From Prediction to Proposal in Catalysis: robust evaluation, LRP explanations, and relevance-guided candidate generation

Parastoo Semnani[1,2,3*] **Mihail Bogojeski**[1,2] **Florian Bley**[1,2] **Zizheng Zhang**[4] **Qiong Wu**[4] **Thomas Kneib**[4] **Jan Herrmann**[5] **Christoph Weisser**[5] **Florina Patcas**[5] **Klaus-Robert Müller**[1,2,6,7]

[1]Machine Learning Group, TU Berlin, 10587 Berlin, Germany [2]Berlin Institute for the Foundations of Learning and Data (BIFOLD), 10587 Berlin, Germany [3]BASLEARN–TU Berlin/BASF Joint Lab for Machine Learning, TU Berlin, 10587 Berlin, Germany [4]Chair of Statistics & Campus Institute Data Science, Georg-August-University Göttingen, 37073 Göttingen, Germany [5]Production AI, BASF SE, 67056 Ludwigshafen, Germany [6]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany [7]Department of Artificial Intelligence, Korea University, 02841 Seoul, South Korea

Abstract

Machine learning (ML) can accelerate experimentation in chemistry and materials, but models trained on small, and highly class imbalanced datasets often look deceptively strong when judged by accuracy alone and provide limited guidance for follow-up simulations or experiments.

We present a robust ML and explainable artificial intelligence (XAI) framework for catalyst yield classification that emphasizes: (i) robust evaluation under class imbalance, (ii) signed, class-aware explanations via Layer-wise Relevance Propagation (LRP) for neural networks and neuralized support vector machines, and (iii) a simple relevance-guided sampler to propose promising compositions. This framework has been implemented on oxidative coupling of methane (OCM) to evaluate the performance of a range of ML models: tree-based models (such as decision trees, random forest, and gradient boosted trees), logistic regression, support vector machines, and neural networks. The proposed framework yields reliable generalization estimates under scarcity and mitigates imbalance during training. The attribution layer interrogates model decisions: tree importances are stable but class-agnostic, whereas signed LRP isolates features that contribute positively to the high-yield class. Using these signed signals to bias a validity-preserving sampler enriches model-predicted high-yield candidates. The resulting workflow forms a practical interface between scalable ML and experimental validation.

1 Introduction

Machine learning (ML) and data-driven methods are reshaping heterogeneous catalyst discovery by exposing nonlinear interactions among elements and supports to shrink the search space and minimize trial-and-error iterations [1–5]. This is critical for systems where synergistic and antagonistic effects complicate yield optimization [6, 7]. Recent applications show that data-driven models can capture subtle composition–performance relationships and guide hypothesis generation [8–10].

However, ML in catalysis is constrained by *data reality*: experimental datasets are typically small, imbalanced, and biased toward historically favored chemistries [11–13]. Even with community efforts to reduce selection bias via unbiased/high-throughput screens (HTS) [14–18], positives remain rare, making accuracy deceptively optimistic and hampering generalization [19]. These factors call

*Corresponding author: p.semnani@tu-berlin.de

Our code can be found at https://github.com/PSemnani/XAI4CatalyticYield

European Conference on Neural Information Processing Systems (EurIPS) Workshop: SimBioChem 2025.

for evaluation protocols and explainability that explicitly address the minority class of high-yield catalysts.

At the same time, interpretability is essential for transferring ML signals into chemically meaningful actions. Tree-based feature importances are class-agnostic; in contrast, propagation-based XAI such as LRP produce *signed, per-sample* attributions that indicate whether a feature promotes or suppresses a target class [20–25]. Recent work extends LRP to non-neural models via neuralization, enabling faithful propagation through, e.g., kernel SVMs. [26]

Contributions. We build on these insights and introduce a framework that (i) stabilizes evaluation under scarcity and imbalance with F1-score-centered evaluation and nested cross-validation (CV) and resampling techniques, and (ii) provides class-aware explanations via LRP for neural networks and a neuralized SVM, complemented by tree importances (Fig. 1) and (iii) finally turn aggregated relevances into priors for a validity-preserving sampler that proposes feasible catalyst compositions. The design mirrors the original unbiased generator so that proposals remain in-distribution. [27] Applied to the unbiased OCM dataset of Nguyen et al. [14], the framework delivers reliable estimates and signed attributions that can be elevated to relevance-guided proposal generation.

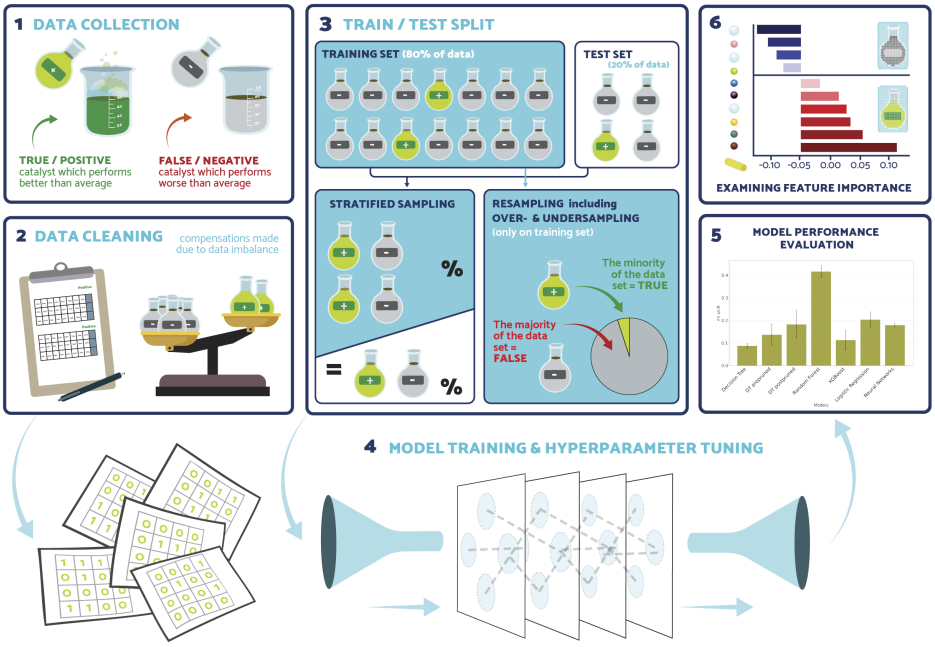


Figure 1: Illustration of the ML framework, starting with data collection and cleaning (steps 1-2), and visualizing the process for obtaining the performance and explanations of a model on a single random train-test split of the dataset (steps 3-6). These training and evaluation steps are then repeated for 100 different train-test splits and the results are aggregated to produce robust performance estimates and feature importance scores.

2 Dataset

Under suitable catalysts, oxidative coupling of methane (OCM) converts CH_4 to C_2 products (C_2H_4, C_2H_6), key building blocks in the chemical industry; performance is typically quantified by the C_2 yield (%) (conversion $\times$ selectivity). Prior catalyst informatics has reported synergistic combinations such as Na–La, Na–Mn, and Ba–Sr [16], but progress is confounded by inconsistent experimental protocols and literature selection bias in component choice [28, 12].

Unbiased HTE dataset. We adopt the process-consistent, randomized OCM screen of Nguyen et al. [14]: 300 quaternary M1–M2–M3/support catalysts drawn uniformly from a declared pool (27 elements + "none"; 9 supports), each evaluated under 135 conditions by high-throughput

experimentation (HTE). For each composition, the best C_2 yield across conditions is retained. To the best of our knowledge, this is the only publicly available OCM dataset with the largest number of unique compositions and an approximately uniform selection frequency of elements/supports—a property that reduces anthropogenic oversampling of "usual suspects" and stabilizes both learning and explanation by limiting spurious frequency-driven importance. The randomized design covers a discrete space of 36,540 combinations with repeated sampling of components [14], enabling fairer evaluation and cleaner class-conditional explanations.

Labels and thresholding. In Nguyen et al. [14], yields were grouped into high/neutral/low bands around the measured catalyst-free baseline (best blank $\approx 10.2\%$ at the reported conditions), using a margin chosen to exceed experimental error. Following the same rationale, we collapse the three bands to a binary target: *high* (positive) if the best C_2 yield is $> 13\%$, and *low* (negative) otherwise. This yields 51 high and 240 low entries (291 total; 9 missing). Element and support presence are encoded as binary indicators. Because process conditions are not provided in the dataset, we focus on composition-only prediction at each composition's best operating condition.

3 Methods

3.1 Models

For classifying the yield of a catalyst, we compare a range of ML models in our framework; including decision trees (pre/post-pruned), Random Forest, XGBoost, Logistic Regression, RBF-SVM, and a small Neural Network (NN); hyperparameters are tuned in the inner loop; results are averaged over 100 train/test splits.

Evaluation protocol. When dealing with small datasets, the performance of the model can depend strongly on the choice of the training and test subsets, making it difficult to obtain a reliable estimate of the model's generalization error. In such cases, providing an accurate and unbiased estimate of the error through cross-validation and hyper-parameter tuning becomes essential, which in turn allows for the selection of the most robust and best-performing model [29, 11]. In our study, we use a variant of nested k-fold cross-validation to reliably evaluate model performance on unseen data [30, 31].

Imbalance handling. ML models often struggle within scenarios with highly imbalanced class distributions [32, 33]. To address imbalance, we oversample with SMOTE [34] and randomly undersample the majority class during training [35–41]. Test folds remain untouched. This shifts the decision boundary toward the minority class, typically increasing Recall and F1-score (while Precision or Accuracy may decrease slightly).

3.2 Explainable AI (XAI)

XAI techniques are playing an increasingly important role in various domains, including catalyst research. [42]

Tree-based models For decision trees, node impurity is

$$\mathrm{Gini}(t) \;=\; 1 - \sum_{c} p_c(t)^2, \tag{1}$$

and feature importance for Random Forest aggregates the total decrease in impurity attributable to feature d across the ensemble:

$$\mathrm{FI}^{\mathrm{RF}}(d) \;=\; \frac{1}{T} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t(d)} \Delta\mathrm{Gini}_t(s). \tag{2}$$

XGBoost reports the summed split gains for feature d:

$$\mathrm{FI}^{\mathrm{XGB}}(d) \;=\; \sum_{s \in \mathcal{S}(d)} \mathrm{Gain}(s). \tag{3}$$

Layer-wise Relevance Propagation (LRP) Layer-wise Relevance Propagation (LRP) redistributes a classifier's output back to intermediate units and ultimately to the inputs via local, conservative propagation rules, yielding *per-sample*, *signed* attributions [20, 21, 43, 23]. Let a_i and a_j denote lower- and upper-layer activations with weights w_{ij}. For a relevance signal R_j at the upper layer, a broad class of rules assigns relevance to lower units as

$$R_i \;=\; \sum_j \frac{\rho(w_{ij})\, a_i}{\sum_{i'} \rho(w_{i'j})\, a_{i'} + \epsilon}\, R_j, \tag{4}$$

where $\rho(\cdot)$ specifies a weighting (e.g., identity, ϵ- or γ-stabilized), and $\epsilon > 0$ prevents division by zero. These rules are chosen to conserve total relevance layerwise (up to numerical stabilization), so that the sum of input relevances equals the chosen output quantity.

To make directionality explicit, we explain a *contrastive* output. In binary classification with evidence units a_+ and a_- (pre-softmax logits), we set the starting relevance to the logit difference

$$\eta \;=\; a_+ - a_- \;=\; \sum_k \big(w_{+,k} - w_{-,k}\big)\, a_k, \tag{5}$$

so that positive input relevance indicates evidence *promoting* the high-yield class and negative relevance indicates evidence *promoting* the low-yield class. For multilayer perceptrons we use the γ-rule on linear layers,

$$\rho(w_{ij}) \;=\; w_{ij} + \gamma\, \max(0, w_{ij}), \qquad \gamma = 0.2, \tag{6}$$

which emphasizes positive contributions and improves stability; activations are handled with the identity rule (treating the nonlinearity locally as constant), and we fall back to ϵ-stabilization when needed [43].

Because part of the relevance can be absorbed by bias terms, raw input relevance may not exactly sum to η. We therefore apply a sign-preserving rebalancing so that positive and negative input relevance match the positive and negative parts of the explained output:

$$\sum_d \rho^+\, \max(R_d, 0) \;=\; \sum_k \max\big((w_{+,k} - w_{-,k})\, a_k,\, 0\big), \tag{7}$$

$$\sum_d \rho^-\, \min(R_d, 0) \;=\; \sum_k \min\big((w_{+,k} - w_{-,k})\, a_k,\, 0\big), \tag{8}$$

yielding input maps whose signed sums recover the contrastive logit while preserving the sign of each R_d.

Since LRP is local (per sample), we form global profiles by averaging rescaled input relevances across test samples and splits:

$$\bar{R}_d \;=\; \frac{1}{S\,N} \sum_{i=1}^{S} \sum_{j=1}^{N} R_d\big(\mathbf{x}_j^{(i)}\big). \tag{9}$$

We report both *signed* $\bar{R}_d$ (class-aware) and *absolute* $|\bar{R}_d|$ (for comparability with nonnegative tree importances).

LRP for neuralized SVMs To enable propagation-based explanations for kernel SVMs, we adopt *neuralization* [44]: the RBF–SVM decision is rewritten as a contrast of class-wise evidence,

$$g(\mathbf{x}) \;=\; \log \sum_{i \in +} \alpha_i \exp\big(-\gamma\, \|\mathbf{x} - \mathbf{x}_i\|^2\big) \;-\; \log \sum_{j \in -} |\alpha_j| \exp\big(-\gamma\, \|\mathbf{x} - \mathbf{x}_j\|^2\big), \tag{10}$$

and realized by a shallow network with a detection layer and soft pooling:

$$g(\mathbf{x}) \;=\; \gamma \cdot \min_j{}^{\gamma} \left(\max_i{}^{\gamma} \big(\mathbf{w}_{ij}^\top \mathbf{x} + b_{ij} \big) \right), \mathbf{w}_{ij} = 2(\mathbf{x}_i - \mathbf{x}_j),\; b_{ij} = \|\mathbf{x}_j\|^2 - \|\mathbf{x}_i\|^2 + \frac{1}{\gamma} \log \frac{\alpha_i}{|\alpha_j|}. \tag{11}$$

This network is *decision-equivalent* to the SVM and admits LRP with conservative rules for soft-min / soft-max layers and LRP-0 on the linear detection layer. We finally apply a sign-preserving rescaling so that the summed positive/negative input relevance equals the corresponding evidence terms in (10). Full rule statements and derivations are provided in Appx. C.

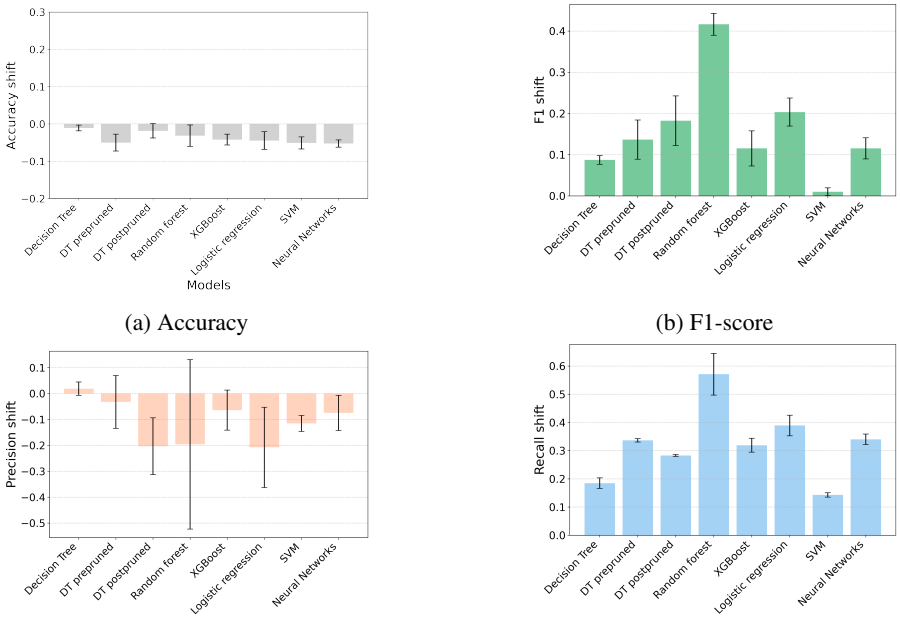
4 Results

We organize results into three parts: (i) yield prediction in a sparse, imbalanced regime (metric lens and the effect of resampling), (ii) interpretability via tree importances and class-aware LRP for neural networks and neuralized SVMs, and (iii) turning explanations into proposals using a relevance-guided sampler.

All experiments were run on a MacBook Pro with an Apple M1 chip and 36 GB RAM. The code for our framework and the experiments is available at `https://github.com/PSemnani/XAI4CatalyticYield`.

4.1 Predicting yield in a sparse and imbalanced regime

Before presenting the final model comparison, we first discuss the choice of performance metrics and the impact of resampling.



(a) Accuracy

(b) F1-score

(c) Precision

(d) Recall

Figure 2: The impact of introducing resampling techniques on key performance metrics (a) Accuracy, (b) F1-score, (c) Precision, (d) Recall.

4.1.1 Metric lens: Accuracy vs. F1-score

On this dataset, a trivial "all-negative" classifier would achieve ≈ 0.82 accuracy, so accuracy can be deceptively optimistic under class imbalance. In contrast, the F1-score directly reflects minority-class utility: for reference, a random classifier yields $\mathrm{F1}{\approx}0.26$, an all-negative classifier 0.0, and an all-positive classifier 0.3. All models substantially exceed these baselines, with $\mathrm{F1}$ in the 0.46–0.52 range, revealing separations that accuracy obscures (Table 1).

4.1.2 Effect of resampling (SMOTE)

Figure 2 summarizes the effect of oversampling the high-yield class (SMOTE) and undersampling the low-yield class on the four metrics. Accuracy (Fig. 2a) shows only a small decrease (-0.01 to -0.05 across models). By rebalancing the training distribution, SMOTE shifts the decision boundary toward the minority class. As the model observes more minority examples, it becomes more likely to predict the positive class; correspondingly, both F1-score (Fig. 2b) and Recall (Fig. 2d) increase markedly for most models, while we observe some negative shifts in the Precision as Fig. 2c depicts. In

design-of-experiments (DOE) for rare high-yield catalysts, recall and F1-score are preferred operating metrics to avoid missing promising candidates.

Final comparison across models. Across 100 stratified outer folds, mean±std are reported in Table 1. F1-scores are close (0.46–0.52) with no single method statistically separating from the rest, likely reflecting data limitations rather than model capacity; with larger, higher-quality data, higher-capacity learners (XGBoost, SVM, NN) are expected to separate from simpler tree/linear baselines.

Table 1: Mean±std over 100 splits with the proposed framework.

Model	Accuracy	F1-score
Decision Tree	0.75 ± 0.05	0.46 ± 0.10
Pre-pruned Tree	0.73 ± 0.06	0.47 ± 0.08
Post-pruned Tree	0.81 ± 0.04	0.50 ± 0.13
Random Forest	0.78 ± 0.05	0.52 ± 0.09
XGBoost	0.77 ± 0.05	0.51 ± 0.09
Logistic Regression	0.78 ± 0.05	0.51 ± 0.10
SVM (RBF)	0.77 ± 0.05	0.49 ± 0.09
Neural Network	0.76 ± 0.05	0.51 ± 0.10

4.2 Interpretability (XAI): Tree importances and LRP relevance scores

4.2.1 Tree-based importances

To aggregate across tree models, we normalize importances to $[0, 1]$ per model and average across splits.

$$\bar{R}_d = \frac{1}{S} \sum_{i=0}^{S} R_d(m^{(i)}), \tag{12}$$

where S is the number of training/test splits, $R_d(m^{(i)})$ is the feature importance for feature d extracted from the model m trained on the training subset from split i.

The resulting global profile (Fig. 3) is stable and highlights globally "important" features (e.g., Mn, Al_2O_3, SiO_2, Ni, CeO_2), but remains *class-agnostic* and therefore does not indicate whether a feature promotes high-yield or explains low-yield.

4.3 LRP for Neural networks and SVM

LRP provides *per-sample* attributions; to obtain a *global* profile we aggregate across both samples and splits. For each split, we compute rescaled input relevances for the high-yield logit and then average over all test samples and all S splits:

$$\bar{R}d; =; \frac{1}{S, N} \sum i = 1^{S} \sum_{j=1}^{N} R_d!\big(\mathbf{x}_j^{(i)}\big), \tag{13}$$

where N is the number of test samples per split and $R_d(\cdot)$ is the relevance of feature d. Unlike tree importances, LRP yields *signed* scores: with the high-yield logit as the starting point, positive input relevance indicates evidence *promoting* the high-yield class, while negative relevance indicates evidence for low yield. For comparability with strictly positive tree importances, we also report global maps from *absolute* LRP values.

For SVMs, explanations are obtained via the neuralization procedure [26] followed by LRP; for NNs, we apply LRP directly. In both cases, input relevances are rescaled to compensate relevance absorbed by bias terms so that evidence is conserved.

Figure 4 contrasts absolute vs. signed relevance scores. The signed view reveals that several features with the largest *absolute* scores actually provide *negative* evidence for the high-yield class (e.g., Mn,

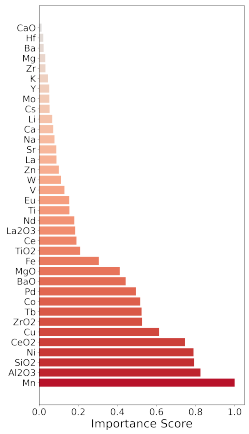


Figure 3: Averaged importance scores for all features across the different tree-based models (Decision tree, DT prepruned, DT postpruned, Random forest, XGBoost)

Al_2O_3, SiO_2), whereas features with smaller absolute scores provide *positive* evidence and thus act as high-yield promoters (notably La_2O_3/BaO). This dataset–specific pattern highlights why class direction matters: class-agnostic absolute importances can be dominated by features that chiefly explain the low-yield class. The dataset's unified element selection further mitigates frequency bias (features do not accrue importance merely by appearing more often), making it well suited for XAI within our framework. Larger datasets will help assess how broadly these observations generalize.

4.4 Predicting promising catalyst compositions via relevance scores

To illustrate the application of feature importances in generating promising catalyst compositions, a generative algorithm was devised that leverages these relevances to bias the generation process toward catalysts predicted by the model to exhibit high yield. All proposals are drawn *within* the experimental design space of Nguyen et al. [14]: one oxide support, up to three distinct elements (sampling without replacement), "none" allowed. This guaranties in-distribution candidates w.r.t. the HTS protocol.

Let $\bar{R}_d$ be the *average* relevance of feature d (tree: impurity/gain; NN/SVM: LRP). Split features into elements $\mathcal{E}$ and supports $\mathcal{S}$, then map relevances to discrete probabilities with a temperatured softmax:

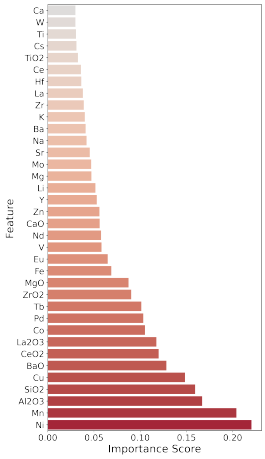
$$\mathrm{softmax}(\mathbf{x}, \beta)_i = \frac{e^{\beta x_i}}{\sum_j e^{\beta x_j}}, \tag{14}$$

Here, $\beta^{\mathcal{E}}$ is used for elements and $\beta^{\mathcal{S}}$ for supports.

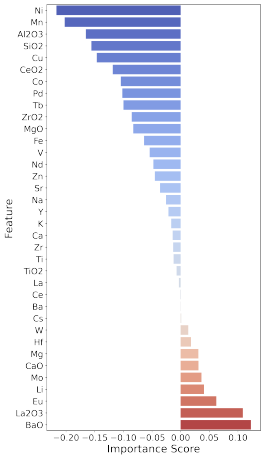
Higher β concentrates mass on features with larger relevance; lower β explores more uniformly. The procedure mirrors the dataset generator in Nguyen et al. [14] so all sampled catalysts are valid under that scheme. A detailed step-by-step description can be found in Appx. Algorithm 1.

Sampling procedure

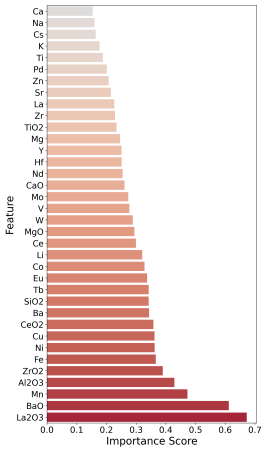
1. Compute $\mathbf{p}^{\mathcal{S}} = \mathrm{softmax}\big(\{\bar{R}_d : d \in \mathcal{S}\}, \beta^{\mathcal{S}}\big)$ and sample one support S^{sel}.
2. For $i = 1, 2, 3$: (a) with probability $1/|\mathcal{E}|$ select "no element" (allowing 1–2 component catalysts); else
   (b) compute $\mathbf{p}^{\mathcal{E}} = \mathrm{softmax}\big(\{\bar{R}_d : d \in \mathcal{E}\}, \beta^{\mathcal{E}}\big)$, sample E_i^{sel}, and remove it from $\mathcal{E}$ (sampling without replacement).



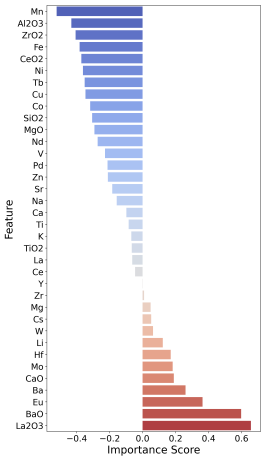
(a) NN — absolute importances (LRP)



(b) NN — signed (class-aware) relevance (LRP)



(c) SVM — absolute importances (LRP)



(d) SVM — signed (class-aware) relevance (LRP)

Figure 4: LRP-based feature analysis (mean) for neural networks and SVMs. (a,b) NN: absolute vs. signed relevance; (c,d) SVM: absolute vs. signed relevance. Signed relevance separates features that promote high yield (positive) from those that promote low yield (negative).

This yields a candidate $(S^{\text{sel}}, E^{\text{sel}}_{1:3})$, which we then score with the originating model.

Using mean relevances over 100 splits, we generated 1000 candidates per setting for: (i) XGBoost with absolute importances; (ii) NN with absolute importances; (iii) NN with *signed, class-aware* LRP relevances. Table 2 reports the fraction of generated candidates predicted as high-yield by the corresponding model.

Table 2: Fraction of generated candidates classified as high-yield (higher is better). As bias increases (larger β), signed LRP enriches positives; absolute importances do the opposite.

Temperature $(\beta^{\mathcal{E}}, \beta^{\mathcal{S}})$	NN: signed	NN: abs.	XGBoost: abs.
$(10,\ 1)$	0.38	0.17	0.31
$(20,\ 2)$	0.49	0.13	0.23
$(40,\ 4)$	0.68	0.04	0.13
$(40,\ 4)$	0.85	0.01	0.05

Absolute importances tend to rank features that are *strong for predicting the majority (low-yield) class*; biasing a generator toward them *reduces* the positive rate. In contrast, signed LRP isolates features that *promote* the desired (high-yield) class, so increasing β *increases* the fraction of high-yield proposals.

For design loops that use explanations to guide simulation or experiment, prefer *class-aware* (signed) relevances over absolute importances when turning explanations into generative priors.

5 Conclusion

We introduced a compact ML and XAI framework tailored to data scarcity and class imbalance in experimental catalyst datasets. While we have chosen to apply the framework to OCM as a representative example in this case, the general design of the framework allows it to be applied for various other catalytic reactions. Using stratified nested cross-validation, training-fold resampling, and F1-score as the primary criterion, we showed that accuracy compresses differences and can be matched by trivial baselines, whereas F1-score reveals meaningful minority-class skill. Resampling improves recall and F1-score for most models, most notably for Random Forest, while SVM changes are minimal, aligning with margin-based behavior.

On the interpretability side, LRP delivers signed, class-aware attributions that isolate high-yield promoters from features that explain low yield. In neural networks and neuralized SVMs, two groups of components, namely rare earth oxides (La and Eu) and alkaline earth metals with a high degree of alkalinity (Ba and Ca) consistently appear as positive contributors. These findings align with chemical intuition and existing OCM literature. Notably, this agreement with established OCM chemistry emerges despite the small dataset used. Future work should scale to larger, more diverse datasets spanning broader compositions and varied process conditions, which would improve model fidelity and explanations and help uncover previously unobserved component–condition interactions.

Turning these signed signals into priors, our relevance-guided sampler biases proposals toward promising compositions, illustrating how explanations can drive candidate generation. This reproducible end-to-end workflow provides a practical template for extending the integration of ML-simulation to larger and more diverse catalytic datasets and process conditions. The resulting blueprint (robust evaluation, class-based explanation and targeted generation) offers an actionable path for scaling to larger datasets and for coupling ML with experimental validation.

Limitations and future work. Limitations include evaluation on a single task and binary composition descriptors. The intent is to support DOE by providing *composition priors*, not full synthesis recipes: catalyst performance also depends on variables not modeled here. Moreover, LRP attributions—though signed and class-aware—depend on model architecture and propagation choices; our stabilization and rebalancing improve faithfulness but do not eliminate method variance. Future work will broaden descriptors and conditions, compare LRP with alternatives (e.g., SHAP; IG), and pursue prospective, closed-loop validation with laboratory partners to quantify enrichment and tighten the ML-to-experiment loop.

Acknowledgments and Disclosure of Funding

This work was supported by BASLEARN, TU Berlin/BASF Joint Laboratory, co-financed by TU Berlin and BASF SE. P.S., M.B., F.B. and K.-R.M. acknowledge support by the German Federal Ministry of Education and Research (BMBF) for BIFOLD (BIFOLD24B). K.-R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0- 00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation) and by the German Federal Ministry for Education and Research (BMBF) under Grants 01IS14013BE and 01GQ1115. C. W. acknowledges support by BASF Data and AI Academy. The authors thank Stef Lenk for illustrating Figure 1.

References

[1] Jake Graser, Steven K Kauwe, and Taylor D Sparks. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chemistry of Materials*, 30(11):3601–3612, 2018.

[2] Mie Andersen, Sergey V Levchenko, Matthias Scheffler, and Karsten Reuter. Beyond scaling relations for the description of catalytic materials. *Acs Catalysis*, 9(4):2752–2759, 2019.

[3] Sicong Ma and Zhi-Pan Liu. Machine learning for atomic simulation and activity prediction in heterogeneous catalysis: current status and future. *ACS Catalysis*, 10(22):13213–13226, 2020.

[4] Mihail Bogojeski, Simeon Sauer, Franziska Horn, and Klaus-Robert Müller. Forecasting industrial aging processes with machine learning methods. *Computers & Chemical Engineering*, 144:107123, 2021.

[5] Johannes T Margraf, Hyunwook Jung, Christoph Scheurer, and Karsten Reuter. Exploring catalytic reaction networks with machine learning. *Nature Catalysis*, 6(2):112–121, 2023.

[6] Aleksandra Vojvodic and Jens K Nørskov. New design paradigm for heterogeneous catalysts. *National Science Review*, 2(2):140–143, 2015.

[7] Bryan R Goldsmith, Jacques Esterhuizen, Jin-Xun Liu, Christopher J Bartel, and Christopher Sutton. Machine learning for heterogeneous catalyst design and discovery. *AIChE Journal*, 2018.

[8] Srinivas Rangarajan. Artificial intelligence in catalysis. In *Artificial Intelligence in Manufacturing*, pages 167–204. Elsevier, 2024.

[9] Keisuke Suzuki, Takashi Toyao, Zen Maeno, Satoru Takakusagi, Ken-ichi Shimizu, and Ichigaku Takigawa. Statistical analysis and discovery of heterogeneous catalysts based on machine learning from diverse published data. *ChemCatChem*, 11(18):4537–4547, 2019.

[10] Song Wang and Jun Jiang. Interpretable catalysis models using machine learning with spectroscopic descriptors. *ACS Catalysis*, 13(11):7428–7436, 2023.

[11] Takashi Toyao, Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu. Machine learning for catalysis informatics: recent applications and prospects. *Acs Catalysis*, 10(3):2260–2297, 2019.

[12] Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron E Ruby, Hao Wang, Sorelle A Friedler, Alexander J Norquist, et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573(7773): 251–255, 2019.

[13] Dean G Brown, Moriah M Gagnon, and Jonas Bostrom. Understanding our love affair with p-chlorophenyl: present day implications from historical biases of reagent selection. *Journal of Medicinal Chemistry*, 58(5):2390–2405, 2015.

[14] Thanh Nhat Nguyen, Sunao Nakanowatari, Thuy Phuong Nhat Tran, Ashutosh Thakur, Lauren Takahashi, Keisuke Takahashi, and Toshiaki Taniike. Learning catalyst design based on bias-free data set for oxidative coupling of methane. *ACS Catalysis*, 11(3):1797–1809, 2021. doi: 10.1021/acscatal.0c04629. URL `https://cads.eng.hokudai.ac.jp/datamanagement/datasources/f7e30001-e440-4c1a-be64-ea866b2f77cb/`. Random catalyst OCM data by HTE.

[15] Thanh Nhat Nguyen, Thuy Tran Phuong Nhat, Ken Takimoto, Ashutosh Thakur, Shun Nishimura, Junya Ohyama, Itsuki Miyazato, Lauren Takahashi, Jun Fujima, Keisuke Takahashi, et al. High-throughput experimentation and catalyst informatics for oxidative coupling of methane. *Acs Catalysis*, 10(2):921–932, 2019.

[16] Ulyana Zavyalova, Martin Holena, Robert Schlögl, and Manfred Baerns. Statistical analysis of past catalytic data on oxidative methane coupling for new insights into the composition of high-performance catalysts. *ChemCatChem*, 3(12):1935–1947, 2011.

[17] Keisuke Takahashi, Lauren Takahashi, Son Dinh Le, Takaaki Kinoshita, Shun Nishimura, and Junya Ohyama. Synthesis of heterogeneous catalysts in catalyst informatics to bridge experiment and high-throughput calculation. *Journal of the American Chemical Society*, 144 (34):15735–15744, 2022.

[18] Shun Nishimura, Son Dinh Le, Itsuki Miyazato, Jun Fujima, Toshiaki Taniike, Junya Ohyama, and Keisuke Takahashi. High-throughput screening and literature data-driven machine learning-assisted investigation of multi-component la 2 o 3-based catalysts for the oxidative coupling of methane. *Catalysis Science & Technology*, 12(9):2766–2774, 2022.

[19] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[20] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140. URL `https://doi.org/10.1371/journal.pone.0130140`.

[21] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[22] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[23] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. doi: 10.1109/JPROC.2021.3060483.

[24] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[25] A Saranya and R Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, page 100230, 2023.

[26] Florian Bley. Explaining kernel classifiers and extensions. Master's thesis, Technical University of Berlin, January 2022.

[27] Parastoo Semnani, Mihail Bogojeski, Florian Bley, Zizheng Zhang, Qiong Wu, Thomas Kneib, Jan Herrmann, Christoph Weisser, Florina Patcas, and Klaus-Robert Muller. A machine learning and explainable ai framework tailored for unbalanced experimental catalyst discovery. *The Journal of Physical Chemistry C*, 128(50):21349–21367, 2024.

[28] Roman Schmack, Alexandra Friedrich, Evgenii V Kondratenko, Jörg Polte, Axel Werwatz, and Ralph Kraehnert. A meta-analysis of catalytic literature data reveals property-performance correlations for the ocm reaction. *Nature communications*, 10(1):441, 2019.

[29] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[30] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, 12(2):181, 2001.

[31] Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O Anatole Von Lilienfeld, Alexandre Tkatchenko, and Klaus-Robert Muller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of chemical theory and computation*, 9(8):3404–3419, 2013.

[32] Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013.

[33] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[34] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[35] Ivan Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, 1976. doi: 10.1109/TSMC.1976.4309452.

[36] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In Douglas H. Fisher, editor, *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, pages 179–186. Morgan Kaufmann, 1997.

[37] Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79, 1998.

[38] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II*, volume 11, 2003.

[39] Alexander Liu, Joydeep Ghosh, and Cheryl Martin. Generative oversampling for mining imbalanced datasets. In Robert Stahlbock, Sven F. Crone, and Stefan Lessmann, editors, *Proceedings of the 2007 International Conference on Data Mining, DMIN 2007, June 25-28, 2007, Las Vegas, Nevada, USA*, pages 66–72. CSREA Press, 2007.

[40] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5375–5384. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.580. URL `https://doi.org/10.1109/CVPR.2016.580`.

[41] Colin Bellinger, Roberto Corizzo, and Nathalie Japkowicz. Remix: Calibrated resampling for class imbalance in deep learning. *arXiv preprint arXiv:2012.02312*, 2020.

[42] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. In Irena Koprinska, Michael Kamp, Annalisa Appice, Corrado Loglisci, Luiza Antonie, Albrecht Zimmermann, Riccardo Guidotti, Özlem Özgöbek, Rita P. Ribeiro, et al., editors, *ECML PKDD 2020 Workshops*, pages 417–431, Cham, 2020. Springer International Publishing. ISBN 978-3-030-65965-3. doi: 10.1007/978-3-030-65965-3_28.

[43] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_10. URL `https://doi.org/10.1007/978-3-030-28954-6_10`.

[44] Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1926–1940, 2024. doi: 10.1109/TNNLS.2022.3185901.

B Evaluation metrics definition

Definition of different commonly used performance evaluation metrics for ML models. Let TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

C LRP for Neuralized SVMs: Detailed Rules and Rescaling

Soft-pooling operators. $\max^\gamma(z) = \frac{1}{\gamma}\log\sum\exp(\gamma z)$ and $\min^\gamma(z) = -\frac{1}{\gamma}\log\sum\exp(-\gamma z)$ [44].

LRP through soft pooling. Deep Taylor–derived conservative propagation:

$$R_j = \frac{\exp(-a_j)}{\sum_{j'}\exp(-a_{j'})}\,R_k, \tag{15}$$

$$R_{ij} = \frac{\exp(a_{ij})}{\sum_{i'}\exp(a_{i'j})}\,R_j. \tag{16}$$

LRP to inputs. LRP-0 on affine units $(\mathbf{w}_{ij}^\top\mathbf{x} + b_{ij})$:

$$R_d \;=\; \frac{w_{ij,d}\,x_d}{\sum_{0,d'} w_{ij,d'}\,x_{d'}}\,R_{ij}. \tag{17}$$

Sign-preserving bias compensation. Rescale the signed input map so that its positive/negative sums recover the class-wise evidence of (10):

$$\sum_d \rho^+\,\max(R_d, 0) = \log\sum_{i\in+}\alpha_i\exp\big(-\gamma\,\|\mathbf{x}-\mathbf{x}_i\|^2\big), \tag{18}$$

$$\sum_d \rho^-\,\min(R_d, 0) = -\log\sum_{j\in-}|\alpha_j|\exp\big(-\gamma\,\|\mathbf{x}-\mathbf{x}_j\|^2\big). \tag{19}$$

Global aggregation. Per-sample input relevances are averaged over test samples and splits to form global SVM profiles (Eq. 13 in the main text).

D Relevance sampling algorithm

Algorithm 1: A simple sampling algorithm for generating promising catalyst combinations based on feature importances provided by an explainability method.

Data:
$\bar{R}_d$ - set of feature importances
$\mathcal{E}$ - feature indices of elements
$\mathcal{S}$ - feature indices of supports
$\beta^{\mathcal{E}}$ - temperature parameter for the softmax applied on element relevances
$\beta^{\mathcal{S}}$ - temperature parameter for the softmax applied on support relevances
Result:
S^{sel} - feature index of sampled support
$E_1^{\text{sel}}, E_2^{\text{sel}}, E_3^{\text{sel}}$ - feature indices of sampled elements

$\bar{R}^{\mathcal{S}} \leftarrow \{\bar{R}_d : d \in \mathcal{S}\}$; `// Select importances of support features`
$\mathbf{p}^{\mathcal{S}} \leftarrow \text{softmax}(\bar{R}^{\mathcal{S}}, \beta^{\mathcal{S}})$; `// Create probability distribution over supports`
$S^{\text{sel}} \sim \mathbf{p}^{\mathcal{S}}$; `// Sample from the probability distribution over supports`
for i in $1 \dots 3$ **do**
    $\bar{R}^{\mathcal{E}} \leftarrow \{\bar{R}_d : d \in \mathcal{E}\}$; `// Select importances of element features`
    $r \sim \text{uniform}(0, 1)$;
    if $r < \frac{1}{|\mathcal{E}|}$ **then** `// No element is selected with a chance of` $1/|\mathcal{E}|$
        $E_1^{\text{sel}} \leftarrow None$;
    else
        `// Sample element and remove it from the list of indices`
        $\mathbf{p}^{\mathcal{E}} \leftarrow \text{softmax}(\bar{R}^{\mathcal{E}}, \beta^{\mathcal{E}})$;
        $E_i^{\text{sel}} \sim \mathbf{p}^{\mathcal{E}}$;
        $\mathcal{E}.\text{remove}(E_i^{\text{sel}})$;

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   The paper claims a robust evaluation protocol under class imbalance, class-aware LRP explanations (NN/SVM), and a relevance-guided sampler; all are implemented and validated on the OCM dataset with scope and limitations clearly stated.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   We explicitly discuss limits (single benchmark; binary composition descriptors; no process-condition features; method-dependence of LRP) and outline future directions (richer descriptors, SHAP/IG comparison, prospective validation).

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   The contribution is empirical/methodological; no new theorems or proofs are introduced beyond standard definitions and cited LRP rules.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   We specify data source and encoding, stratified nested CV, SMOTE+undersampling applied to training folds only, model families, Optuna tuning, metrics, and 100 randomized splits; code and scripts are released to reproduce tables/figures.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Code: `https://github.com/PSemnani/XAI4CatalyticYield`. Data: Nguyen *et al.*, ACS Catal. 2021, 11, 1797–1809, DOI: `http://dx.doi.org/10.1021/acscatal.0c04629`; "Not for redistribution"—reuse requires citation and contacting the dataset owners as stated.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   We report stratified nested CV (outer/inner), 100 splits, SMOTE+undersampling on training folds, model families and tuning strategy, metrics (Accuracy/Precision/Recall/F1), and provide additional settings in the supplement.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

We report mean±std across 100 randomized splits and use error bars reflecting split-to-split variability; the text clarifies that bars represent standard deviation.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Experiments ran on an Apple M1 laptop with 36 GB RAM and no GPU. The released scripts reproduce all figures/tables.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   We use a public experimental dataset and standard ML methods; no human subjects or sensitive data are involved; we adhere to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Potential positives include more efficient DOE and reduced experimental waste; risks include misprioritization if models are over-trusted. We mitigate via class-aware explanations, reporting variability, and advocating prospective validation and training on a bigger data sets.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    No high-risk models or scraped web-scale datasets are released; we rely on a curated experimental dataset from prior work.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    We cite Nguyen *et al.* (ACS Catal. 2021, 11, 1797–1809; DOI: `http://dx.doi.org/10.1021/acscatal.0c04629`); the dataset is "Not for redistribution" and reuse requires citation and contacting the owners as specified.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    We release a public repository (`https://github.com/PSemnani/XAI4CatalyticYield`) with instructions, environment details, and scripts to reproduce the main tables/figures.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

No human subjects or crowdsourcing are involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    No human-subjects research was conducted.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [NA]

    No LLMs were used as core, non-standard components of the methodology; any language editing did not affect the scientific method.