Explainable Machine Learning Framework for Decoding Zhang–Rice Singlet Formation and Oxygen Evolution Reaction Kinetics

Haoranyue Shen
Department of Computer Science, University of Houston, Houston, TX, USA.
hellohaoranyue@uh.edu

Longjin Zheng
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
longzheng04@binghamton.edu

Rilliot Mershell
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
emarshall@uab.edu

Abstract

Deciphering the electronic structure origin of catalytic activity in the oxygen evolution reaction (OER) remains a central challenge for sustainable energy technologies. The Zhang–Rice singlet, a two-hole bound state on a single transition metal site, has emerged as a key descriptor linking local electronic structure to OER kinetics, yet its formation pathways under operando conditions are highly complex. This paper presents a system-level explainable machine learning framework that decodes Zhang–Rice singlet formation and OER kinetics from multimodal operando spectroscopy data. The architecture couples deep representation learning with post-hoc interpretability techniques, enabling the extraction of physically meaningful features while preserving high predictive fidelity. We examine structural trade-offs between model accuracy, transparency, and computational cost across a distributed data infrastructure built on cloud-native, FAIR-compliant pipelines. The analysis extends to governance and fairness concerns, including dataset bias, underrepresentation of catalyst families, and the need for federated learning to protect intellectual property while enabling collaborative model refinement. Robustness is addressed through uncertainty quantification and adversarial validation protocols, while sustainability considerations motivate efficiency-aware model selection and resource budgeting. Policy implications concerning open science mandates, reproducibility, and equitable access to advanced catalyst informatics tools are discussed. By integrating technical and socio-technical dimensions, the proposed framework establishes a principled pathway toward interpretable, responsible, and scalable data-driven discovery in electrocatalysis.

Keywords

explainable machine learning, Zhang–Rice singlet, oxygen evolution reaction, operando spectroscopy, catalyst informatics, model interpretability, data infrastructure, governance.

1. Introduction

The development of efficient catalysts for the oxygen evolution reaction (OER) is fundamental to the viability of water electrolysis, metal–air batteries, and solar fuel generation. In transition metal oxides, OER activity is intimately governed by the electronic structure of the active site, where subtle modifications in metal–oxygen hybridization can dramatically alter reaction barriers. Among the most intriguing electronic motifs is the Zhang–Rice singlet state, originally formulated to describe the low-energy excitations in hole-doped cuprates [4]. The concept has recently been extended to electrocatalysis, where the formation of Zhang–Rice singlets through ligand holes on oxygen ligands has been linked to enhanced OER performance. Machine learning (ML) is increasingly deployed to navigate the vast compositional and operational space of heterogeneous catalysts, leveraging high-throughput computational data and experimental characterizations to predict activity trends [1]. However, the opaque nature of state-of-the-art deep neural networks obstructs the extraction of causal physical mechanisms, creating a pressing need for explainable artificial intelligence (XAI) methods such as local interpretable model-agnostic explanations [2] and Shapley additive explanations [8] that can illuminate the learned relationships between spectroscopic signatures and catalytic function.

The complexity of OER is compounded by the fact that active site evolution occurs under electrochemical polarization and cannot be captured by static ground-state calculations alone. Operando X-ray absorption and photoemission spectroscopies have revealed dynamic changes in oxidation state, covalency, and local symmetry that precede and accompany oxygen–oxygen bond formation [5]. Recent advances have demonstrated that the Zhang–Rice singlet state can emerge through a sequential two-step oxidation process, wherein a ligand hole first delocalizes across metal–oxygen bonds and subsequently couples antiferromagnetically with a metal center to stabilize a highly active configuration [7]. This mechanistic nuance underscores the necessity of interpretable frameworks that can decompose spectroscopic time series into specific electronic and structural events.

Computational catalyst design has historically relied on scaling relations and descriptor-based approaches that simplify the multi-dimensional nature of reaction kinetics [6]. While these models offer physical transparency, they often fail to capture the configurational flexibility and non-ideal surface dynamics present in real electrocatalysts. Conversely, deep learning models trained on operando data can uncover higher-order correlations but risk becoming black-box oracles. The resulting tension between accuracy and interpretability is not merely a technical inconvenience; it manifests at the system level, influencing how research communities build, validate, and trust data-driven insights. This paper introduces a holistic explainable machine learning framework that situates model architecture within a broader infrastructure of data pipelines, governance protocols, robustness assessments, and sustainability constraints, aiming to decode Zhang–Rice singlet formation and OER kinetics in a manner that is both physically insightful and institutionally responsible.

2. Background on Zhang–Rice Singlets and OER Kinetics

The Zhang–Rice singlet was originally introduced to explain the electronic structure of superconducting copper oxides, where a doped hole residing primarily on oxygen ligands forms a spin singlet with the central Cu 3d9 hole [4]. In the context of OER, an analogous configuration emerges when strong metal–oxygen covalency enables substantial ligand-hole character on the oxygen ions adjacent to the active metal site. Spectroscopic signatures of this state include characteristic pre-edge features in X-ray absorption spectra and shifts in the oxygen K-edge that reflect unoccupied O 2p states hybridized with transition metal d states.

Operando X-ray absorption near-edge structure and resonant inelastic X-ray scattering have become indispensable tools for monitoring such electronic rearrangements under reaction conditions [5].

The mechanistic role of the Zhang–Rice singlet in OER kinetics is governed by the interplay between electron removal steps and the stabilization of oxyl intermediates. The two-step oxidation pathway identified in recent operando studies reveals that the first oxidation creates a ligand hole that is not yet antiferromagnetically coupled, while the second oxidation triggers the formation of the truly singlet-coupled state, which drastically lowers the thermodynamic barrier for O–O bond formation [7]. This sequence implies that OER activity does not simply scale with the number of formal oxidation equivalents but depends critically on the temporal ordering and electronic correlation of charge-transfer events. Consequently, decoding singlet formation requires models that can parse time-resolved spectral data into discrete electronic transitions and attribute kinetic relevance to each step.

Traditional descriptor-based models, such as the oxygen p-band center relative to the metal d-band center, condense the electronic structure into a single scalar and have been effective for rationalizing bulk trends [3]. Yet they cannot resolve the transient population of intermediate spin states or the local environment effects that differentiate surface and subsurface sites. Advanced theoretical methods, including constrained density functional theory and dynamical mean-field theory, can simulate singlet formation but remain computationally prohibitive for high-throughput screening [6]. Machine learning offers a middle ground by learning complex mappings from operando spectral features to kinetic observables, provided that the learned representations remain open to physical inspection. Bridging the gap between physical modeling and data-driven prediction is therefore a central motivation for the framework developed here.

3. Architecture of the Explainable Machine Learning Framework

The proposed framework is built on a modular architecture that separates representation learning, prediction, and explanation into interconnected but independently configurable components. At the representation layer, a combination of convolutional neural networks and graph neural networks operates on operando spectra and local coordination graphs extracted from X-ray absorption fine structure analyses. This multimodal encoder is designed to capture both local electronic signatures and longer-range structural motifs that influence the stability of ligand holes. The decoder head maps the learned latent representation to kinetic observables including Tafel slopes, turnover frequencies, and faradaic efficiency. Crucially, the framework does not treat prediction as an end in itself; it embeds explanation modules that can be invoked post-training without retraining the core model.

For interpretability, the framework integrates two complementary families of explanation methods. Gradient-weighted class activation mapping enables the visualization of the spatial and spectral regions most influential for a given prediction, directly linking peaks in the X-ray absorption pre-edge region to singlet-associated transitions [9]. Shapley additive explanations decompose the prediction into additive feature contributions, quantifying the global importance of orbital hybridization metrics and geometry parameters while also providing instance-level justifications [8]. The architecture supports both post-hoc local explanations and inherently interpretable surrogate models such as decision trees fitted on the latent features, allowing users to choose the trade-off between fidelity and simplicity based on the downstream task.

A distinctive architectural choice is the inclusion of a physics-informed inductive bias through an auxiliary loss that encourages the latent space to preserve known physical constraints, such as charge neutrality sum rules and invariance to permutational symmetries of the ligand environment. This constraint not only improves generalization under limited training data but also structures the latent space in a way that aligns with interpretable physical variables, facilitating the extraction of reaction coordinates. The architecture thus navigates the classic trade-off between black-box accuracy and white-box transparency by enforcing a latent geometry that is both discriminative and physically meaningful, a design principle that has been advocated for scientific machine learning applications [12].

4. System-Level Infrastructure and Data Pipelines

Deploying an explainable ML framework for operando catalysis research requires a robust infrastructure that can ingest, harmonize, and serve heterogeneous data streams from geographically distributed synchrotron facilities and electrochemical testing laboratories. The framework leverages cloud-native storage and compute resources, adopting a microservices architecture in which data preprocessing, model training, inference, and explanation services are containerized and orchestrated through Kubernetes [20]. All experimental data, including raw beamline files, are converted into standardized hierarchical data formats and annotated with rich metadata capturing experimental conditions, sample provenance, and instrument configurations. This infrastructure aligns with the FAIR guiding principles—making data findable, accessible, interoperable, and reusable—which are essential for collaborative catalyst discovery [11].

A central design consideration is the latency and throughput trade-off between on-demand explainability and batch prediction. Operando experiments generate streaming data at kilohertz rates, and the framework supports near-real-time inference with explanation budgets controlled by an adaptive gating mechanism that decides when a detailed SHAP or Grad-CAM analysis is triggered, balancing computational cost against the need for immediate physical insight. Data provenance is tracked through a lineage graph stored in a distributed ledger, enabling the full reproducibility of any model output from raw spectrum to explanation. This lineage is critical when findings must be audited by experimental collaborators or journal reviewers, linking the high-level interpretation of singlet formation back to the original spectroscopic measurements.

The framework also interfaces with public materials databases such as the Materials Project, which provide precomputed electronic structure descriptors that can be used to augment sparse experimental datasets [10]. By automatically fetching reference data and aligning it with measured spectra through a feature alignment pipeline, the system reduces the manual overhead of data wrangling and minimizes errors introduced by inconsistent energy calibration. The infrastructure is designed to be modular, allowing individual research groups to contribute new data connectors and processing algorithms while maintaining a shared, versioned model registry that serves as the single source of truth for all deployed predictive and explanatory models.

5. Model Governance, Fairness, and Data Representativeness

Data-driven catalyst discovery is susceptible to systematic biases that arise from the uneven exploration of chemical space. Most operando OER studies to date have concentrated on first-row transition metal oxides under narrow electrolyte conditions, leading to datasets that are heavily skewed toward Co, Ni, and Fe systems while underrepresenting precious metal oxides

and multi-metallic perovskites. When an explainable ML model trained on such data is queried for predictions on underrepresented compositions, the explanations it generates may be physically plausible within the training manifold but fail outside it, yielding overconfident and misleading interpretations. Addressing this challenge requires a governance framework that audits dataset composition, quantifies representation gaps, and enforces fairness-aware evaluation protocols [14].

Federated learning protocols offer a promising avenue for collaboratively training models across multiple institutions without sharing raw experimental data, thereby mitigating privacy and intellectual property concerns that often impede data aggregation in competitive research domains [16]. In the proposed architecture, participating laboratories train local model replicas on their proprietary operando datasets and periodically share encrypted gradient updates with a central aggregation server. The central model then disseminates a consensus representation that captures a broader diversity of catalyst families while preserving the confidentiality of individual experimental conditions. This federated arrangement directly addresses the fairness problem by reducing the dominance of well-studied systems and enabling the continuous inclusion of novel catalyst chemistries as they emerge from exploratory synthesis efforts.

Governance extends beyond training to the deployment lifecycle of models and explanations. A model card system, analogous to those advocated for high-stakes machine learning applications, is implemented to document the intended use, performance characteristics, and known limitations of each model version. For catalyst informatics, the model card records the composition range, pH window, and potential range over which the Zhang–Rice singlet explanations are valid, alongside benchmark metrics on held-out operando datasets. Such transparency enables end users—including synthetic chemists and device engineers—to assess the trustworthiness of both predictions and the accompanying mechanistic narratives, fostering responsible adoption.

6. Robustness, Validation, and Uncertainty Quantification

Robustness in the context of catalyst informatics encompasses both the stability of predictions under perturbations of input spectra and the generalization of explanations across different catalyst compositions and reaction environments. Spectral data from operando experiments are inherently noisy due to photon statistics, beam instability, and sample heterogeneity, making it essential to evaluate whether small input fluctuations can alter the assigned feature importance rankings and thereby change the mechanistic story. The framework incorporates adversarial robustness testing by constructing worst-case perturbations within physically plausible noise envelopes and requiring that explanation consistency, measured via explanation similarity metrics, remains above a predefined threshold [13]. This ensures that minor experimental variability does not lead to qualitatively different assignments of rate-determining electronic steps.

Uncertainty quantification is embedded as a first-class architectural component rather than an afterthought. Bayesian deep learning approxi-mations, such as Monte Carlo dropout, are applied to the representation encoder, producing predictive distributions for kinetic parameters and enabling the estimation of epistemic uncertainty arising from data sparsity in certain regions of catalyst space [13]. The explanation module propagates these uncertainties to feature importance scores, marking spectral regions where the model's attribution is unreliable and should be interpreted with caution. Such uncertainty-aware explanations are particularly valuable when analyzing borderline cases where the Zhang–Rice singlet feature is

only subtly developed, helping researchers distinguish genuine physical signatures from artifacts.

External validation is performed against independent experiments not included in the training or federated aggregation processes. A curated holdout set comprising operando data from iridium-based model catalysts, where the ligand-hole activation mechanism has been thoroughly characterized, serves as a benchmark for framework performance [18]. The ability of the framework to recover the known two-step oxidation pathway and correctly identify the emergence of the singlet-coupled state without prior mechanistic prompting is used as a definitive validation criterion. These validation protocols are codified in an automated continuous integration pipeline that runs whenever new data or model updates are registered, ensuring that system-level reliability is maintained over the entire research lifecycle.

7. Sustainability and Computational Trade-offs

The computational demands of training large-scale multimodal models and repeatedly computing post-hoc explanations can be substantial, raising sustainability concerns in an era where the carbon footprint of artificial intelligence is under increasing scrutiny. The framework adopts a green AI philosophy by integrating energy-aware scheduling, model distillation, and dynamic explanation depth controls [15]. Before training a full-capacity model, a lightweight screening model based on random forests or boosted trees is used to identify candidate catalysts of high interest, thereby reserving expensive deep learning computations for the most promising or perplexing cases. Explanation computations are similarly tiered: rapid gradient-based saliency maps are generated by default, while full Shapley value computations are selectively triggered only when high-confidence mechanistic interpretation is required.

Architectural choices carry direct carbon implications. The use of graph neural networks, while expressive, can lead to exponential growth in computation with the size of the local coordination environment; thus, the framework enforces a maximum cutoff radius and a maximum atom count per graph, striking a balance between representational fidelity and computational tractability. Periodic assessment of model efficiency using metrics such as floating-point operations per explanation and energy consumption per insight enables the research community to make informed trade-offs between model complexity and environmental cost. These assessments are made transparent through sustainability reports attached to each model release, encouraging accountability.

From an infrastructure perspective, the choice of cloud region, energy mix, and hardware accelerator type can significantly alter the carbon intensity of model operations. The framework's orchestration layer is configured to preferentially schedule training and large-scale explanation jobs in data centers powered by renewable energy sources, when available, and to exploit spot instances and low-utilization periods to reduce overall energy consumption. Such system-level optimizations align the pursuit of catalyst discovery with broader societal goals of decarbonization, ensuring that the tools used to design efficient water-splitting catalysts do not themselves become an unsustainable burden.

8. Policy Implications and Open Science

The convergence of machine learning and operando electrocatalysis raises fundamental policy questions regarding data ownership, reproducibility, and equitable access to advanced computational tools. National and international materials genome initiatives have called for open data ecosystems to accelerate discovery, yet competitive pressures often incentivize data

hoarding [21]. The presented framework addresses this tension by enabling tiered data sharing: metadata and derived features can be made openly accessible while raw proprietary spectra remain under institutional control, with federated learning bridging the gap. Policy frameworks that recognize and credit such intermediate forms of data sharing are essential to incentivize participation without compromising commercial interests.

Reproducibility has emerged as a major concern in machine learning research, and its importance is amplified when models are used to derive mechanistic claims that influence experimental directions and funding decisions. The framework's lineage tracking, containerization, and model card system collectively provide a robust reproducibility infrastructure that surpasses current norms in catalysis informatics [17]. Journals and funding agencies could mandate the deposition of model cards and explanation trails alongside traditional supplementary information, establishing a new standard for computational evidence in mechanistic studies.

Open science principles further demand that the explanation outputs themselves be understandable to domain experts who may not be machine learning specialists. The framework generates interactive explanation dashboards that display spectral overlays with attribution heatmaps, allowing researchers to visually confirm that the model is focusing on physically expected pre-edge features rather than spectral artifacts. Such tools democratize access to mechanistic interpretation, reducing the asymmetry between groups that possess in-house machine learning expertise and those that do not [22]. The long-term vision is an open, community-governed repository of validated catalyst explanations that can serve as a collective knowledge base for the OER community, accelerating the design cycle while maintaining rigorous standards of evidence.

9. Conclusion

This paper has presented a comprehensive system-level perspective on the development and deployment of an explainable machine learning framework for decoding Zhang–Rice singlet formation and oxygen evolution reaction kinetics. By architecting a modular pipeline that tightly integrates representation learning, post-hoc interpretability, and physics-informed constraints, the framework navigates the inherent trade-offs between predictive accuracy and mechanistic transparency. The analysis has demonstrated that realizing the full potential of such a framework demands far more than algorithmic innovation; it requires deliberate attention to infrastructure design, from cloud-native data pipelines to federated governance protocols that address dataset bias and fairness. Robustness and uncertainty quantification mechanisms ensure that the mechanistic narratives derived from the model remain reliable under realistic experimental variability, while sustainability considerations guide the responsible use of computational resources. Policy dimensions surrounding open science, reproducibility, and equitable access have been woven into the fabric of the framework, positioning it as a socio-technical artifact that can foster collaborative, trustworthy, and green catalyst discovery. Future work will extend the framework to encompass other multi-electron electrocatalytic reactions and integrate autonomous experiment steering, further closing the loop between data-driven insight and experimental realization.

References

1. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J., & Sutton, C. (2018). Machine learning for heterogeneous catalyst design and discovery. AIChE Journal, 64(7), 2311–2323.

2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

3. Rossmeisl, J., Qu, Z.-W., Zhu, H., Kroes, G.-J., & Nørskov, J. K. (2007). Electrolysis of water on oxide surfaces. Journal of Electroanalytical Chemistry, 607(1–2), 83–89.

4. Zhang, F. C., & Rice, T. M. (1988). Effective Hamiltonian for the superconducting Cu oxides. Physical Review B, 37(7), 3759–3761.

5. Frenkel, A. I., Rodriguez, J. A., & Chen, J. G. (2012). Synchrotron-based techniques for in situ and operando characterization of catalytic materials. ACS Catalysis, 2(11), 2269–2280.

6. Nørskov, J. K., Bligaard, T., Rossmeisl, J., & Christensen, C. H. (2009). Towards the computational design of solid catalysts. Nature Chemistry, 1(1), 37–46.

7. Peng, C. K., Lin, Y. C., Chiang, C. L., Qian, Z., Huang, Y. C., Dong, C. L., ... & Lin, Y. G. (2023). Zhang-Rice singlets state formed by two-step oxidation for triggering water oxidation under operando conditions. Nature Communications, 14(1), 529.

8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

9. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

10. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., ... & Ceder, G. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials, 1(1), 011002.

11. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

12. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071–22080.

13. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 1050–1059.

14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

15. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

16. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

17. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d’Alché-Buc, F., ... & LeCun, Y. (2021). Improving reproducibility in machine learning research.

Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks.

18. Grimaud, A., Demortière, A., Saubanère, M., Dachraoui, W., Duchamp, M., Doublet, M.-L., & Tarascon, J.-M. (2017). Activation of surface oxygen sites on an iridium-based model catalyst for the oxygen evolution reaction. Nature Energy, 2(1), 1–10.
19. Esterhuizen, J. A., Goldsmith, B. R., & Linic, S. (2022). Interpretable machine learning for knowledge generation from heterogeneous catalysis. Nature Catalysis, 5(3), 175–185.
20. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50–58.
21. de Pablo, J. J., Jackson, N. E., Webb, M. A., Chen, L.-Q., Moore, J. E., Morgan, D., ... & Crespi, V. H. (2019). New frontiers for the materials genome initiative. npj Computational Materials, 5(1), 41.
22. Molloy, J. C. (2011). The open knowledge foundation: open data means better science. PLoS Biology, 9(12), e1001195.