Research Article
doi.org/10.1002/elsa.202300028

Chemistry Europe
European Chemical
Societies Publishing

Received: 22 December 2023
Revised: 5 February 2024
Accepted: 25 February 2024

Explainable AI for optimizing oxygen reduction on Pt monolayer core–shell catalysts

Noushin Omidvar | Shih-Han Wang | Yang Huang | Hemanth Somarajan Pillai | Andy Athawale | Siwen Wang | Luke E. K. Achenie | Hongliang Xin

Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

Correspondence
Hongliang Xin, Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA.
Email: hxin@vt.edu

Funding information
National Science Foundation, Grant/Award Number: CBET-2245402; US Department of Energy, Office of Basic Energy Sciences, Grant/Award Number: DE-SC0023323

Abstract

As a subfield of artificial intelligence (AI), machine learning (ML) has emerged as a versatile tool in accelerating catalytic materials discovery because of its ability to find complex patterns in high-dimensional data. While the intricacy of cutting-edge ML models, such as deep learning, makes them powerful, it also renders decision-making processes challenging to explain. Recent advances in explainable AI technologies, which aim to make the inner workings of ML models understandable to humans, have considerably increased our capacity to gain insights from data. In this study, taking the oxygen reduction reaction (ORR) on {111}-oriented Pt monolayer core–shell catalysts as an example, we show how the recently developed theory-infused neural network (TinNet) algorithm enables a rapid search for optimal site motifs with the chemisorption energy of hydroxyl (OH) as a single descriptor, revealing the underlying physical factors that govern the variations in site reactivity. By exploring a broad design space of Pt monolayer core–shell alloys ($\sim 17,000$ candidates) that were generated from $\sim$ 1500 thermodynamically stable bulk structures in existing material databases, we identified novel alloy systems along with previously known catalysts in the goldilocks zone of reactivity properties. SHAP (SHapley Additive exPlanations) analysis reveals the important role of adsorbate resonance energies that originate from sp-band interactions in chemical bonding at metal surfaces. Extracting physical insights into surface reactivity with explainable AI opens up new design pathways for optimizing catalytic performance beyond active sites.

KEYWORDS
d-band theory, electrocatalysis, interpretable deep learning, Newns–Anderson model, oxygen reduction reaction

1 | INTRODUCTION

Machine learning (ML) is increasingly being leveraged as a sophisticated predictive tool in many subdomains of materials science.[1–6] Among these applications, a notable advancement is its ability to accurately estimate the binding strength of reaction intermediates on heterogeneous catalyst surfaces, a capability that is significantly accelerating the discovery and development of novel catalytic materials and offering a transformative approach to

Research Article
doi.org/10.1002/elsa.202300028

catalysis science. With the availability of ever-growing datasets in open-access repositories, such as Catalysis Hub,[7] Computational Materials Repository,[8] ioChem-BD,[9] and Open Catalyst Project,[10] deep learning has demonstrated a dramatically superior performance over traditional ML models with hand-crafted feature descriptors. These data-driven models, though, have been viewed as black-box nonlinear function estimators, without explanations for the prediction unlike their classical equivalents, such as linear regression and its variants. The concerns on lack of interpretability in the implementation of deep learning have lately been highlighted in the catalysis field.[11,12] This concern is especially relevant in predicting surface reactivity because of the intrinsic complexity of chemical bonding at active sites; consequently, physical factors governing chemisorption remain hidden within the black-box ML models. Uncovering these factors is undoubtedly important to advance catalysis science but remains notoriously difficult. On the other hand, physics-based models, for example, the *d*-band theory of chemisorption in heterogeneous catalysis,[13] directly provides valuable insights into governing factors of catalytic outcomes, for example, the filling, center, and higher-order moments of the *d*-states distribution projected onto site atoms.[14] However, the use of physical models to catalytic sites with strongly perturbed properties is hampered by the lack of meaningful model parameters and its limited accuracy.[15]

Explainable AI aims to make AI systems more transparent by providing explanations for their predictions.[16] One approach to creating more explainable AI is incorporating physical principles into ML models. Physics-inspired ML involves encoding physical knowledge and constraints into the structure and learning process of models, providing both the accuracy and explanatory insights. In this endeavor, a theory-infused neural networks (TinNet) approach was developed with chemisorption processes at surfaces as an example.[17] TinNet integrates physics-based and data-driven modules within a graph neural network (GNN) architecture,[18] which combines domain theories with the flexibility of neural networks to model complex systems. The approach can be used to elucidate the fundamental mechanisms of chemical bonding at surfaces while ensuring accurate predictions of site reactivity. TinNet plays a pivotal role by extracting crucial electronic factors, akin to the genetic information in biology that defines a catalyst's unique characteristics. The comprehension of these key elements in catalysis opens new pathways to rationally manipulate and design catalytic functions. Central to this approach is the application of model-agnostic, explainable AI, with techniques such as SHapley Additive exPlanations (SHAP) being particularly valuable. These methods help unravel the complex electronic factors influencing chemisorption, providing deep insights into the behavior and interactions of a catalyst with reactive species at the orbital level.

Herein, we focus on the oxygen reduction reaction (ORR) in proton-exchange membrane (PEM) fuel cells as a prime example for the application of our approach. Initially, we showcase the effectiveness of TinNet in the discovery of catalysts for ORR. TinNet is employed to swiftly predict OH-binding energies, serving as a critical descriptor for ORR activity on metal catalysts, in conjunction with a high-throughput active learning strategy. The integration of physics-informed ML with quantum-chemical simulations paves the way for the expedited discovery of efficient catalysts. By going beyond the black-box predictions, we identify the dominant factors influencing the ORR. To achieve this, we utilize two types of model explanation methods: local instance-level explanations and global-level explanations. For global explanations, we apply Shapley values to quantify the contribution of electronic factors to bonding strength, thereby highlighting key descriptors that influence reactivity trends. For local explanations, we develop justifications for individual alloy surfaces to offer deep chemical insights into site perturbation relative to Pt. By merging global visualization techniques with local structural explanations, our approach enables a thorough extraction of knowledge from sophisticated deep learning models. This comprehensive method enhances our understanding of the governing factors in ORR and guides the development of more efficient catalysts in PEM fuel cells.

2 | METHODS

2.1 | DFT methods

Quantum ESPRESSO was used to conduct spin-polarized DFT calculations on *OH adsorption systems with ultrasoft pseudopotentials. With revised Perdew-Burke-Ernzerhof,[19] the exchange-correlation was approximately calculated using the generalized gradient approximation. The Pt_{ML} alloys were simulated using (2×2) supercells with six layers and a 15-Å vacuum in between. The top three layers and adsorbates were let to relax till a force threshold of 0.1 eV/Å, while the bottom three layers were fixed. The energy cut-off for plane waves was 500 eV. For molecules and radicals, just the Gamma point was employed, while the Brillouin zone was sampled using a Monkhorst-Pack mesh of $6 \times 6 \times 1$. With a smearing parameter of 0.1 eV for adsorbate systems and 0.001 eV for molecules, the Methfessel-Paxton smearing strategy was used. The projected atomic and molecular density of states were produced by projecting the complete system's eigenvectors onto the ones of the portion, as estimated by gas-phase calculations, at a denser k-point sampling ($12 \times 12 \times 1$) with an energy spacing of 0.01 eV.

2.2 | Theory-infused neural networks

GNN-based models were used as the ML framework to predict formation energies of *OH on Pt_{ML} alloy surfaces. The GNN architecture consists of a convolutional neural network for feature extraction from graph representations of adsorbate–substrate systems, and a fully connected neural network mapping these features to the target property.

To enable model interpretability, the GNNs were integrated with a theory module based on the Newns-Anderson model Hamiltonians in a TinNet architecture.[17] The theory module imposes scientific constraints during training to produce physics-informed predictions.

Hyperparameter optimization of the three key GNN hyperparameters (layer numbers, neuron counts, learning rate) was performed using Bayesian optimization with the Ray Tune library. A 10-fold cross-validation method was employed, wherein models were trained on 81% of data, validated on 9% to enable early stopping, and tested on the remaining 10%. The hyperparameters with the lowest average validation loss were selected.

Rigorous nested 10-fold cross-validation was then conducted, generating 100 total models. 90 models were used to evaluate in-sample accuracy. The other 10 models were applied to an alloy test set to demonstrate out-of-sample performance and suggest candidates for future DFT calculations.

Model optimization was performed using the AdamW algorithm on mini-batches of 64 examples. The multiobjective loss function contained mean squared error terms for adsorption energies as well as d-band moments and projected density of states onto 3σ, 1π, and $4\sigma^*$ orbitals from the theory module. This ties model predictions directly to the underlying theory of chemisorption.

2.3 | SHAP analysis

To explore how much each of the electronic factors and interaction parameters is responsible for the change of OH adsorption energies upon alloying, we leveraged the SHAP,[20–22] an additive feature attribution method for post hoc analysis. SHAP provides a model-agnostic approach based on cooperative game theory to explain the output of ML models. Specifically, SHAP values indicate the contribution of each feature to a particular prediction, relative to a baseline. Being developed based on the Shapley values, SHAP takes each feature (i) as a player in the game and assigns an importance value, ϕ_i, to them which indicates their fair contribution to the output of the model. The SHAP simplifies the original model into an explanation model, g:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i, \tag{1}$$

where M is the number of simplified input features, $x'_i \in 0,1$ is the coalition vector, ϕ_i is the feature attribution for feature i, Shapley values, and ϕ_0 represents the model output with all the simplified inputs being 0. For the instance of interest, the explanation model uses simplified inputs x' that map to the original inputs through a mapping function $x = h_x(x')$. The function maps an entry of 0 to replace the feature value with a base value and maps 1 to keep the feature value as it is.

We have estimated the Shapley values by exact explainers as implemented in the SHAP library.[22] As Shapley values are measures of the contribution of each feature in the change of chemisorption energy, a positive (negative) sign for these values refers to the contribution in weakening (strengthening) of the OH adsorption to a Pt site. The summation of all Shapley values is the total estimated change in the chemisorption energy relative to the pure Pt. Analyzing and comparing the SHAP values for different sites provides detailed, local interpretations for the variations in reactivity and breaks down the contributing roles of specific electronic structure changes.

3 | RESULTS AND DISCUSSION

3.1 | OH chemisorption as a reactivity descriptor for ORR

The sluggish kinetics of the ORR at the cathode is a major source of voltage loss in PEM fuel cells, limiting their widespread adoption.[23–25] Even the state-of-the-art Pt nanoparticle catalyst exhibits ORR overpotentials of ∼ 300 mV at practical current densities.[25]Oxygen reduction is generally accepted to follow an associative mechanism on Pt-based catalysts at fuel cell operating potentials, whereby O_2 adsorbs through a proton-coupled electron transfer to form *OOH, followed by electrochemical reduction to *O and *OH and ultimately formation of H_2O.[26–31] This mechanism is supported by density functional theory (DFT) calculations of the free energy diagram on Pt(111), which shows that all elementary steps become exergonic at $\sim 0.8V_{RHE}$ with a theoretical overpotential of -0.43 V (Figure 1a).[23] However, recent studies propose a more nuanced understanding of these mechanisms.

Emerging research, including that by Keith and Jacob, introduces additional intermediates and pathways,

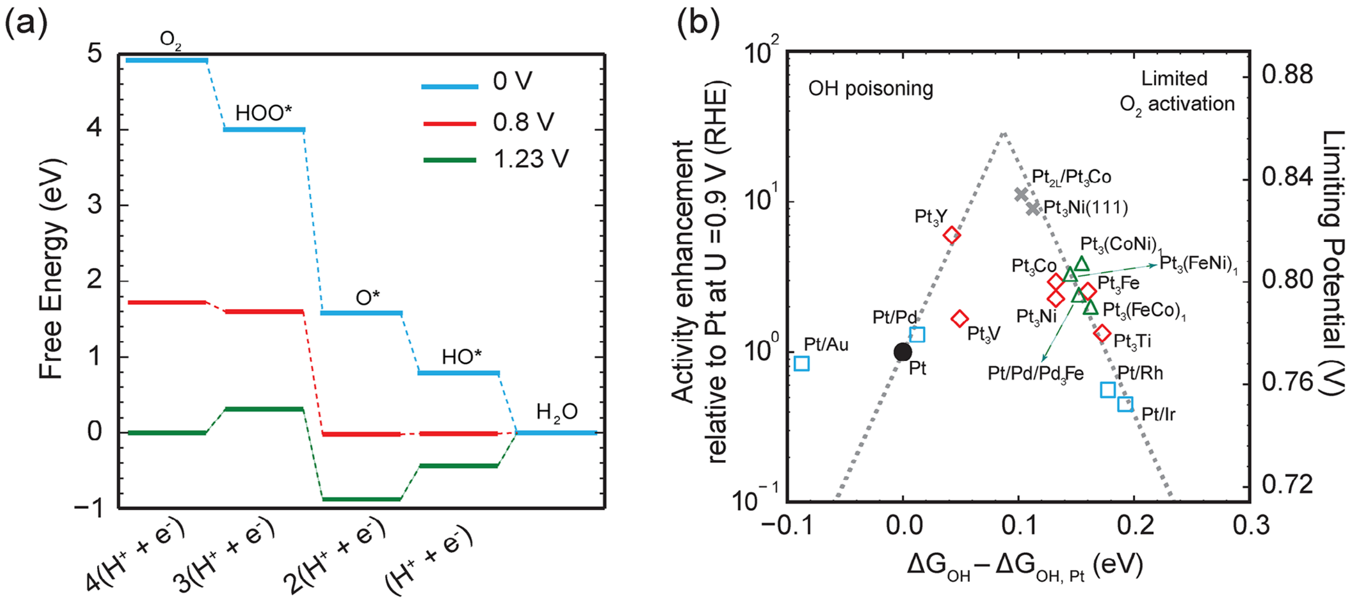

FIGURE 1 (a) DFT-calculated free energy of ORR on Pt(111) at relevant potentials. The computational reversible hydrogen electrode (RHE) is used for relating the free energy of a (H$^+$ + e$^-$) pair at any pH value to that of H_2(g, 1.0 bar) at 300 K. The zero-point energy and entropic contributions to the free energy of an adsorbed species were taken into account.[23,32] The solvation energy of adsorbed species (*O: −0.04 eV, *OH: −0.6 eV, *OOH: −0.3 eV) is assumed to be independent of metal surfaces.[23] (b) The ORR rate enhancement (left axis) and limiting potential (right axis) as a function of DFT-calculated OH binding energies. The dashed lines reflect the theoretical trend assuming a basic kinetic model with rate limited by either O_2 activation or OH elimination from the surface. The experimentally measured rate enhancement of various (111) surfaces: Pt_{ML} on pure metal (□), binary alloy (⋄, x), and ternary alloy (△) surfaces shows a volcano-shaped dependency. The rates were obtained from the literature,[33–35] which were measured at 0.9 V with respect to the RHE in acidic conditions.

underscoring that the ORR on Pt may involve a variety of mechanisms beyond the direct *OOH → *O → *OH sequence.[36] Furthermore, evidence from Exner suggests that the oxygen reduction volcano plot likely represents overlapping free energy landscapes of multiple pathways, rather than a singular, uniform mechanism.[37] This indicates that while associative steps similar to those we have considered may dominate on more strongly binding metals (left of the volcano peak), significant branching to new intermediates likely occurs in the weaker binding regime (right of the optimal activity).

This complex interplay of mechanisms across the volcano plot suggests that the governing steps driving catalysis vary considerably. The possibility that cooperative effects between competing routes near the apex might enhance turnover rates of top catalysts, rather than a single dominant mechanism, introduces a new layer of complexity to our understanding of ORR landscapes.[38]

While we continue to use the *OOH/*O/*OH intermediate scaling as a reasonable approximation for estimating ORR catalysis from *OH-binding energies, we acknowledge the limitations of this approach in fully capturing the intricacies of ORR activity. The field is progressively uncovering the rich and varied landscape of ORR pathways, underscoring the importance of considering these dynamics in interpreting catalytic activity trends.

Late transition and noble metals for oxygen reduction illustrate a strong linear scaling constraint of the free formation energies of the *OH and *OOH intermediates in an aqueous environment, $\Delta G_{*OOH} = \Delta G_{*OH} + \sim 3.2$ eV.[39–43] Therefore, the free energy of *OH formation can serve as a representative descriptor for ORR activity on metal surfaces.[44–48] Figure 1b depicts the theoretical activity volcano along with experimentally measured ORR activity of several known surfaces as a function of the DFT-calculated *OH free formation energies relative to H_2 O(g, 0.035 bar) and H_2(g, 1.0 bar) at 300 K. The volcano peak represents the optimal region of *OH-binding energies to balance O_2 activation to *OOH against surface poisoning by *OH. Alloy catalysts that bind *OH ∼ 0.1 eV weaker than Pt(111) are predicted to be highly active. Overall, this illustrates how the theoretical scaling relations and activity descriptor(s) provide fundamental insights into ORR kinetics and guide the design of improved ORR electrocatalysts.

3.2 | High-throughput screening of Pt_{ML} core–shell catalysts for ORR

Intuitively, the binding energy of *OH can be tuned by controlling the lattice strain (the bond distances of an active site with neighboring atoms) and the metal ligand (the nature of atoms surrounding a catalytic center).[49,50] Pt monolayer (Pt_{ML}) core–shell electrocatalysts have shown promise for enhancing the ORR due to their ability to tune both strain and ligand effects.[51–57] By modifying the core composition and structure, the ORR performance and stability of the Pt_{ML} shell can be optimized. However, rationally modifying the alloying configurations presents challenges, as the complex interplay of strain and ligand effects in site reactivity is unknown a priori.[49,50,58–60]

Extensive work has investigated first- and second-generation Pt_{ML} core–shell alloys to optimize Pt surface activity while reducing Pt content. First-generation alloys have a Pt_{ML} pseudomorphically deposited on a core metal or alloy, while second generation contains a buffer layer beneath the Pt. The vast possibilities for metal selections, compositions, and configurations can theoretically lead to an infinitely large design space for Pt_{ML} electrocatalysts. To efficiently navigate this vast materials space, we generated a database of first- and second-generation Pt_{ML} core–shell alloys by systematically enumerating different combinations of 26 transition metals in AB and A_3 B type structures with face-centered cubic (FCC), body-centered cubic (BCC), hexagonal closed packing (HCP), and simple tetragonal (ST) phases, based on the Materials Project database.[61] Environmentally unfriendly elements were excluded, for example, lead (Pb), resulting in 1518 bulk alloy structures. These were used to construct ~17,000 nearly close-packed Pt_{ML} slab models with 29,040 distinct surface sites.

Although high-throughput DFT screening has seen some successes in catalytic materials design,[62] its extensive cost poses significant challenges. To address this, we adopt an activity model that uses *OH-binding energy as a descriptor for the ORR to quickly screen potential Pt_{ML} catalysts. Concurrently, the TinNet model to be discussed in the next section is employed for efficient prediction of *OH-binding energies, thereby bypassing the need for exhaustive DFT calculations. Figure 2 illustrates our refined high-throughput screening process, which is designed to expedite the discovery of optimal Pt_{ML} electrocatalysts.

When investigating an alloy material for catalytic purposes, the initial consideration is its synthesizability, which can be partly determined by its bulk thermodynamic stability. Therefore, the first phase of our high-throughput strategy involves selecting structures with core alloys that exhibit a high degree of thermodynamic stability. The potential for two metals to form a stable intermetallic compound is often gauged by the alloy's formation energy, denoted as ΔE_f. However, the formation energy alone does not fully capture the thermodynamic stability of a material. An alternative metric, the convex hull distance, defined as the decomposition energy of a phase into its most stable constituents (ΔE_{hull}), also serves as a vital measure for assessing the stability of bulk alloys. In our analysis, we used both the formation energy criterion ($\Delta E_f \leq 0$ eV) and the convex hull distance criterion ($\Delta E_{hull} \leq 0.1$ eV/atom) to evaluate materials stability. The hull distance threshold aims to \retaining potentially synthesizable candidates while screening extremely unlikely structures, considering errors in computational methods.[63,64] Our thermodynamic stability analysis revealed that, within our initial design space, a total of 830 bulk systems, 8610 slabs, and 15,960 unique sites meet the criteria for thermodynamic stability, making them suitable as candidates for further screening.

Descriptor-based analysis facilitates the evaluation of alloy catalysts for enhanced catalytic activity. Adhering to the established criteria for an active ORR catalyst, specifically ($0 < \Delta E_{\mathrm{OH}} - \Delta E_{\mathrm{OH,\,Pt}} < 0.15$) eV, we advance to the next level of high-throughput screening for Pt_{ML} catalysts. In this stage, we estimate the OH adsorption energies of the candidate catalysts within the thermodynamically stable design space. These estimations are performed using a TinNet model that has been trained with DFT data (see the Methods section).[17] TinNet was trained using an active learning framework that minimizes reliance on extensive data. Instead, it focuses on progressively constructing the minimal dataset required for effective candidate exploration. This approach incorporates uncertainty estimates derived from nested 10-fold cross-validation into the criteria for data acquisition. The active learning process is concisely summarized in Table 1.Figure 3 presents a statistical analysis of the model's accuracy and effectiveness within the active learning framework. Through iterative learning of the structure–reactivity relationships among the candidate materials, the uncertainty in predictions was progressively reduced to below 0.1 eV. To assess the model's precision, we calculated the mean absolute error (MAE) of the predictions and their standard deviation using a nested 10-fold cross-validation method. The model demonstrated a high level of accuracy, achieving an MAE of 0.05 eV with a standard deviation of 0.04 eV.

Using the active learning scheme, TinNet models rapidly identified alloy systems with the *OH adsorption energy being 0–0.15 eV weaker than that on Pt(111), which is the ideal range for optimal activity. The models confirmed known electrocatalysts for oxygen reduction that have lower overpotential than pure Pt, such as Pt_3 $B@Pt_{ML}$ (B: Co, Ni, and Ti),[65,66] $Pd_3Fe - Pd@Pt_{ML}$.[53] Additionally, the model uncovered several new first-generation Pt_{ML} core–shell alloys nanostructures using earth-abundant metals, like $Co_3B@Pt_{ML}$ (B: Ni, V, Ti, Mo), $MnB@Pt_{ML}$ (B: Sc and Ni), and $FeB@Pt_{ML}$ (B: Ti, V, Ir),

Research Article
doi.org/10.1002/elsa.202300028

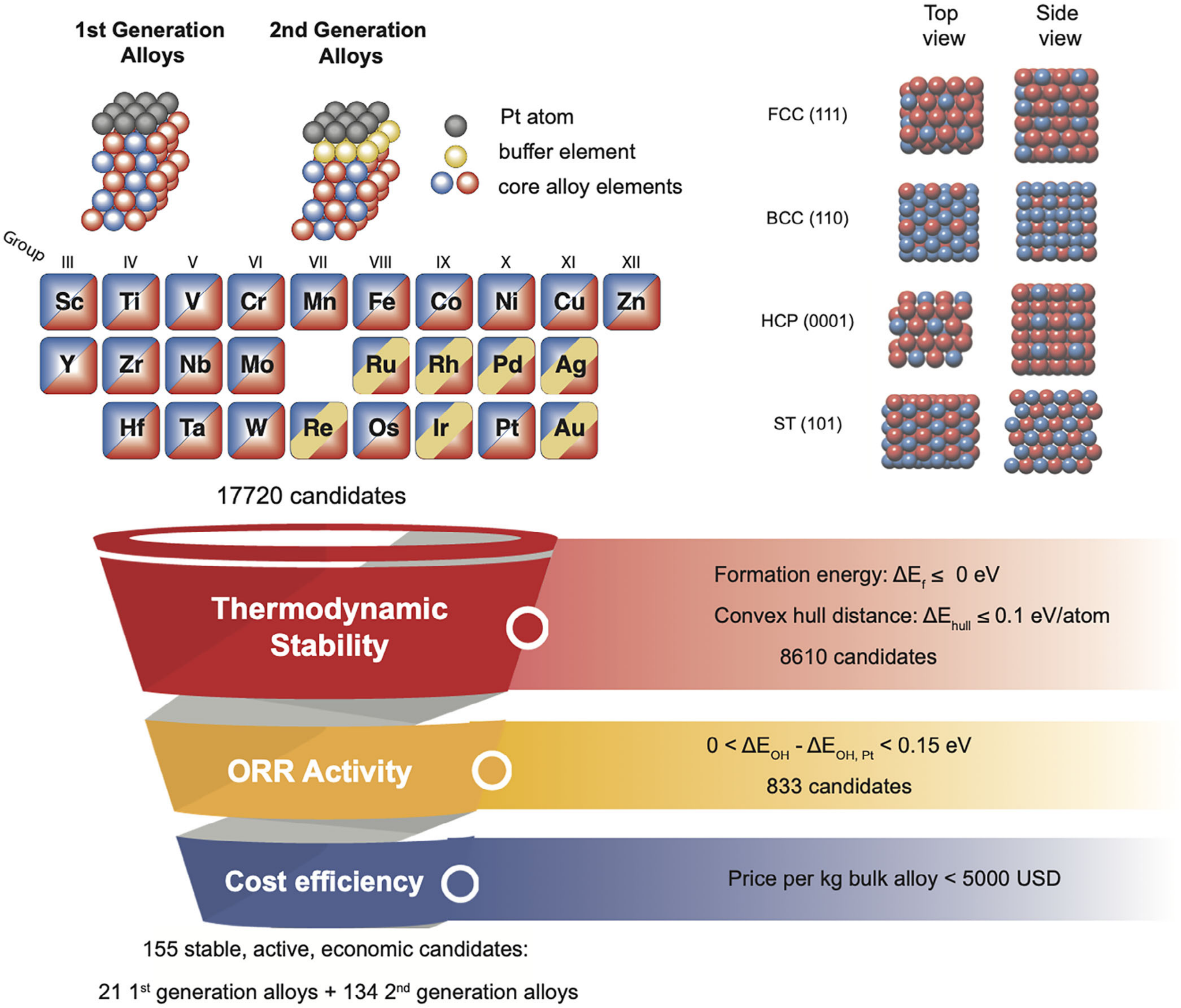

FIGURE 2 The design space for ORR catalysts. First- and second-generation Pt_{ML} core–shell alloys were generated by systematically enumerating different combinations of 26 transition metals in AB- and A_3 B-type structures with face-centered cubic (FCC), body-centered cubic (BCC), hexagonal closed packed (HCP), and simple tetragonal (ST) phases, based on the Materials Project database. ORR, oxygen reduction reaction.[61].

which exhibit near-optimal *OH adsorption energies and are more cost-effective.

To narrow down the selected active and stable candidates into candidates with low-cost core bulk alloys, the prices of the selected core bulk alloys are estimated by the price of the two metals that make up them. This assessment is predicated on the assumption that the cost of the catalyst is independent of the process used to prepare it and is instead determined only by the cost of pure metals. The cut-off value was set at \$5000 kg^{-1} to get rid of alloys that contain precious metals. The final low-cost candidates include five AB_3-type core bulk alloys and 12 AB-type core bulk alloys. Most of the predicted stable, active, and low-cost alloys contain earth-abundant elements such as Fe, Co, and Ni in their core alloy composition. The full catalyst screening results from our high-throughput framework are listed in Table S1 in the Supporting Information.

Figure 3c displays the mapping of some notable known and new core–shell Pt_{ML} nanostructures onto the ORR activity volcano. The TinNet models also predict that the ligand effect from a buffer metal leads to shifts in OH chemisorption energies in the second-generation alloys compared to first-generation counterparts. Specifically, our model suggests that adding Ru as a buffer layer in state-of-the-art $Pt_3Co@Pt_{ML}$ and $Pt_3Ni@Pt_{ML}$ structures can propel them to the top of the ORR activity volcano.

TABLE 1 An active learning algorithm for training TinNet models.

Step	Description
1	**Initialization**:
	- Generate initial dataset D_0 by randomly sampling candidate structures and computing DFT energies.
	- Train initial TinNet model f_θ on D_0.
	- Set $\epsilon \leftarrow \epsilon_0$, iteration $t \leftarrow 1$.
2	**Active Learning Loop**:
	While $t \leq T$:
	- Use current f_θ to predict energies $\hat{y}_i$ and uncertainties σ_i for remaining candidates.
	- **For** $i = 1$ **to** n_b:
	- Sample random $r \sim \text{Uniform}(0, 1)$.
	- **If** $r < \epsilon$: Sample a random candidate.
	- **Else**: Sample a high uncertainty candidate.
	- Add the sampled candidate to batch B_t.
	- Compute DFT energies for batch B_t.
	- Update dataset $D_t = D_{t-1} \cup B_t$.
	- Retrain f_θ on D_t.
	- Decay $\epsilon = \epsilon_0 \cdot \gamma^t, \gamma < 1$.
	- $t \leftarrow t + 1$.
3	**Termination Check**:
	- Compute the ratio r of reliable predictions.
	- **While** $r < 0.98$: Continue the active learning loop.
4	**Return** the final trained model f_θ.

3.3 | Development of TinNet models of surface chemisorption of OH

The application of TinNet for estimating OH adsorption energies offers distinct benefits compared to conventional neural network models, particularly due to its interpretable architecture. Specifically, TinNet integrates a GNN for quantitative prediction with a physics-based theory module for interpretability. The theory module leverages the Newns–Anderson model Hamiltonians to characterize the adsorption process through the *d*-band theory. In this framework, the interaction between the adsorbate and the transition metal is described in two steps. Initially, the electronic states of the adsorbate broaden into resonances and experience a downward energy shift as they interact with the metal's broad *sp*-band near the surface.

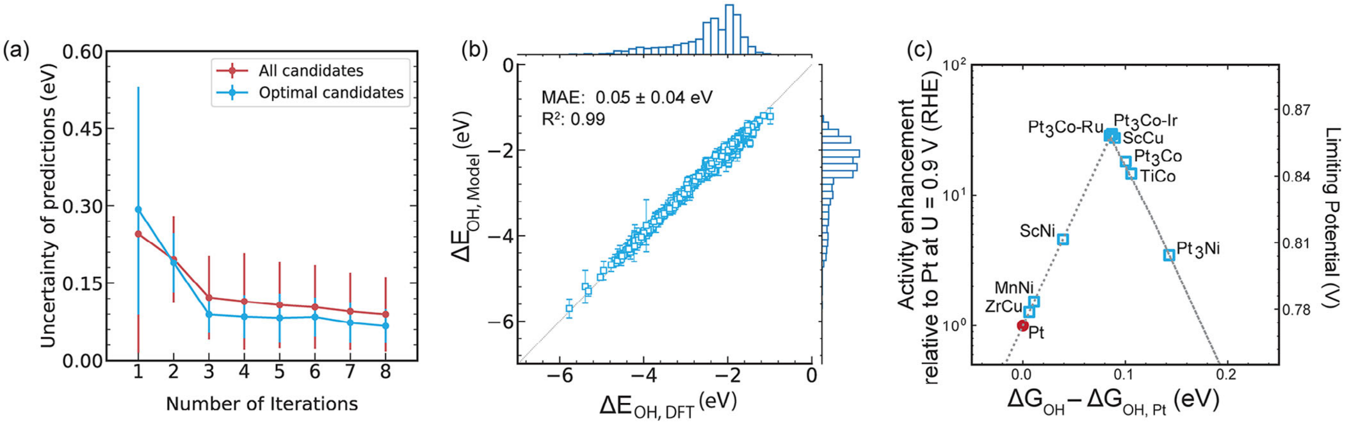

FIGURE 3 Model performance: (a) Evolution of prediction uncertainty across iterations, (b) parity plots comparing DFT-calculated and TinNet-predicted adsorption free energies on Pt_{ML} surfaces for *OH, and (c) Representation of identified optimal structures on the activity volcano plot. Pt_{ML} on bimetallic metal alloy surfaces is represented by the core alloy composition. MAE, mean absolute error; RHE, reversible hydrogen electrode.

Subsequently, a further interaction occurs between the d-states of metals and renormalized adsorbate states which leads to upshifts in energy levels of adsorbate orbitals (orthogonalization) and splitting of the adsorbate density of states into bonding and antibonding orbitals (hybridization). The chemisorption energy is, therefore, governed by two contributions from both the sp-band and d-band of the metal site:

$$\Delta E = \Delta E_{sp} + \Delta E_d. \tag{2}$$

Since the d-block metals have a similar, free-electron like sp-band, ΔE_{sp} can be approximated as a surface-independent constant while still having the largest contribution to the bonding.[67] The chemisorption energy differences from one metal to another for a given facet and site, thus, are governed mainly by the hybridization energy gain and orthogonalization energy cost of the interactions of adsorbate states with metal d-states[68–70]:

$$\Delta E_d = \Delta E_d^{orth} + \Delta E_d^{hyb}. \tag{3}$$

The orthogonalization cost of interacting orbitals, ΔE_d^{orth}, is proportional to the coupling integral, V, and orbital overlap integral, S:

$$\Delta E_d^{orth} = 2(\tilde{n}_a + f)SV. \tag{4}$$

The constant 2 considers spin degeneracy of the orbital, $\tilde{n}_a$ is the occupancy of the renormalized adsorbate states, and f is the idealized d-band filling of the metal atom. The orbital overlap integral, S, is linearly proportional to the coupling integral V by the coefficient α for a given orbital. α is a metal-independent parameter that is constant for a given adsorbate orbital and site type. V^2 can also be written as βV_{ad}^2, where β denotes the orbital coupling coefficient when the atoms are aligned along the z-axis. The standard values V_{ad}^2 for d-metals relative to Cu are readily available on the solid-state table.[71] The hybridization energy gain, ΔE_d^{hyb}, can be calculated from one-electron eigenenergies using Green's function approach[72]:

$$\begin{aligned}\Delta E_d^{hyb} = {} & \frac{2}{\pi}\int_{-\infty}^{\epsilon_f} \arctan\left(\frac{\Delta(\epsilon)}{\epsilon - \epsilon_a - \Lambda(\epsilon)}\right) d\epsilon \\ & - \frac{2}{\pi}\int_{-\infty}^{\epsilon_f} \arctan\left(\frac{\Delta_0(\epsilon)}{\epsilon - \epsilon_a}\right) d\epsilon,\end{aligned} \tag{5}$$

where $\Delta(\epsilon)$ is the chemisorption function consisting of the sp- and d-states contributions and $\Lambda(\epsilon)$ is its Hilbert transform. ϵ_a is the effective energy level of the renormalized adsorbate density of states. (Interested readers can refer to Refs. [69, 73] for further details on deriving the theory.)

A widely accepted concept from the d-band theory of chemisorption, commonly employed in catalysis and surface chemistry, correlates the binding energy of an adsorbate with the position of the electronic d-states' center, projected onto a metal site relative to the Fermi level.[74] According to this inference, a metal site with a higher d-band center binds adsorbates more strongly than geometrically equivalent sites on metals with the d-band center further down the Fermi level.[75,76] Although this simple take from the d-band theory has been able to capture the variations of chemisorption energy of a given adsorbate across metal sites, it is still inadequate to explain the complete chemisorption behavior of complex adsorbate-surface systems. Attaining a thorough understanding of reactivity trends can be made possible by further considering all related electronic factors and interaction parameters in the Newns−Anderson model. The TinNet framework allows us to gain a deep understanding by learning the local interaction parameters of the individual adsorbate orbitals with the metal states, including the substrate function of sp-states (Δ_0), renormalized OH orbital energies (ϵ_a^i), and effective coupling coefficients (β_i) as well as the d-band moments of site atoms if not given. For OH chemisorption on d-metals, the filled 3σ, double degenerate, partially-filled 1π, and empty $4\sigma^*$ electronic states are the frontier molecular orbitals of a gas phase OH radical that are known to be responsible for its bonding to the metal substrate. Therefore, current TinNet models have explicitly accounted for the interaction of these orbitals with the substrate sp- and d-states.

3.4 | Revealing reactivity origin via explainable AI

SHAP analysis provides critical insights into the origin of $Pt_3Co@Pt_{ML}$ enhanced ORR activity by explaining its weaker OH binding than pure Pt(111). As a near-optimal ORR catalyst, $Pt_3Co@Pt_{ML}$ binds OH ∼ 0.13 eV weaker. The change in *OH chemisorption energy on $Pt_3Co@Pt_{ML}$ relative to Pt is derived from a set of changes in electronic factors such as the coupling strength of adsorbate 1π resonance orbital with metal d-states, the center position of d-states, the width, etc. As shown in Figure 4a, the $1\pi{-}d$ coupling strength coefficient (β) and d-band center are major contributors to weakening OH binding on $Pt_3Co@Pt_{ML}$. The increased $1\pi{-}d$ coupling strength coefficient (β) by +0.15 leads to a 0.05 eV weaker bond and the downshifted d-band center by −0.16 eV also contributes to the weakening of the bond ∼0.05 eV. The increase in the coupling strength in this case is due to the extended d-states of metal sites influenced by the strain effect. The biaxial compressive strain of 2.19% from the Pt_3Co core

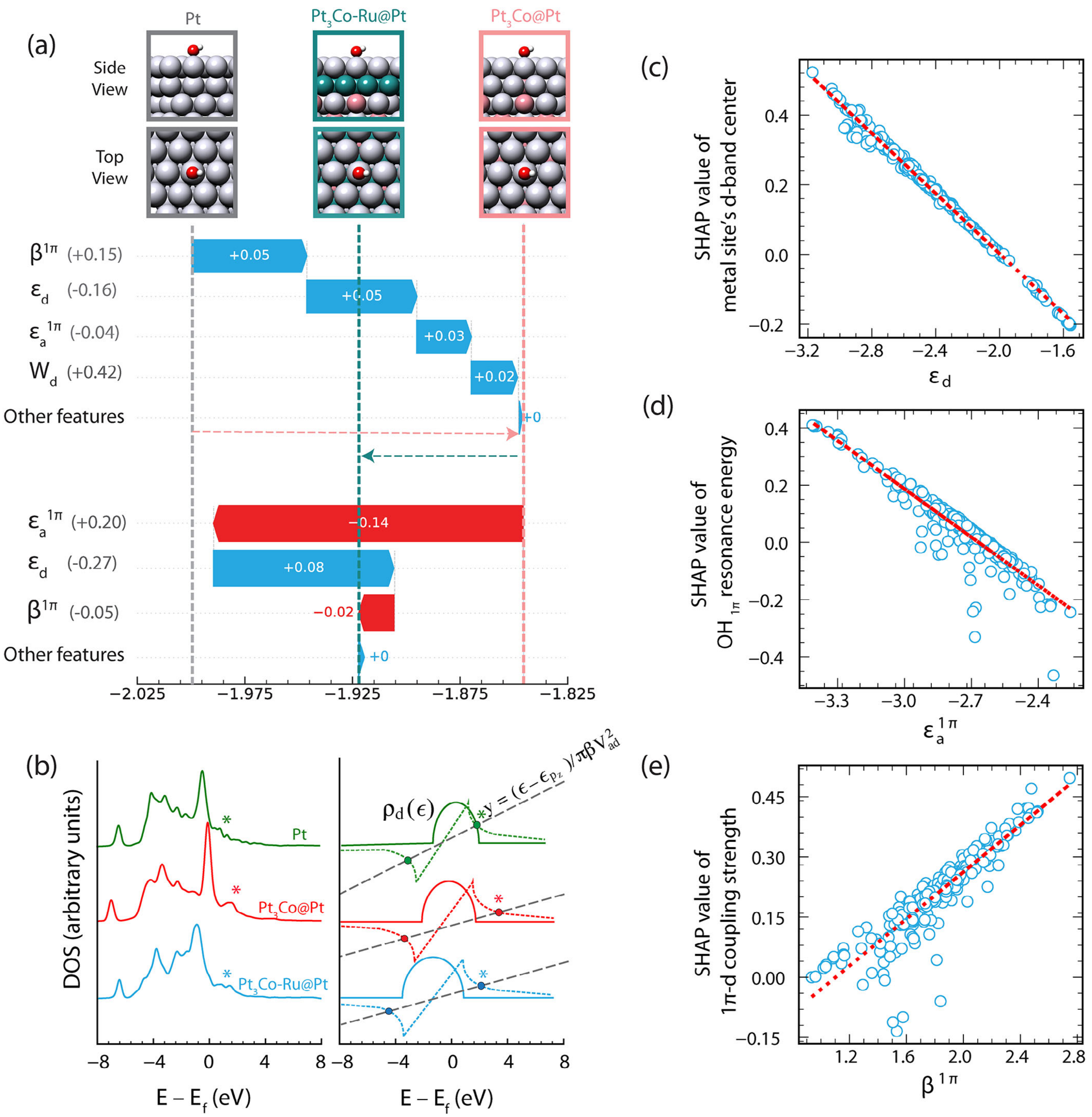

FIGURE 4 (a) Local interpretation of OH adsorption on Pt$_3$Co@Pt$_{ML}$ (111) relative to Pt(111) and Pt$_3$ Co-Ru@Pt$_{ML}$ (111) to Pt$_3$Co@Pt$_{ML}$ (111). The blue color indicates the features contributing to the weakening of the bond (positive Shapley value). The red color displays the parameters and factors pushing toward strengthening the bond (negative Shapley value), and the length of the ribbons is a direct measure of these contributions. (b) A graphical illustration of the Newns–Anderson model for an adsorbate resonance state interacting with metal *d*-states. DFT-calculated density of states (DOS) projected onto the 1π orbital of *OH on Pt$_3$ Co-Ru@Pt$_{ML}$, Pt$_3$Co@Pt$_{ML}$, and Pt are shown with adsorbate-metal antibonding states highlighted for comparison. (c) Global SHAP analysis of TinNet models of OH adsorption showing a strong correlation of the machine-learned *d*-band center (c), adsorbate 1π resonance energy (d), and $1\pi - d$ coupling strength with the adsorption-energy contribution, that is, the SHAP value. SHAP, SHapley Additive exPlanations.

on the Pt monolayer reduces the internuclear distances between Pt atoms on the surface. It intuitively increases the overlap between the *d*-states of metals and the adsorbate resonance states that leads to enhanced Pauli repulsion, making the Pt site less reactive. The downshift in the *d*-band center of metal sites is also a major contributor to the weakening of the bond as it leads to a lower lying, more occupied adsorbate-metal antibonding state.

Our screening framework has also identified Pt_3 Co-Ru@Pt_{ML} as a second-generation Pt_{ML} core–shell catalyst in the optimal region of theoretical activity volcano (Figure 3c). Looking into the SHAP analysis for OH chemisorption model for this structure compared with Pt_3Co@Pt_{ML} as a base, as illustrated in Figure 4a, it shows that the impact of the hybridization originated from the upshift of the renormalized 1π state contributes to the attractive binding energy and overcompensates the increase of repulsive contribution due to the downshift in the *d*-band center. The energy-level shift of renormalized adsorbate states is a function of the electron transfer from substrate *sp*-states. Being more electronegative than Co and less electronegative than Pt, Ru in this structure will lose more electrons to the surface Pt sites and it makes the surface Pt *sp*-bands more electron dense compared to Pt_3Co@Pt_{ML}. This electron transfer shifts the $\epsilon_a^{1\pi}$ up and eventually results in a more attractive contribution due to the less occupation of the adsorbate-metal antibonding states.

We reason that the change in the Pt−OH bond strength is primarily a function of the position of the metal *d*-states (ϵ_d) and renormalized oxygen 1π states ($\epsilon_a^{1\pi}$), and the Pt−OH coupling strength ($\beta_{1\pi-d}$), which together determine the filling of Pt−OH antibonding states. The partial dependency of model prediction on these factors is shown in Figure 4c–e. These plots illustrate the trend of variations of governing factors' contribution to OH chemisorption energy versus their values. The novel insights gained into the chemisorption process at metal sites are crucial for the systematic development of catalysts with improved activities. This understanding enables the manipulation of catalytic properties at specific sites by externally adjusting the key influencing factors, for example, by using electrolyte molecules or ions to exert an additional coupling term with the adsorbate energy level. Exploring those strategies to fine-tune electronic factors in chemical bonding can open up innovative approaches in catalyst design beyond active sites.

4 | CONCLUSION

In our study, we utilized the TinNet interpretable framework combined with an active learning approach to swiftly navigate through the vast design space of Pt-active sites. This enabled us to identify sites with optimal catalytic properties positioned near the peak of the activity volcano, while also elucidating the electronic factors driving reactivity trends. Our findings highlight the potential of modulating active sites through core–shell alloying and incorporating buffer layers of Ir, Rh, and Ru, offering a promising strategy to enhance surface reactivity and reduce Pt usage for the ORR. However, crafting second-generation Pt_{ML} core–shell alloy nanoparticles, smaller than 5 nm, with stable buffer layers and earth-abundant core elements, demands near-atomic precision in catalyst synthesis, posing a significant challenge in the field.

The development of user-centric explanations and comprehensive metrics is crucial for interpretable catalyst discovery. By applying explainable AI techniques, we gained insights into the electronic factors influencing the surface reactivity of the identified structures. Our research indicates that the adsorption energy of adsorbate–substrate pairs cannot be solely attributed to the *d*-band center's position. Other factors, including renormalized adsorbate states originating from *sp*-band interactions, are also critical in understanding catalysis functions.

ACKNOWLEDGMENTS

The project is partially supported by the NSF CBET Catalysis program (CBET-2245402) and the US Department of Energy, Office of Basic Energy Sciences under contract no. DE-SC0023323. The computational resource used in this work is provided by the advanced research computing at Virginia Polytechnic Institute and State University.

CONFLICT OF INTEREST STATEMENT

No conflict of interest was declared.

DATA AVAILABILITY STATEMENT

All data are available from the GitHub repository: https://github.com/hlxin/orr

ORCID

Noushin Omidvar https://orcid.org/0000-0001-6766-8548

REFERENCES

1. N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, G. Ceder, *Nature* **2023**, *624*, 86-91.
2. H. Xin, T. Mou, H. S. Pillai, S.-H. Wang, Y. Huang, *Acc. Mater. Res.* **2023**, *5*(1), 22-34.
3. J. Moon, W. Beker, M. Siek, J. Kim, H. S. Lee, T. Hyeon, B. A. Grzybowski, *Nat. Mater.* **2023**, *23*(1), 108–115.
4. H. Wang, J. Feng, Z. Dong, L. Jin, M. Li, J. Yuan, Y. Li, *npj Comput. Mater.* **2023**, *9*(1), 200.

5. E. A. Pogue, N. Alexander, K. McElroy, N. Q. Le, M. J. Pekala, I. McCue, E. Gienger, J. Domenico, E. Hedrick, T. M. McQueen, B. Wilfong, C. D. Piatko, C. R. Ratto, A. Lennon, C. Chung, T. Montalbano, G. Bassen, C. D. Stiles, *npj Comput. Mater.* **2023**, *9*, 1.
6. H. Choubisa, P. Todorović, J. M. Pina, D. H. Parmar, Z. Li, O. Voznyy, I. Tamblyn, E. H. Sargent, *npj Comput. Mater.* **2023**, *9*(1), 117.
7. K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, T. Bligaard, *Sci. Data* **2019**, *6*, 75.
8. D. D. Landis, J. S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov, K. W. Jacobsen, *Comput. Sci. Eng* **2012**, *14*, 51.
9. M. Álvarez-Moreno, C. de Graaf, N. López, F. Maseras, J. M. Poblet, C. Bo, *J. Chem. Inf. Model.* **2015**, *55*, 95.
10. L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, Z. Ulissi, *ACS Catal.* **2021**, *11*, 6059.
11. N. Omidvar, H. S. Pillai, S.-H. Wang, T. Mou, S. Wang, A. Athawale, L. E. Achenie, H. Xin, *J. Phys. Chem. Lett.* **2021**, *12*, 11476.
12. J. A. Esterhuizen, B. R. Goldsmith, S. Linic, *Nat. Catal.* **2022**, *5*, 175.
13. B. Hammer, J. K. Nørskov, *Surf. Sci.* **1995**, *343*, 211.
14. H. Xin, A. Vojvodic, J. Voss, J. K. Nørskov, F. Abild-Pedersen, *Phys. Rev. B Condens. Matter* **2014**, *89*, 115114.
15. X. Ma, Z. Li, L. E. K. Achenie, H. Xin, *J. Phys. Chem. Lett.* **2015**, *6*, 3528.
16. W. Samek, K.-R. Müller, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, (Eds.: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller), Springer International Publishing, Cham **2019**, pp. 5–22.
17. S.-H. Wang, H. S. Pillai, S. Wang, L. E. Achenie, H. Xin, *Nat. Commun.* **2021**, *12*, 1.
18. T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
19. B. Hammer, L. B. Hansen, J. K. Nørskov, *Phys. Rev. B* **1999**, *59*, 7413.
20. S. M. Lundberg, S.-I. Lee, *Advances in Neural Information Processing Systems*, *30*. (Eds.: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., Red Hook, NY, **2017**.
21. S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al.*Nat. Biomed. Eng.* **2018**, *2*, 749.
22. S. M. Lundberg, S.-I. Lee, in *Advances in Neural Information Processing Systems 30*, (Eds.: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., Red Hook, NY, **2017**, pp. 4765–4774.
23. J. K. Norskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard, H. Jonsson, *J. Phys. Chem. B* **2004**, *108*, 17886.
24. J. X. Wang, N. M. Markovic, R. R. Adzic, *J. Phys. Chem. B* **2004**, *108*, 4127.
25. N. M. Markovic, T. J. Schmidt, V. Stamenkovic, P. N. Ross, *Fuel Cells (Weinheim, Ger)* **2001**, *1*, 105.
26. N. M. Marković, P. N. Ross, *Surf. Sci. Rep.* **2002**, *45*, 117.
27. B. N. Grgur, N. M. Marković, P. N. Ross, *Can. J. Chem.* **1997**, *75*, 1465.
28. M. Chatenet, L. Genies-Bultel, M. Aurousseau, R. Durand, F. Andolfatto, *J. Appl. Electrochem.* **2002**, *32*, 1131.
29. B. B. Blizanac, C. A. Lucas, M. E. Gallagher, M. Arenz, P. N. Ross, N. M. Marković, *J. Phys. Chem. B* **2004**, *108*, 625.
30. R. R. Adžić, J. X. Wang, *Solid State Ionics* **2002**, *150*, 105.
31. S. Štrbac, N. A. Anastasijević, R. R. Adžić, *Electrochim. Acta* **1994**, *39*, 983.
32. V. Viswanathan, H. A. Hansen, J. Rossmeisl, J. K. Nørskov, *ACS Catal.* **2012**, *2*, 1654.
33. J. Zhang, M. B. Vukmirovic, Y. Xu, M. Mavrikakis, R. R. Adzic, *Angew. Chem.* **2005**, *117*, 2170.
34. C. Wang, D. Li, M. Chi, J. Pearson, R. B. Rankin, J. Greeley, Z. Duan, G. Wang, D. Van der Vliet, K. L. More, et al.*J. Phys. Chem. Lett.* **2012**, *3*, 1668.
35. D. Wang, H. L. Xin, R. Hovden, H. Wang, Y. Yu, D. A. Muller, F. J. DiSalvo, H. D. Abruña, *Nat. Mater.* **2013**, *12*, 81.
36. J. A. Keith, T. Jacob, *Angew. Chem., Int. Ed.* **2010**, *49*, 9521.
37. K. S. Exner, *ChemCatChem* **2023**, *15*, e202201222.
38. K. S. Exner, *Mater. Horiz.* **2023**, *10*, 2086.
39. F. Calle-Vallejo, A. Krabbe, J. M. García-Lastra, *Chem. Sci.* **2017**, *8*, 124.
40. H. A. Hansen, V. Viswanathan, J. K. Nørskov, *J. Phys. Chem. C* **2014**, *118*, 6706.
41. E. M. Fernández, P. G. Moses, A. Toftelund, H. A. Hansen, J. I. Martínez, F. Abild-Pedersen, J. Kleis, B. Hinnemann, J. Rossmeisl, T. Bligaard, J. K. Nørskov, *Angew. Chem. Int. Ed.* **2008**, *47*, 4683.
42. J. Rossmeisl, Z.-W. Qu, H. Zhu, G.-J. Kroes, J. K. Nørskov, *J. Electroanal. Chem.* **2007**, *607*, 83.
43. F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. R. Munter, P. G. Moses, E. Skulason, T. Bligaard, J. K. Norskov, *Phys. Rev. Lett.* **2007**, *99*, 016105.
44. J. Greeley, I. E. L. Stephens, A. S. Bondarenko, T. P. Johansson, H. A. Hansen, T. F. Jaramillo, J. Rossmeisl, I. Chorkendorff, J. K. Nørskov, *Nat. Chem.* **2009**, *1*, 552.
45. J. Zhang, M. B. Vukmirovic, Y. Xu, M. Mavrikakis, R. R. Adzic, *Angew. Chem., Int Ed Engl.* **2005**, *44*, 2132.
46. A. S. Bandarenka, H. A. Hansen, J. Rossmeisl, I. E. L. Stephens, *Phys. Chem. Chem. Phys.* **2014**, *16*, 13625.
47. V. Viswanathan, H. Hansen, J. Nørskov, *225th ECS Meet. Abstr.* **2014**, *MA2014-01*, 901.
48. J. Greeley, N. M. Markovic, *Energy Environ. Sci.* **2012**, *5*, 9246.
49. M. Mavrikakis, B. Hammer, J. K. Nørskov, *Phys. Rev. Lett.* **1998**, *81*, 2819.
50. J. R. Kitchin, J. K. Nørskov, M. A. Barteau, J. G. Chen, *Phys. Rev. Lett.* **2004**, *93*, 156801.
51. F. H. B. Lima, J. Zhang, M. H. Shao, K. Sasaki, M. B. Vukmirovic, E. A. Ticianelli, R. R. Adzic, *J. Phys. Chem. C* **2007**, *111*, 404.
52. M. Asano, R. Kawamura, R. Sasakawa, N. Todoroki, T. Wadayama, *ACS Catal.* **2016**, *6*, 5285.
53. W.-P. Zhou, X. Yang, M. B. Vukmirovic, B. E. Koel, J. Jiao, G. Peng, M. Mavrikakis, R. R. Adzic, *J. Am. Chem. Soc.* **2009**, *131*, 12755.
54. A. U. Nilekar, M. Mavrikakis, *Surf. Sci.* **2008**, *602*, L89.
55. J. Zhang, M. B. Vukmirovic, K. Sasaki, A. U. Nilekar, M. Mavrikakis, R. R. Adzic, *J. Am. Chem. Soc.* **2005**, *127*, 12480.
56. J. Wu, H. Yang, *Acc. Chem. Res.* **2013**, *46*, 1848.
57. D. Friebel, V. Viswanathan, D. J. Miller, T. Anniyev, H. Ogasawara, A. H. Larsen, C. P. O'Grady, J. K. Nørskov, A. Nilsson, *J. Am. Chem. Soc.* **2012**, *134*, 9664.

58. J. R. Kitchin, J. K. Nørskov, M. A. Barteau, J. G. Chen, *J. Chem. Phys.* **2004**, *120*, 10240.
59. N. Schweitzer, H. Xin, E. Nikolla, J. T. Miller, S. Linic, *Top. Catal.* **2010**, *53*, 348.
60. M. P. Hyman, J. W. Medlin, *J. Phys. Chem. C* **2007**, *111*, 17052.
61. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al.*APL Mater.* **2013**, *1*, 011002.
62. J. K. Nørskov, F. Abild-Pedersen, F. Studt, T. Bligaard, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 937.
63. W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, G. Ceder, *Sci. Adv.* **2016**, *2*, e1600225.
64. A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, C. Wolverton, *Chem. Mater.* **2016**, *28*, 5621.
65. V. Stamenkovic, B. S. Mun, K. J. Mayrhofer, P. N. Ross, N. M. Markovic, J. Rossmeisl, J. Greeley, J. K. Nørskov, *Angew. Chem.* **2006**, *118*, 2963.
66. V. R. Stamenkovic, B. S. Mun, K. J. Mayrhofer, P. N. Ross, N. M. Markovic, *J. Am. Chem. Soc.* **2006**, *128*, 8813.
67. B. Hammer, Y. Morikawa, J. K. Norskov, *Phys. Rev. Lett.* **1996**, *76*, 2141.
68. B. Hammer, J. K. Norskov, in *Chemisorption and Reactivity on Supported Clusters and Thin Films: Toward an Understanding of Microscopic Processes in Catalysis* (Eds: R. M. Lambert, G. Pacchioni), Springer, Dordrecht, **1997**, pp. 285.
69. S. Wang, H. S. Pillai, H. Xin, *Nat. Commun.* **2020**, *11*, 1.
70. A. Vojvodic, J. K. Nørskov, F. Abild-Pedersen, *Top. Catal.* **2014**, *57*, 25.
71. W. A. Harrison, *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond*, Courier Corporation, Chelmsford, MA **2012**.
72. D. M. Edwards, D. M. Newns, *Phys. Lett. A* **1967**, *24*, 236.
73. D. Newns, *Phys. Rev.* **1969**, *178*, 1123.
74. B. Hammer, J. K. Nørskov, *Surf. Sci.* **1996**, *359*, 306.
75. J. R. Kitchin, J. K. Norskov, M. A. Barteau, J. G. Chen, *J. Chem. Phys.* **2004**, *120*, 10240.
76. H. Xin, S. Linic, *J. Chem. Phys.* **2010**, *132*, 221101.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: N. Omidvar, S.-H. Wang, Y. Huang, H. S. Pillai, A. Athawale, S. Wang, L. E. K. Achenie, H. Xin, *Electrochem. Sci. Adv.* **2024**, e2300028. https://doi.org/10.1002/elsa.202300028