbates (coverages of 1/12 to ½ ML), on a $Pt_3Sn(111)$ surface, and 2) OH* surface configurations on Pt(100) and Pt(221), also encompassing 1-6 adsorbates (coverages of 1/12 to ½ ML) - see below for unit cell details). The graph enumeration code, SurfGraph, is used to identify the binding sites and to generate the high coverage configurations which are converted to a graph object through ACE-GCN for property prediction. The target property of choice is the binding energy of the adsorbates, normalized to the number of adsorbates considered in the facet:

$$BE_{NO} = \frac{E_{n-NO/Slab} - E_{Slab} - nE_{NO(g)}}{n_{NO}}$$

$$BE_{OH} = \frac{E_{n-OH/Slab} + \frac{n}{2}E_{H_2(g)} - E_{Slab} - nE_{H_2O(g)}}{n_{OH}}$$

DFT methods

The simulations for NO* on $Pt_3Sn(111)$ were adopted from previous publications.[23] For the case of OH* adsorption on Pt(221), the simulations are performed within the framework of periodic density functional theory with the Vienna Ab Initio Simulation Package (VASP) . [49] The energies and geometries of the most stable configurations of OH* on the Pt(221) surface are obtained through minimization of the total energy with respect to geometry by spin polarized generalized gradient approximation calculations (GGA-PBE). [50] The projected augmented wave (PAW) method is used to account for the effect of core electrons on the valence electron density. [51] A PBE-calculated lattice constant of 3.97 Å for pure Pt is employed. The Pt(221) surface is represented by a 3x3 unit cell with 4 layers (total of 33 atoms per unit cell). A vacuum equivalent to 13 Å is applied between any two successive slabs, and surface relaxation is allowed in the top three layers. A planewave energy cutoff of 300 eV is used for the high-throughput calculations. A minimum k-point grid sampling of 3x3x1 is employed. For selected cases reported in the phase diagram in Figure 5, a larger unit cell containing 60 Pt atoms is utilized, and a planewave energy cutoff of 400 eV, along with k-point grid sampling 4x4x1, is employed. It is observed that between the two different kinds of models and simulation parameters utilized, the trends in the adsorption energies of OH* remains the same, with minimal (~ 0.1 eV) change in relative adsorption energies. The electronic occupancies are determined according to a Methfessel– Paxton scheme with an energy smearing of 0.2 eV. Dipole corrections are used in all cases to decouple the electrostatic interactions between the periodically repeated slabs. Structures are fully relaxed until the Hellmann– Feynman forces acting on the atoms are smaller than 0.05 eV/Å. Atomic configrations were visualized using Atomic Simulation Environment (ASE) and Ovito. [52,53]

Adsorbate subgraph generation

Adsorbate subgraphs were generated using the SurfGraph algorithm. [41] Initially, for a given unit cell, a full graph incorporating all the atoms in the cell is generated. Adsorbate nodes are then identified, and a subgraph is generated with each identified adsorbate node as the center. The subgraphs are generated such that they incorporate the information of the surface atoms immediately adjacent to the adsorbate along with other adsorbate atoms interacting with these surface atoms.

Hydrogen bond generation with directed graphs

All hydrogen atoms with a bond distance greater than 1.3 Å and less than 2.1 Å to a given oxygen atom are constituted as hydrogen bonds. To construct combinations of possible pairs of H-bonds between a set of oxygen atoms, all possible hydrogen bonds are initially identified using the rule explained in the previous sentence. Then, all possible directed graphs are generated between the identified pairs, using the rule that each OH adsorbate can only donate one hydrogen bond and accept multiple hydrogen bonds. The directed graph combinations with the maximum number of hydrogen bond pairs are then selected for property prediction or to perform DFT simulations.

Model architecture and implementation

Graph neural networks (GNN), also known as message-passing neural networks, [39,54] have been previously proposed for computer vision, natural language processing, generating molecular fingerprints, predicting crystal bulk properties, and predicting binding energy on surface slab models. The network developed in this work is the extension of the graph convolution neural network (GCN) approach introduced by Xie et. al. [25] The GCN framework is coupled with a sub-graph generation routine to systematically encode complex high coverage surface configurations. The subgraphs capture important features of the high coverage geometries, and at the same time, the versatility of the neural networks provides nonlinear mapping between the chemical fingerprints and the target property. Hence, it is possible to strike a balance between end-to-end feature learning, provided by deep neural networks, and chemical intuition found in ‘hand-engineered’ features.

Each crystal lattice entry is split into smaller network motifs (subgraphs) as per the number of unique adsorbates identified by SurfGraph. Each subgraph is an adsorbate-centered undirected graph (ego-graph) with nodes representing the atoms and edges representing the connection between the neighboring atoms in the lattice. The chemical identity of each node in this subgraph is represented by a feature vector generated based on its elemental identity using a combination of chemical and geometric features. These attributes are encoded as one-hot encoding. The edge connecting two nodes is described by edge attributes based on the spatial pairwise atom distance. This feature can be expressed either as a Gaussian feature expansion, as done in previous implementations, [25] or as one-hot encoding, as implemented in the current version. The reason for using the one-hot encoding expression of the spatial bond distance is to modulate model’s sensitivity to bond fluctuations arising out of structure optimization. A full list of chemical and geometric properties used is provided in the Supplemental Information S2. Next, the bond distance and the one hot encodings are used to create an adjacency matrix for each subgraph. An indexing scheme is generated to account for various neighbors of a given node; each node index is superseded by the adsorbate index based on the number of unique adsorbates in each crystal entry. Likewise, for every node atom and its corresponding neighbors, the atom indices are superseded by a supplemental indexing linking the neighboring atoms to its parent node. This indexing strategy facilitates the subsequent hierarchical pooling operations, enabling the network to account for arbitrary sized subgraphs. A schematic of this pooling operation strategy is provided in the Supplemental Information S2. Model training starts by embedding node attributes in subgraph embeddings. The graph convolution layers iteratively update the node feature vectors by performing convolutions with surrounding nodes in the subgraphs using.

$$Z_{(i,j)_k}^{(t)} = v_j^{(t)} \oplus u_{(i,j)_k} \qquad (1)$$

$$v_i^{(t+1)} = g_{act}\left[\left(\sum_{j,k}^{N(v_i),\ \ E(v_i)} W_c^{(t)} Z_{(i,j)_k}^{t}\right) + W_s^{(t)} v_i^{(t)} + b^{(t)}\right] \qquad (2)$$

$$g_{act}(x) = \ln(1 + e^x) \qquad (3)$$

$$V_G = \frac{1}{N_P} \sum_P^{N_P} V_G^{(P)} \qquad (4)$$

Equation (1) is the new fingerprint vector formed by concatenation of corresponding neighbor and edge features for each node. Equation (2) shows the graph convolution equation used for iterating the node features in each message-passing round. This equation is inspired from work for predicting small molecule and bulk crystal properties. Here, W_x and b are the shared weights and biases for the graph convolution module, while g_{act} is the softplus activation function, a smooth approximation of the ReLU (rectified linear unit).

The hierarchical pooling is implemented using PyTorch scatter module's scatter method. Through this method, elements in the input matrix of known dimensions can be reduced (summed or normalized) by explicitly specifying the indices which have been used for the said reduction. As a result, arbitrarily sized subgraphs are collapsed into a single user-defined n-sized vector fingerprint equivalent to the atom embeddings defined for each atom node at the start. Following the convolution and mean pooling operations, the fingerprint vector is supplied to fully connected layers to capture the mapping of configuration to the target property). The creation of graph objects for the high coverage configurations is parallelized across multiple CPU cores using DASK. [55]

Model training

The network performance is evaluated using three common metrics based on the model's residuals, the mean absolute error (MAE), the root mean-squared error (RMSE), and the mean absolute percentage error (MAPE). A train-validation-test scheme is adopted for choosing the best model for prediction. During the training phase, the data is randomly split into a train-validation-test split where the test set is kept aside for final evaluation. The model weights are iteratively updated by minimizing the loss function (MSE in this case) associated with predicting the target in the training data, and the validation set is scored after each epoch (as per the MAE). The Adam optimizer as implemented in PyTorch is used for the training. After model training for predefined epochs, the model with best validation score is selected for evaluation on the test set. A complete list of hyperparameters is provided in the Supplemental Information S3. Model training and validation was carried on local CPU cores and Tesla P100 GPU cores provided by the Purdue's Research Computing Facility.

Acknowledgements

The authors acknowledge the United States Department of Energy through the Office of Science, Office of Basic Energy Sciences (BES), Chemical, Biological, and Geosciences Division, Data Science Initiative, grant DE-SC0020381. Use of the Center for Nanoscale Materials, a U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences User Facility under Contract No. DE-AC02-06CH11357, and of computational resources from the National Energy Research Scientific Computing Center, is also acknowledged.

References

1. Greeley, J. *et al.* Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. *Nature Chemistry* **1**, 552–556 (2009).
2. Bligaard, T. *et al.* The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis. *Journal of Catalysis* **224**, 206–217 (2004).
3. Nørskov, J. K. *et al.* Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode. *The Journal of Physical Chemistry B* **108**, 17886–17892 (2004).
4. Lansford, J. L., Mironenko, A. V. & Vlachos, D. G. Scaling relationships and theory for vibrational frequencies of adsorbates on transition metal surfaces. *Nature Communications* **8**, 016105 (2017).
5. Jacobs, R., Hwang, J., Shao-Horn, Y. & Morgan, D. Assessing Correlations of Perovskite Catalytic Performance with Electronic Structure Descriptors. *Chemistry of Materials* **31**, 785–797 (2019).
6. Back, S. *et al.* Convolutional Neural Network of Atomic Surface Structures To Predict Binding Energies for High-Throughput Screening of Catalysts. *The Journal of Physical Chemistry Letters* **10**, 4401–4408 (2019).
7. Batchelor, T. A. A. *et al.* High-Entropy Alloys as a Discovery Platform for Electrocatalysis. *Joule* (2019) doi:10.1016/j.joule.2018.12.015.
8. Moosavi, S. M. *et al.* Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications* **11**, 2618–10 (2020).
9. Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat Commun* **11**, 6280 (2020).
10. Saidi, W. A., Shadid, W. & Castelli, I. E. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *Npj Comput Mater* **6**, 36 (2020).
11. Back, S., Tran, K. & Ulissi, Z. W. Discovery of Acid-Stable Oxygen Evolution Catalysts: High-Throughput Computational Screening of Equimolar Bimetallic Oxides. *Acs Appl Mater Inter* **12**, 38256–38265 (2020).
12. Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. *The Journal of Physical Chemistry Letters* **6**, 3528–3533 (2015).
13. Lu, Z., Chen, Z. W. & Singh, C. V. Neural Network-Assisted Development of High-Entropy Alloy Catalysts: Decoupling Ligand and Coordination Effects. *Matter* **3**, 1318–1333 (2020).
14. Lu, Z., Yadav, S. & Singh, C. V. Predicting aggregation energy for single atom bimetallic catalysts on clean and O* adsorbed surfaces through machine learning models. *Catal Sci Technol* **10**, 86–98 (2019).
15. Chowdhury, A. J. *et al.* Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *J Phys Chem C* **122**, 28142–28150 (2018).

16. Ghanekar, P. *et al.* Catalysis at Metal/Oxide Interfaces: Density Functional Theory and Microkinetic Modeling of Water Gas Shift at Pt/MgO Boundaries. *Topics in Catalysis* **63**, 673–687 (2020).

17. Deshpande, S. & Greeley, J. First-Principles Analysis of Coverage, Ensemble, and Solvation Effects on Selectivity Trends in NO Electroreduction on Pt 3Sn Alloys. *ACS Catalysis* 9320–9327 (2020) doi:10.1021/acscatal.0c01380.

18. Bruix, A., Margraf, J. T., Andersen, M. & Reuter, K. First-principles-based multiscale modelling of heterogeneous catalysis. *Nature Catalysis* **9**, 17–12 (2019).

19. Bhandari, S., Rangarajan, S. & Mavrikakis, M. Combining Computational Modeling with Reaction Kinetics Experiments for Elucidating the In SituNature of the Active Site in Catalysis. *Accounts of Chemical Research* **53**, 1893–1904 (2020).

20. Dionigi, F. *et al.* In-situ structure and catalytic mechanism of NiFe and CoFe layered double hydroxides during oxygen evolution. *Nature Communications* **11**, 4347 (2020).

21. Yan, B. *et al.* Surface Restructuring of Nickel Sulfide Generates Optimally Coordinated Active Sites for Oxygen Reduction Catalysis. *Joule* **1**, 600–612 (2017).

22. Lansford, J. L. & Vlachos, D. G. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nat Commun* **11**, 1513 (2020).

23. Deshpande, S., Maxson, T. & Greeley, J. Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis. *npj Computational Materials* **6**, 79 (2020).

24. Boes, J. R., Mamun, O., Winther, K. & Bligaard, T. Graph Theory Approach to High-Throughput Surface Adsorption Structure Generation. *The Journal of Physical Chemistry A* **123**, 2281–2285 (2019).

25. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **120**, 1929 (2018).

26. Battaglia, P. W. *et al.* Relational inductive biases, deep learning, and graph networks. *Arxiv* **cs.LG**, (2018).

27. Cybulskis, V. J. *et al.* Zinc Promotion of Platinum for Catalytic Light Alkane Dehydrogenation: Insights into Geometric and Electronic Effects. *ACS Catalysis* **7**, 4173–4181 (2017).

28. Greeley, J. & Mavrikakis, M. Alloy catalysts designed from first principles. *Nature Materials* **3**, 810–815 (2004).

29. Purdy, S. C. *et al.* Origin of Electronic Modification of Platinum in a Pt 3V Alloy and Its Consequences for Propane Dehydrogenation Catalysis. *ACS Applied Energy Materials* **3**, 1410–1422 (2020).

30. Purdy, S. C. *et al.* Structural trends in the dehydrogenation selectivity of palladium alloys. *Chemical Science* **11**, 5066–5081 (2020).

31. Clayborne, A., Chun, H.-J., Rankin, R. B. & Greeley, J. Elucidation of Pathways for NO Electroreduction on Pt(111) from First Principles. *Angewandte Chemie* **127**, 8373–8376 (2015).

32. Zeng, Z., Chang, K.-C., Kubal, J., Markovic, N. M. & Greeley, J. Stabilization of ultrathin (hydroxy)oxide films on transition metal substrates for electrochemical energy conversion. *Nature Energy* **2**, 17070 (2017).

33. Haid, R. W., Kluge, R. M., Liang, Y. & Bandarenka, A. S. In Situ Quantification of the Local Electrocatalytic Activity via Electrochemical Scanning Tunneling Microscopy. *Small Methods* 2000710 (2020) doi:10.1002/smtd.202000710.

34. McCrum, I. T. & Koper, M. T. M. The role of adsorbed hydroxide in hydrogen evolution reaction kinetics on modified platinum. *Nature Energy* **39**, 163–9 (2020).

35. Wei, J. *et al.* The Dynamic Nature of CO Adlayers on Pt(111) Electrodes. *Angewandte Chemie* **132**, 6241–6245 (2020).

36. Iovanac, N. C. & Savoie, B. M. Improving the generative performance of chemical autoencoders through transfer learning. *Mach Learn Sci Technology* **1**, 045010 (2020).

37. Smith, J. S. *et al.* Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* **10**, 2903 (2019).

38. Hutchinson, M. L. *et al.* Overcoming data scarcity with transfer learning. *Arxiv* (2017).

39. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. in 1263–1272 (PMLR, 2017).

40. Ahmad, Z., Xie, T., Maheshwari, C., Grossman, J. C. & Viswanathan, V. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Central Science* **4**, 996–1006 (2018).

41. Deshpande, S., Maxson, T. & Greeley, J. Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis. *npj Computational Materials* **6**, 4981 (2020).

42. Zeng, Z. & Greeley, J. Characterization of oxygenated species at water/Pt(111) interfaces from DFT energetics and XPS simulations. *Nano Energy* **29**, 369–377 (2016).

43. Pfisterer, J. H. K., Liang, Y., Schneider, O. & Bandarenka, A. S. Direct instrumental identification of catalytically active surface sites. *Nature* **549**, 74–77 (2017).

44. Cheula, R., Soon, A. & Maestri, M. Prediction of morphological changes of catalyst materials under reaction conditions by combined ab initiothermodynamics and microkinetic modelling. *Catal. Sci. Technol.* **54**, 3465 (2018).

45. Müller, A., Comas-Vives, A. & Copéret, C. Shape and Surface Morphology of Copper Nanoparticles under CO 2 Hydrogenation Conditions from First Principles. *J Phys Chem C* **125**, 396–409 (2020).

46. Bandarenka, A. S., Hansen, H. A., Rossmeisl, J. & Stephens, I. E. L. Elucidating the activity of stepped Pt single crystals for oxygen reduction. *Physical Chemistry Chemical Physics* **16**, 13625–13629 (2014).

47. Deshpande, S., Kitchin, J. R. & Viswanathan, V. Quantifying Uncertainty in Activity Volcano Relationships for Oxygen Reduction Reaction. *ACS Catalysis* **6**, 5251–5259 (2016).

48. Heenen, H. H., Gauthier, J. A., Kristoffersen, H. H., Ludwig, T. & Chan, K. Solvation at metal/water interfaces: An ab initiomolecular dynamics benchmark of common computational approaches. *The Journal of Chemical Physics* **152**, 144703 (2020).

49. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initiototal-energy calculations using a plane-wave basis set. *Physical Review B* **54**, 11169–11186 (1996).

50. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77**, 3865–3868 (1996).

51. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **59**, 1758–1775 (1999).

52. Larsen, A. H. *et al.* The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017).

53. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Model Simul Mater Sc* **18**, 015012 (2010).

54. Río, E. G. del, Mortensen, J. J. & Jacobsen, K. W. Local Bayesian optimizer for atomic structures. *Physical Review B* **100**, 104103 (2019).

55. Team, D. D. Dask: Library for dynamic task scheduling. https://dask.org (2016).