

A Multimodal Semantic Learning Framework for Data-Driven Discovery of Nitrate-to-Ammonia Electrocatalysts

Jiajia Liu^{a#}, Yuying Jiang^{b#}, Ran Ma^a, Zhenhui Ma^{b*}, Shuren Wang^{a, c, d*}, Yanglong Hou^{a, c, d*}

- a. School of Materials Science and Engineering, Peking University, Beijing, 100871, China
- b. Department of Physics, Beijing Technology and Business University, Beijing, 100048, China
- c. School of Materials, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, China
- d. State Key Laboratory of Optoelectronic Materials and Technologies, Sun Yat-sen University, Guangzhou, 510275, China

[#] These authors contributed equally to this work.

*Corresponding author: E-mail: mazh@btbu.edu.cn (Zhenhui Ma), shurenwang_wsr@pku.edu.cn (Shuren Wang), hou@pku.edu.cn (Yanglong Hou)

Supplementary Notes

Supplementary Note 1: Exact Prompts for LLM-Assisted Data Extraction.

<Advanced_eNO₃RR_Extraction>

<Task>

Extract only explicitly stated information from the provided literature excerpt related to nitrate reduction reaction (NO₃RR) electrocatalysts.

Do not infer, guess, calculate, or hallucinate missing information.

If a field is not explicitly mentioned in the source text, output "not mentioned".

Preserve the original terminology, units, and numerical values exactly as reported in the source text.

</Task>

<Output_Rules>

<Rule_1>Use only the fields defined in this schema.</Rule_1>

<Rule_2>Do not generate free text outside the schema.</Rule_2>

<Rule_3>Do not convert or normalize values unless explicitly required by the source text.</Rule_3>

<Rule_4>Do not assign default labels when information is absent.</Rule_4>

<Rule_5>For categorical fields, use "not mentioned" if the category cannot be identified from the text.</Rule_5>

<Rule_6>For numerical fields, preserve the original unit and numerical precision exactly as reported.</Rule_6>

<Rule_7>For multiple values, list only those explicitly reported in the source text.</Rule_7>

<Rule_8>For reaction mechanisms, pathways, energies, adsorption values, or kinetic parameters, extract only when the source text explicitly reports them.</Rule_8>

</Output_Rules>

<Extraction_Schema>

<Intelligent_Induction>

<Catalyst_Type>not mentioned</Catalyst_Type>

<Synthesis_Method>not mentioned</Synthesis_Method>

<Regulation_Strategy>not mentioned</Regulation_Strategy>

</Intelligent_Induction>

<Catalyst_Properties>

<Composition>not mentioned</Composition>

<Architecture>not mentioned</Architecture>

<Defect_Engineering>not mentioned</Defect_Engineering>

</Catalyst_Properties>

<Reaction_Environment>

<Cell_Type>not mentioned</Cell_Type>

<Electrolyte_Profile>

<Base_Electrolyte>not mentioned</Base_Electrolyte>

<Nitrate_Source>not mentioned</Nitrate_Source>

<pH_Value>not mentioned</pH_Value>

</Electrolyte_Profile>

</Reaction_Environment>

<Performance_Insights>

<FE_Matrix>not mentioned</FE_Matrix>

<Byproduct_Spectrum>not mentioned</Byproduct_Spectrum>

<Stability_Indicator>not mentioned</Stability_Indicator>

</Performance_Insights>

<Reaction_Mechanism>

<Reaction_Pathway>not mentioned</Reaction_Pathway>

<Adsorption_Energy>not mentioned</Adsorption_Energy>

<Reaction_Energy_Barrier>not mentioned</Reaction_Energy_Barrier>

<Electrochemical_Kinetics>not mentioned</Electrochemical_Kinetics>

</Reaction_Mechanism>

</Extraction_Schema>

<Classification_Guidelines>

<Catalyst_Type>

Identify the catalyst family only when it is explicitly stated or unambiguously described in the text.

Examples: "Cu-based", "FeCo alloy", "single-atom catalyst", "Co₃O₄-derived".

If the text does not clearly support a family label, output "not mentioned".

</Catalyst_Type>

<Synthesis_Method>

Extract the synthesis method only if explicitly stated.

Preserve the original wording, such as "hydrothermal method", "electrodeposition", or "microwave-assisted method".

If absent, output "not mentioned".

</Synthesis_Method>

<Regulation_Strategy>

Extract only strategies explicitly described in the text.

Examples: "alloying", "single-atom engineering", "defect engineering", "heteroatom doping", "interfacial electronic modulation".

Do not invent strategy labels from interpretation.

If multiple strategies are explicitly stated, list them as a semicolon-separated string.

If absent, output "not mentioned".

</Regulation_Strategy>

<Cell_Type>

Extract only if explicitly stated.

Allowed examples include "H-cell", "flow-cell", and "MEA".

If the cell configuration is not explicitly reported, output "not mentioned".

</Cell_Type>

<Electrolyte_Profile>

<Base_Electrolyte>

Extract the electrolyte composition and concentration only if explicitly reported.

Example: "1 M KOH", "0.1 M Na₂SO₄".

</Base_Electrolyte>

<Nitrate_Source>

Extract the nitrate source and concentration only if explicitly reported.

Example: "0.1 M KNO₃".

</Nitrate_Source>

<pH_Value>

Extract the pH only if explicitly reported in the text.

Do not calculate pH from other parameters.

</pH_Value>

</Electrolyte_Profile>

<FE_Matrix>

Extract Faradaic efficiency values only if explicitly reported.

Preserve the original format when possible.

Example: "-0.7 V: 98%; -0.5 V: 85.5%".

If the text reports only one value, extract only that value.

</FE_Matrix>

<Byproduct_Spectrum>

Extract only byproducts explicitly reported in the text.

Preserve the original descriptors and proportions if provided.

Example: "N₂ (<2%); NO₂⁻ trace".

If absent, output "not mentioned".

</Byproduct_Spectrum>

<Stability_Indicator>

Extract only explicitly reported stability metrics, such as current retention, cycle number, or testing duration.

Example: "95% current retention after 30 cycles".

</Stability_Indicator>

<Reaction_Pathway>

Extract only pathways explicitly reported in the text.

Preserve the original symbolic notation and ordering.

Do not infer missing intermediates or mechanistic steps.

If the text discusses a proposed pathway rather than a confirmed one, still extract it as reported, without judgment.

</Reaction_Pathway>

<Adsorption_Energy>

Extract adsorption energies only if explicitly reported.

Example: " *NO_2 : -1.2 eV; *NH_2 : -0.8 eV".

</Adsorption_Energy>

<Reaction_Energy_Barrier>

Extract activation barriers or reaction barriers only if explicitly reported.

Example: " *NO_3^- *NO_2 : 0.45 eV".

</Reaction_Energy_Barrier>

<Electrochemical_Kinetics>

Extract kinetic descriptors only if explicitly reported.

Examples: "Tafel slope: 38 mV dec⁻¹ at -0.5 V", "R_{ct}: 12 ohm".

</Electrochemical_Kinetics>

</Classification_Guidelines>

<Conflict_Handling>

If the same catalyst is reported with different values in different excerpts, extract only the values present in the current input text.

Do not resolve conflicts across documents.

Do not choose a preferred value unless the source text explicitly indicates one.

</Conflict_Handling>

<Quality_Control>

Before finalizing the output, verify that:

1. Every field in the schema is present.
2. All values are directly supported by the source text.
3. No inferred, normalized, or speculative content has been added.

4. Every missing field is marked "not mentioned".

</Quality_Control>

<Example_Input>

Cu single-atom catalyst (Cu-SAs/NC) synthesized by a microwave-assisted method exhibits excellent performance in the flow cell. The Faradaic efficiency of NH_3 reaches 98% at -0.7 V, with trace N_2 (<2%) as the main byproduct. Activity is enhanced via interfacial electronic modulation and defect engineering.

</Example_Input>

<Example_Output>

<Advanced_eNO₃RR_Extraction>

<Intelligent_Induction>

<Catalyst_Type>Cu-based</Catalyst_Type>

<Synthesis_Method>microwave-assisted method</Synthesis_Method>

<Regulation_Strategy>single-atom catalyst; interfacial electronic modulation; defect engineering</Regulation_Strategy>

</Intelligent_Induction>

<Catalyst_Properties>

<Composition>Cu-SAs/NC</Composition>

<Architecture>single-atom catalyst</Architecture>

<Defect_Engineering>defect engineering</Defect_Engineering>

</Catalyst_Properties>

<Reaction_Environment>

<Cell_Type>flow-cell</Cell_Type>

<Electrolyte_Profile>

<Base_Electrolyte>not mentioned</Base_Electrolyte>

<Nitrate_Source>not mentioned</Nitrate_Source>

<pH_Value>not mentioned</pH_Value>

</Electrolyte_Profile>

</Reaction_Environment>

<Performance_Insights>

<FE_Matrix>-0.7 V: 98%</FE_Matrix>

<Byproduct_Spectrum> N_2 (<2%)</Byproduct_Spectrum>

<Stability_Indicator>not mentioned</Stability_Indicator>

</Performance_Insights>

<Reaction_Mechanism>

<Reaction_Pathway>not mentioned</Reaction_Pathway>

<Adsorption_Energy>not mentioned</Adsorption_Energy>

<Reaction_Energy_Barrier>not mentioned</Reaction_Energy_Barrier>

<Electrochemical_Kinetics>not mentioned</Electrochemical_Kinetics>

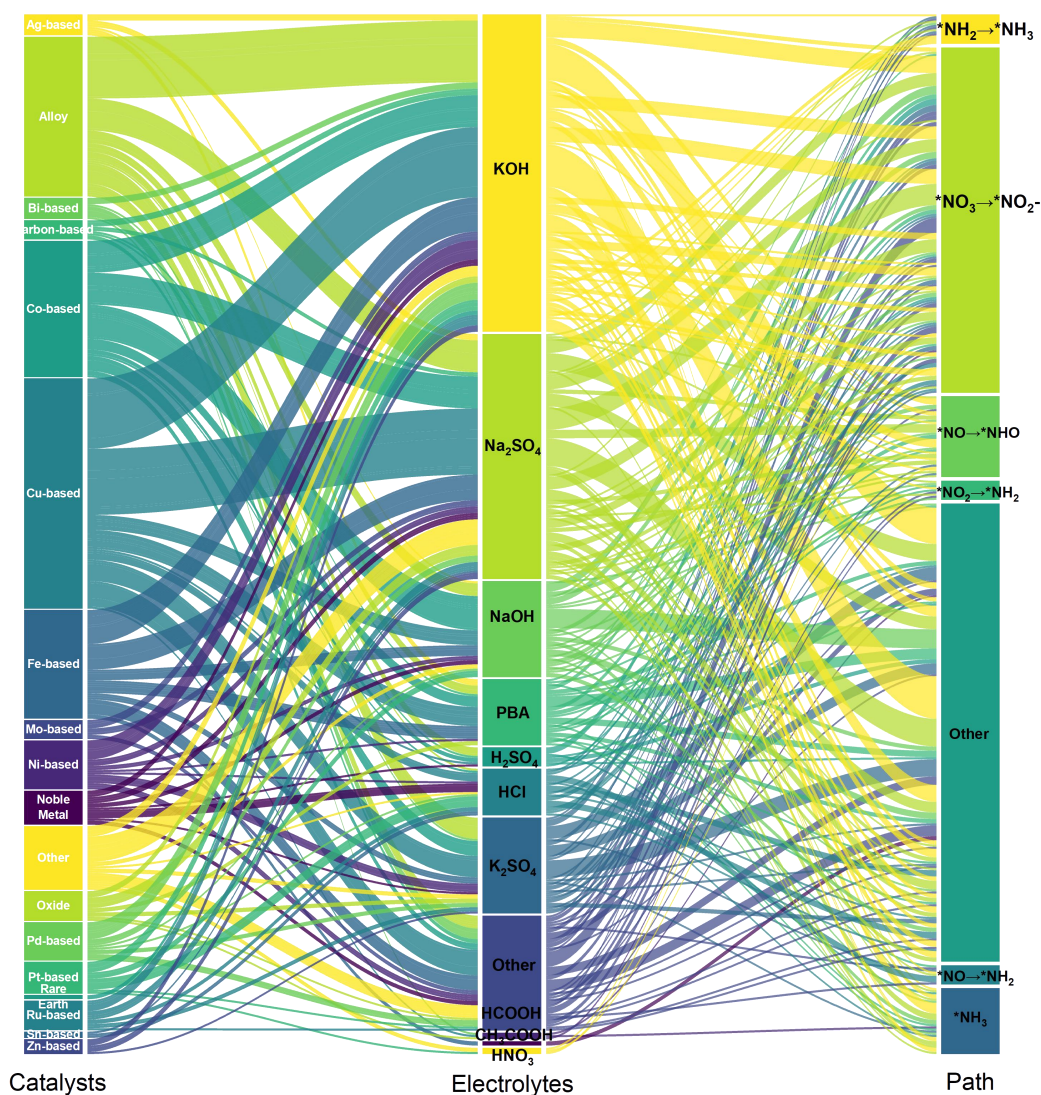
</Reaction_Mechanism>

</Advanced_eNO₃RR_Extraction>

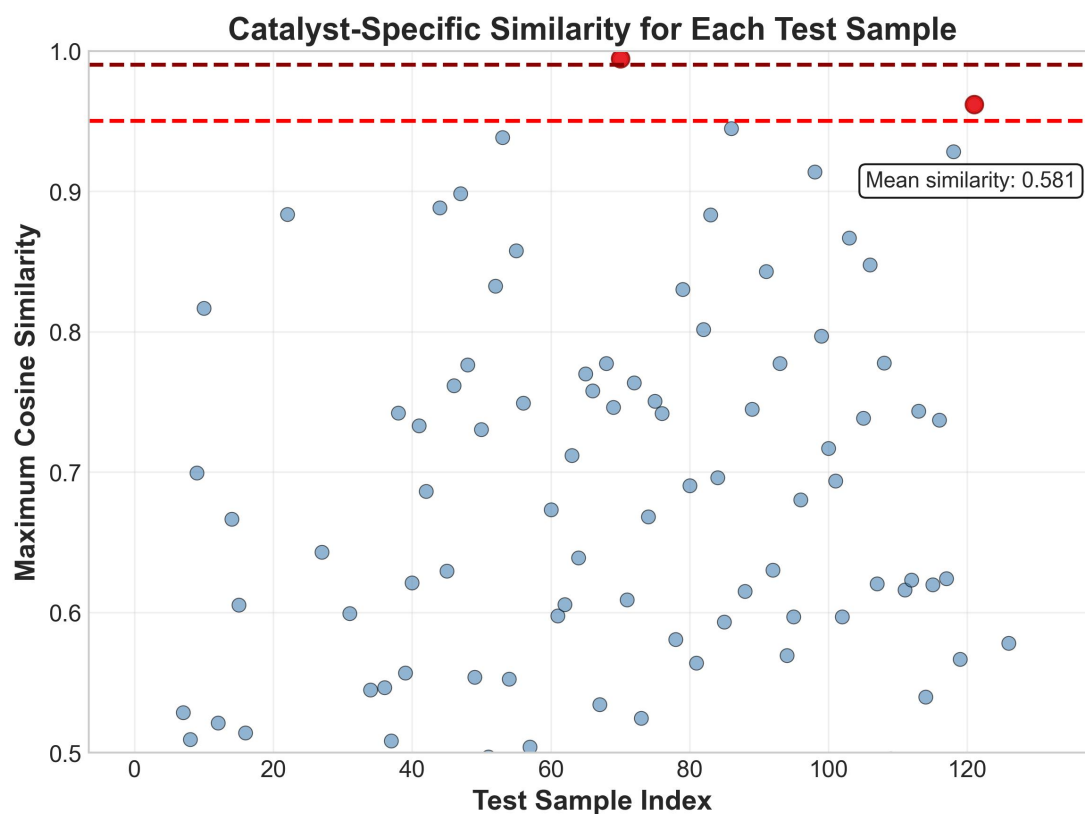
</Example_Output>

</Advanced_eNO₃RR_Extraction>

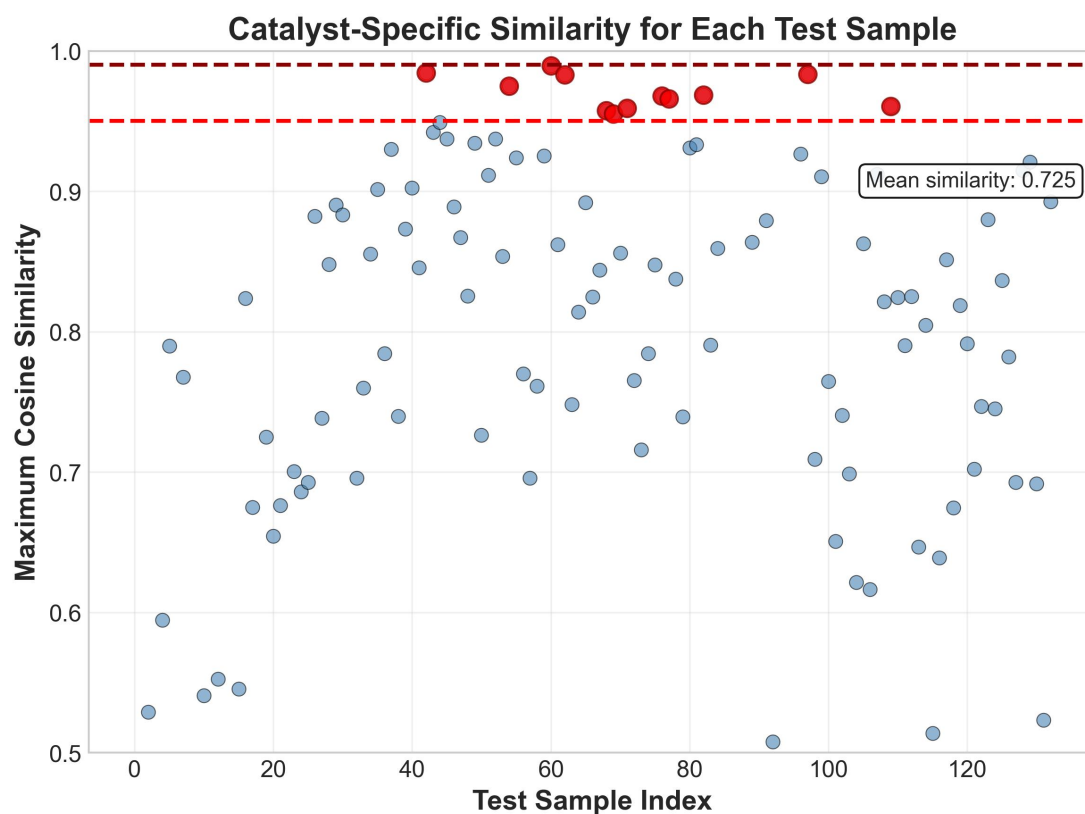
Supplementary Figures



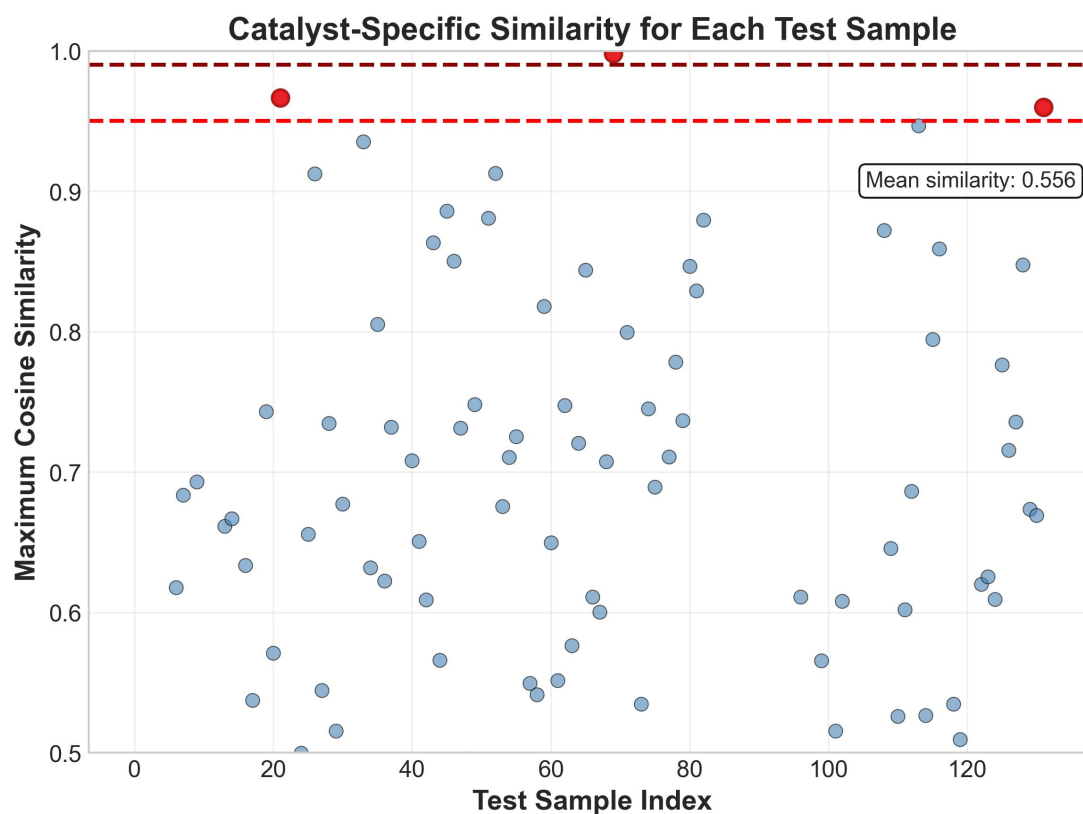
Supplementary Figure 1. A mapping of catalyst type, electrolyte environment, and nitrogen-containing reaction pathways. The Sankey diagram shows the distribution of reported relationships among catalyst categories, electrolyte systems, and nitrogen-containing reaction pathways, with emphasis on Co-based catalyst families within the broader dataset. The left, middle, and right columns represent catalyst types, electrolytes, and reaction pathways, respectively, and ribbon width indicates the relative frequency of each connection. This figure was generated using Origin under a valid institutional license.



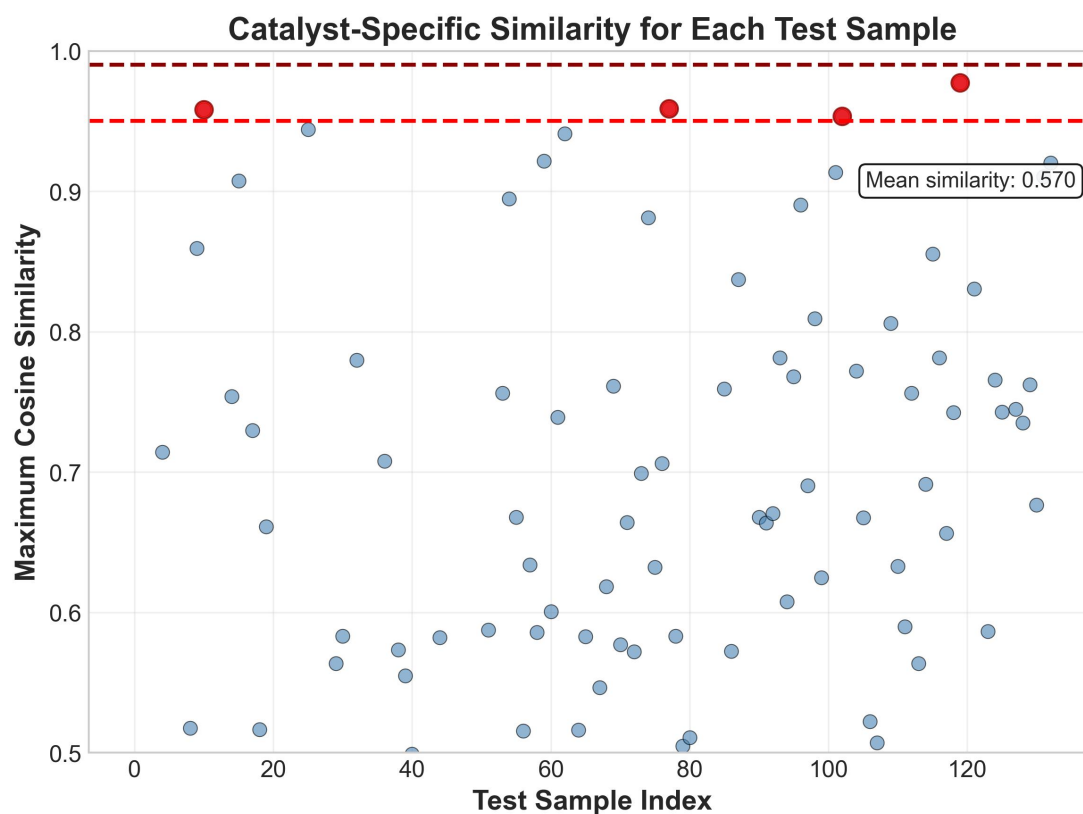
Supplementary Figure 2. MaterialsBERT-based similarity analysis between training and test samples. The scatter plot shows the maximum cosine similarity between each test sample and the training set, calculated using the MaterialsBERT encoder. Blue circles represent individual test samples, and red circles highlight test samples with very high similarity to training samples. The red dashed horizontal lines indicate high-similarity thresholds used to identify potentially overlapping samples.



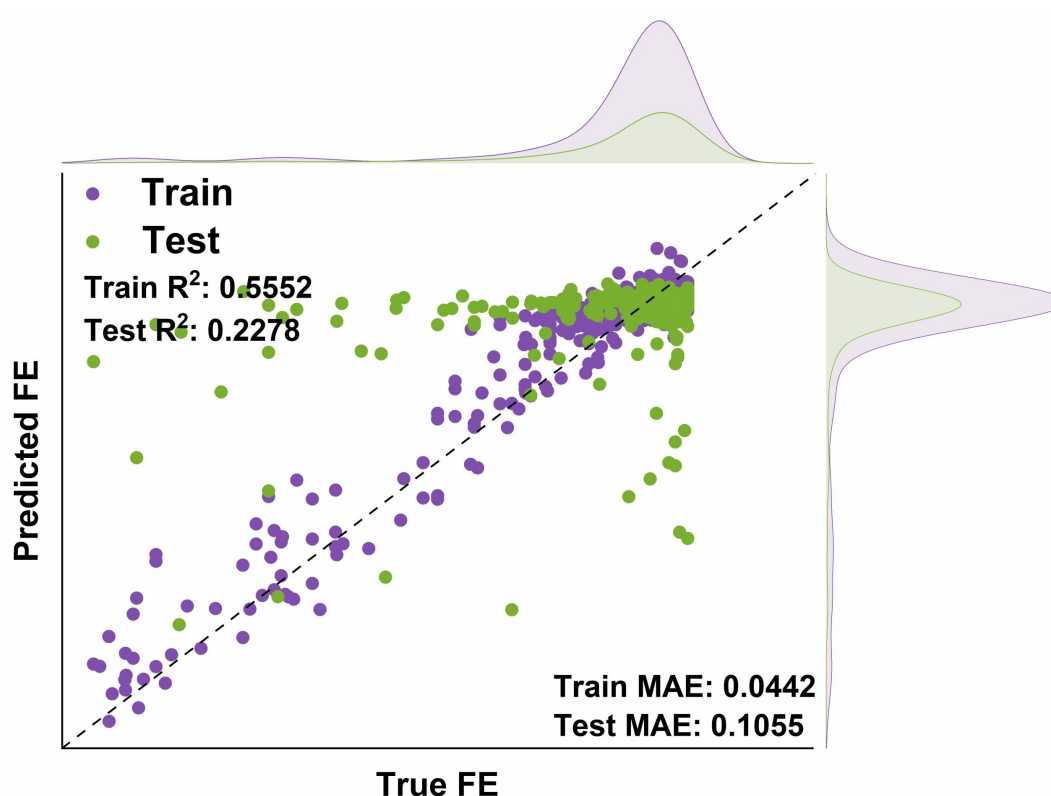
Supplementary Figure 3. MatBERT-based similarity analysis between training and test samples. The scatter plot shows the maximum cosine similarity between each test sample and the training set, calculated using the MatBERT encoder. Blue circles represent individual test samples, and red circles highlight test samples with very high similarity to training samples. The red dashed horizontal lines indicate high-similarity thresholds used to identify potentially overlapping samples.



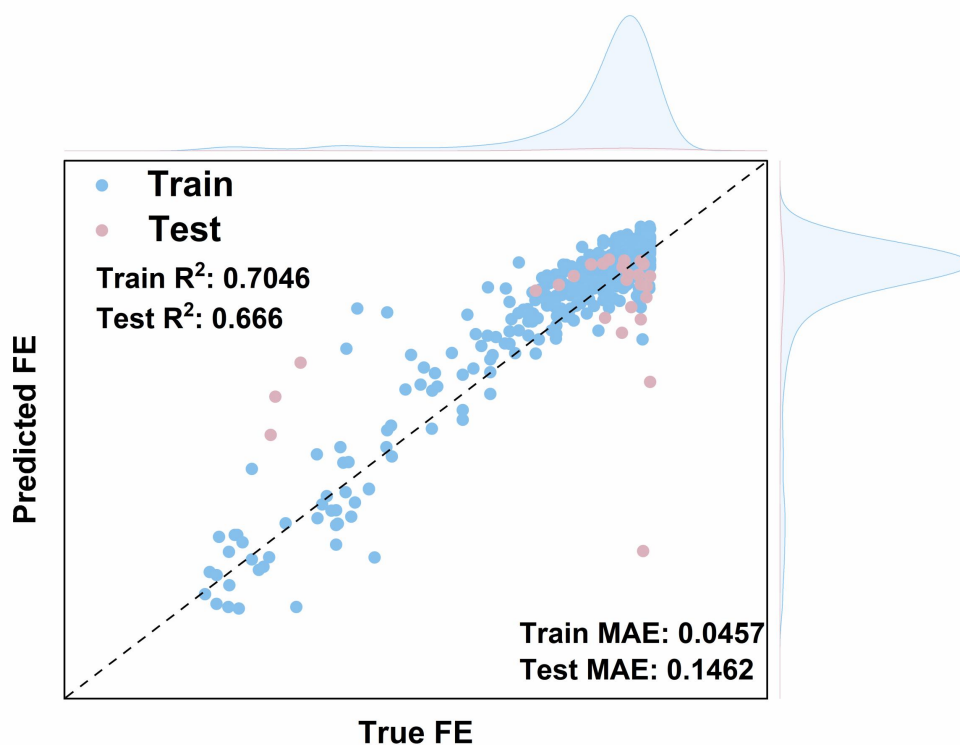
Supplementary Figure 4. Text-embedding-3-large-based similarity analysis between training and test samples. The scatter plot shows the maximum cosine similarity between each test sample and the training set, calculated using the Text-embedding-3-large encoder. Blue circles represent individual test samples, and red circles highlight test samples with very high similarity to training samples. The red dashed horizontal lines indicate high-similarity thresholds used to identify potentially overlapping samples.



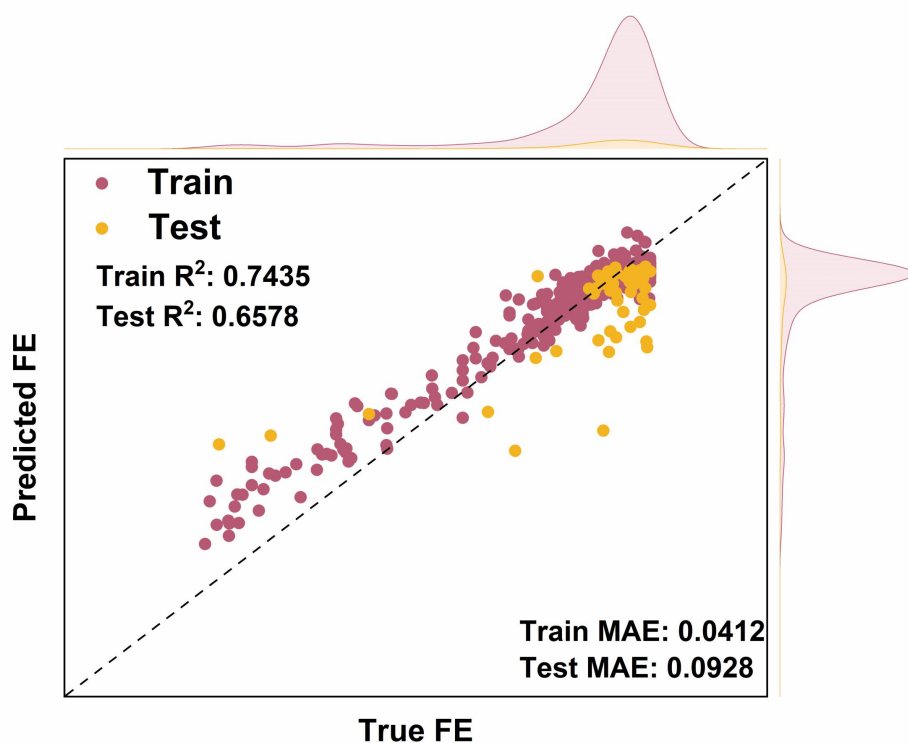
Supplementary Figure 5. SciBERT-based similarity analysis between training and test samples. The scatter plot shows the maximum cosine similarity between each test sample and the training set, calculated using the SciBERT encoder. Blue circles represent individual test samples, and red circles highlight test samples with very high similarity to training samples. The red dashed horizontal lines indicate high-similarity thresholds used to identify potentially overlapping samples.



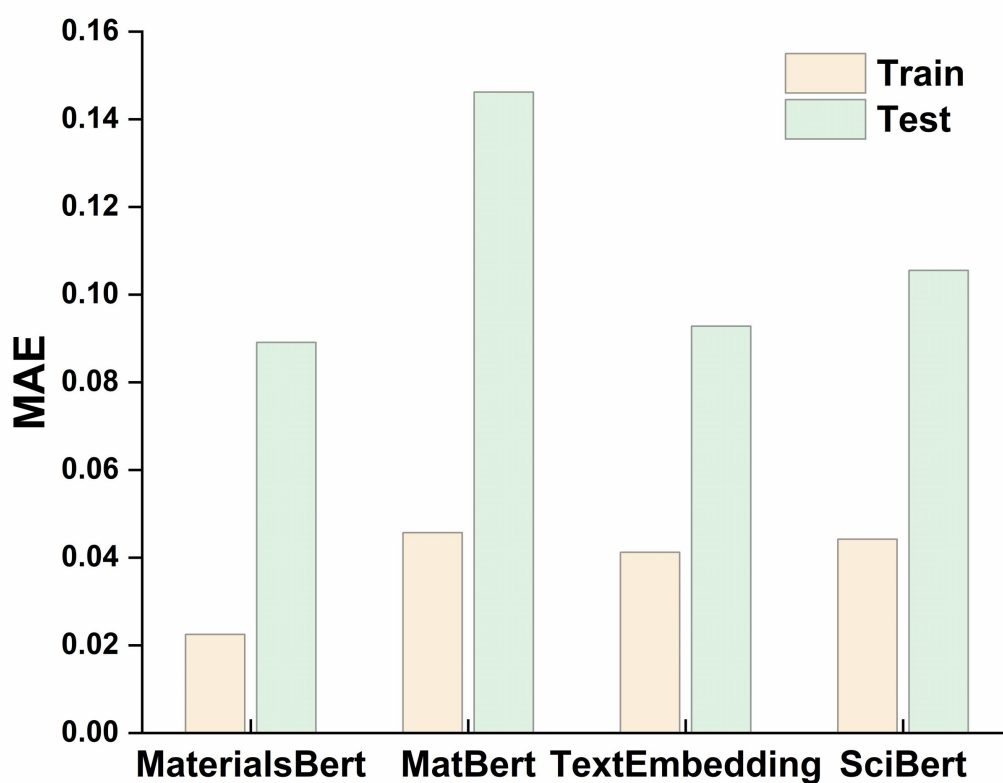
Supplementary Figure 6. SciBERT-ReactionLSTM performance for predicting faradaic efficiency. The parity plot compares predicted and true faradaic efficiency values for the model using the SciBERT encoder and ReactionLSTM module. Purple circles represent training samples, and green circles represent test samples. The black dashed diagonal line indicates ideal agreement between predicted and true values. Marginal density curves along the top and right axes show the distributions of true and predicted faradaic efficiency values, respectively, for the training and test sets. This figure was generated using Origin under a valid institutional license.



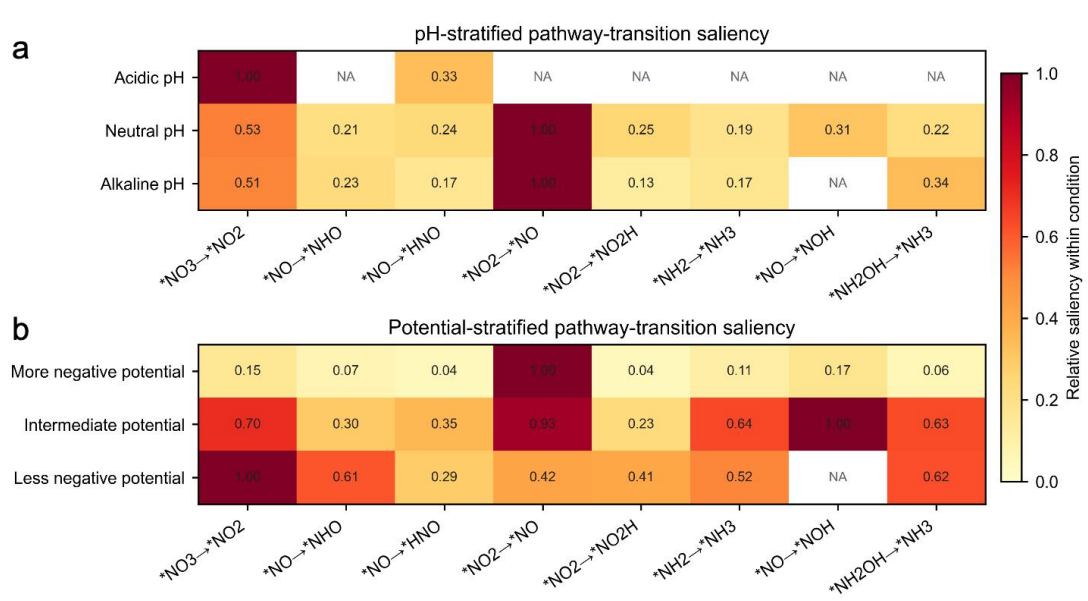
Supplementary Figure 7. MatBERT-ReactionLSTM performance for predicting faradaic efficiency. The parity plot compares predicted and true faradaic efficiency values for the model using the MatBERT encoder and ReactionLSTM module. Blue circles represent training samples, and pink circles represent test samples. The black dashed diagonal line indicates ideal agreement between predicted and true values. Marginal density curves along the top and right axes show the distributions of true and predicted faradaic efficiency values, respectively, for the training and test sets. This figure was generated using Origin under a valid institutional license.



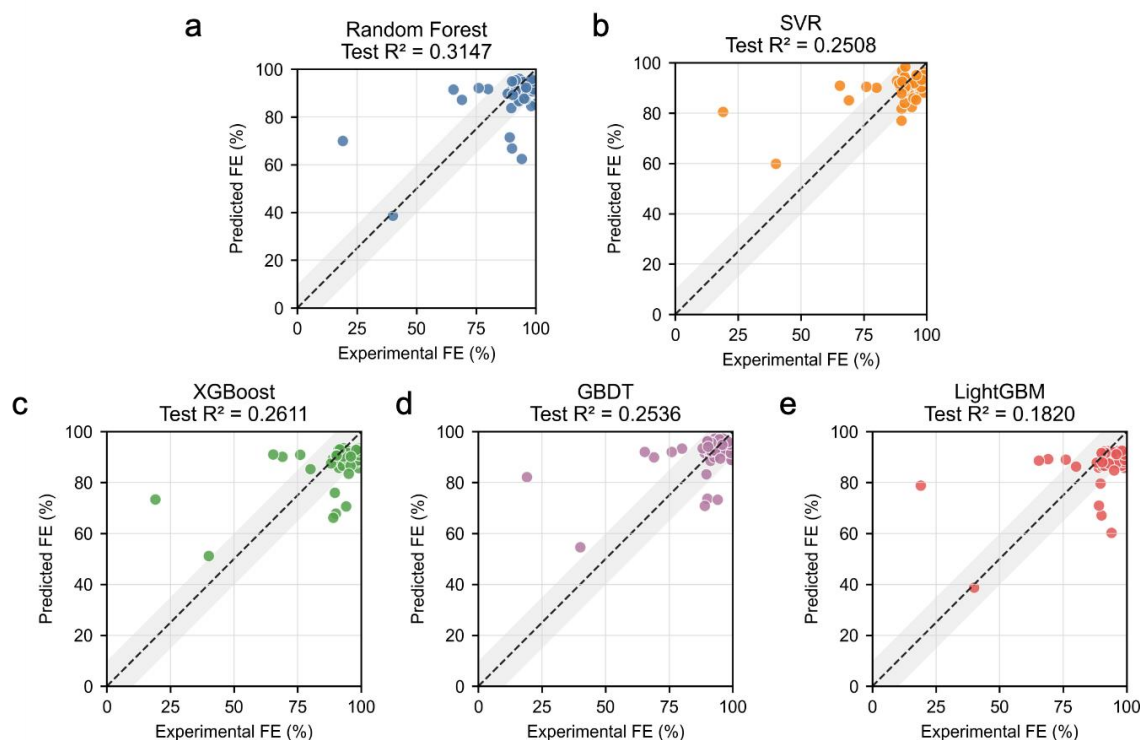
Supplementary Figure 8. Text-embedding-3-large-ReactionLSTM performance for predicting faradaic efficiency. The parity plot compares predicted and true faradaic efficiency values for the model using the Text-embedding-3-large encoder and ReactionLSTM module. Red circles represent training samples, and orange circles represent test samples. The black dashed diagonal line indicates ideal agreement between predicted and true values. Marginal density curves along the top and right axes show the distributions of true and predicted faradaic efficiency values, respectively, for the training and test sets. This figure was generated using Origin under a valid institutional license.



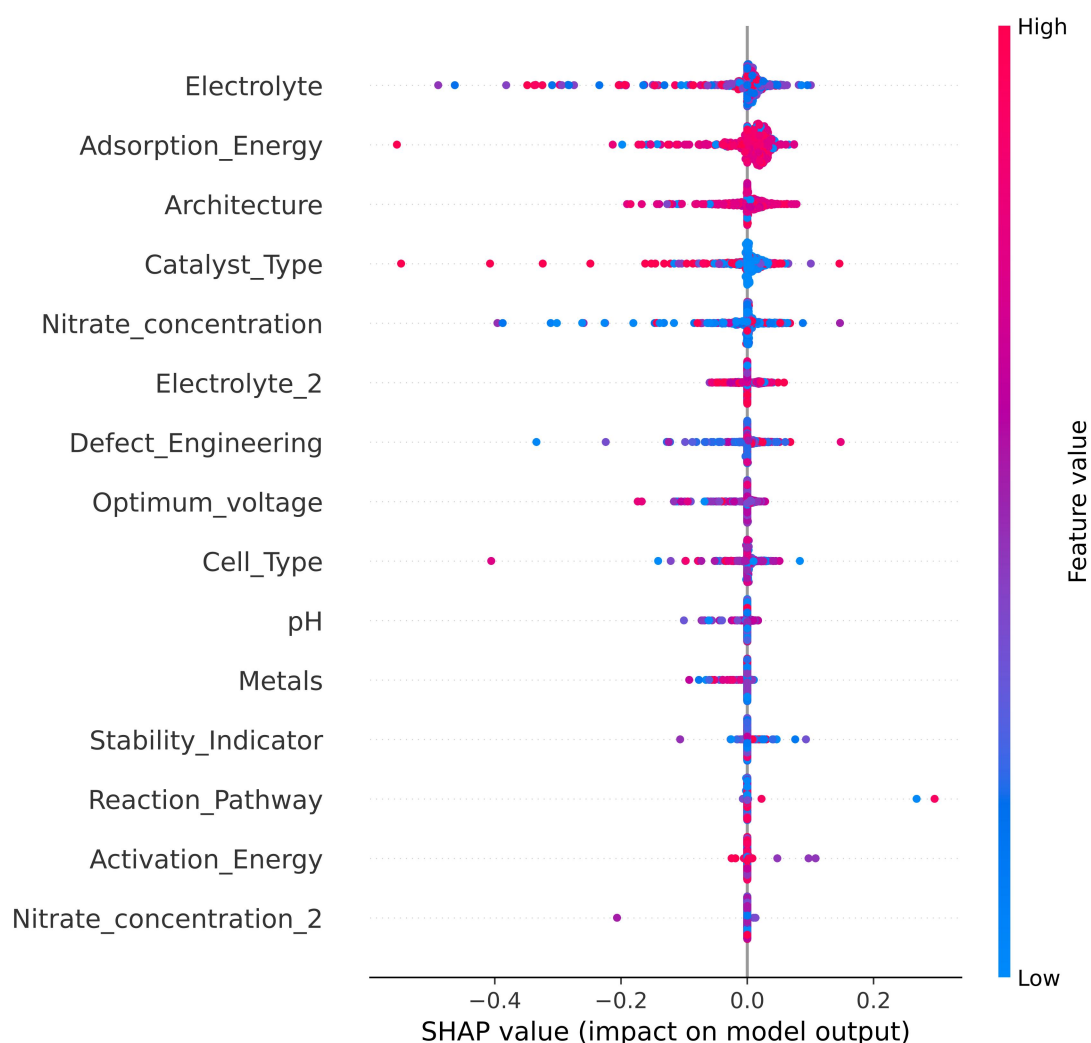
Supplementary Figure 9. MAE comparison across semantic encoder variants. The bar chart compares the training and test MAE values of model variants using different semantic encoders, including MaterialsBERT, MatBERT, Text-embedding-3-large, and SciBERT. Beige bars represent the training set, and light green bars represent the test set. This figure was generated using Origin under a valid institutional license.



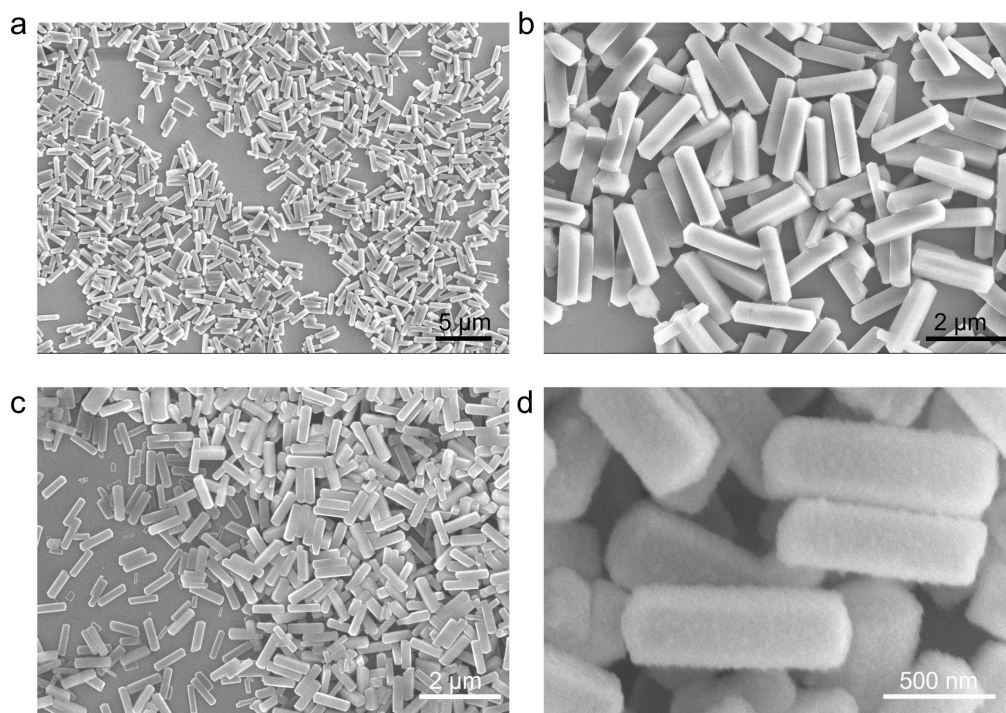
Supplementary Figure 10. Condition-stratified saliency analysis identifies reaction-pathway transitions with condition-dependent influence on model predictions. (a) Heatmap showing the relative prediction sensitivity to occluding specific reaction-pathway transitions under acidic, neutral, and alkaline pH conditions. (b) Heatmap showing the relative prediction sensitivity to occluding the same reaction-pathway transitions under more negative, intermediate, and less negative applied-potential conditions. Values are normalized within each condition group, with darker red indicating higher relative saliency and lighter yellow indicating lower relative saliency. White cells labeled “NA” indicate transition-condition pairs with fewer than three samples.



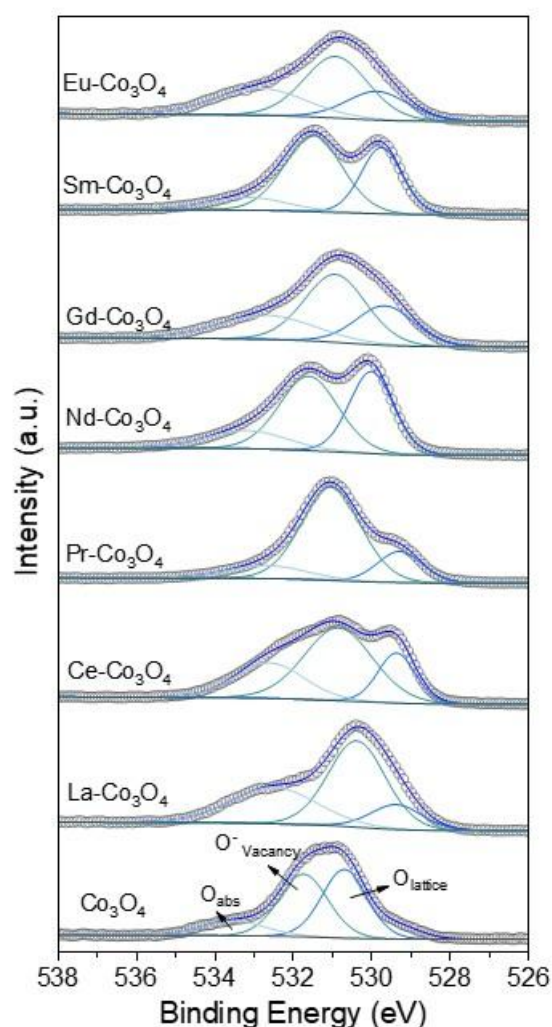
Supplementary Figure 11. Traditional machine-learning baseline models show limited prediction accuracy on the grouped test set. Parity plots compare predicted and experimental faradaic efficiency values for five traditional machine-learning baseline models on the grouped test set. (a) Random Forest model, shown with blue circles. (b) Support vector regression model, shown with orange circles. (c) XGBoost model, shown with green circles. (d) Gradient boosting decision tree model, shown with purple circles. (e) LightGBM model, shown with red circles. The black dashed diagonal line indicates perfect agreement between predicted and experimental values, and the gray shaded band indicates a deviation of plus or minus 10 percentage points from this line.



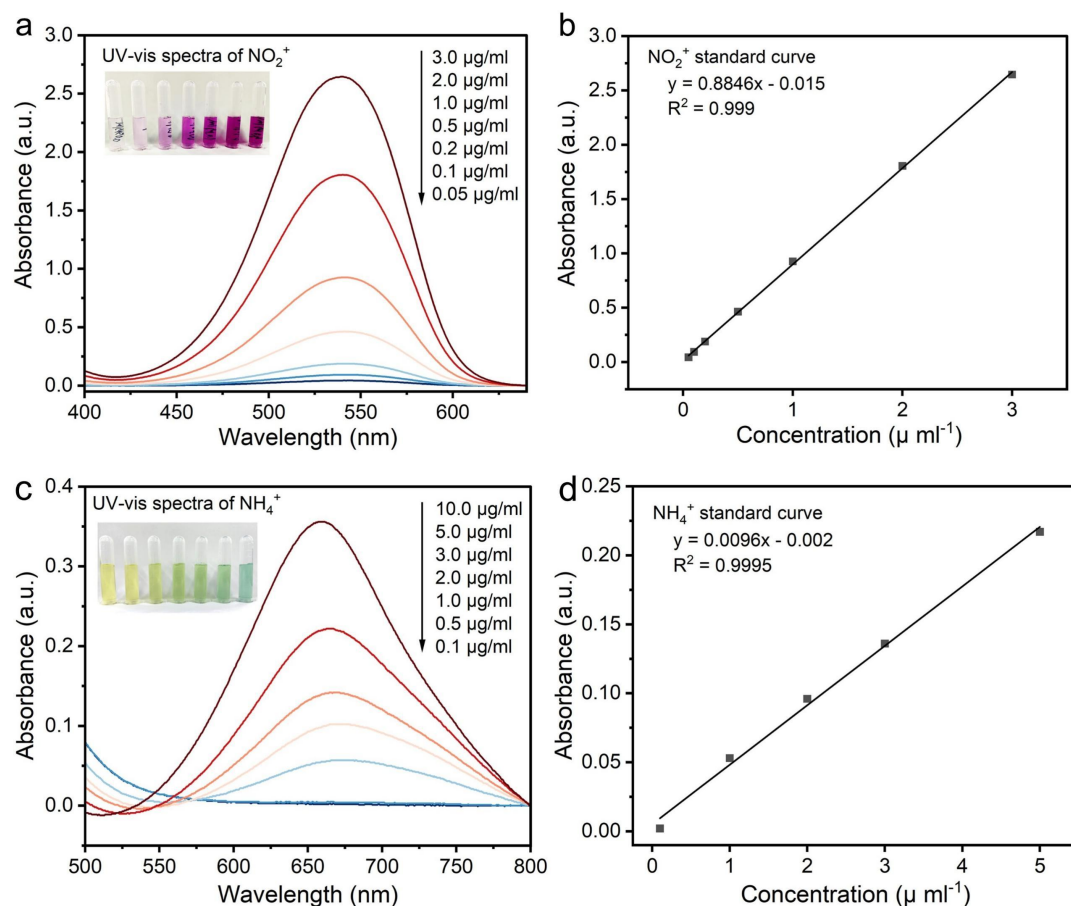
Supplementary Figure 12. SHAP feature importance analysis chart of the Co-Based catalyst prediction model. The SHAP summary plot shows the relative contribution of input features to the Co-based catalyst prediction model. Features are ranked from top to bottom according to their overall importance, with electrolyte showing the largest impact on model output. Each point represents one sample, and its horizontal position indicates the SHAP value, where larger absolute values correspond to stronger effects on the prediction. The color scale represents feature value, with blue indicating low values and red indicating high values.



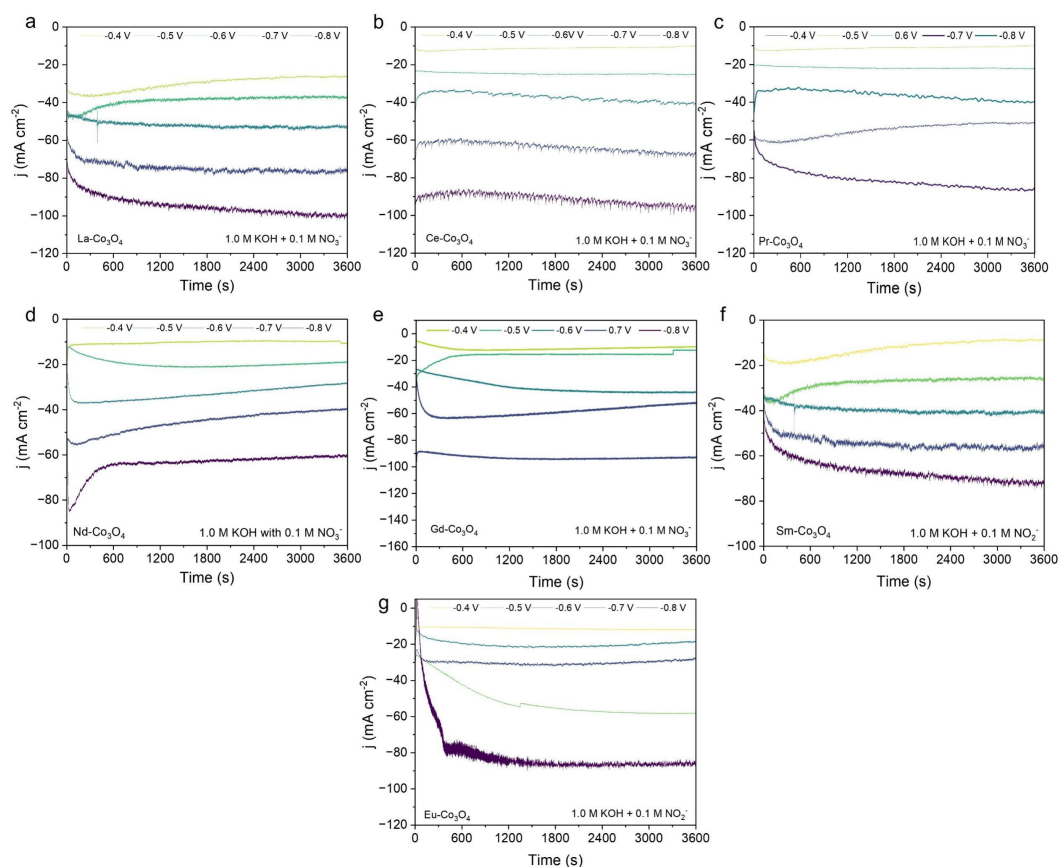
Supplementary Figure 13. SEM images of Co_3O_4 and $\text{La-Co}_3\text{O}_4$. Scanning electron microscopy images of (a), (b) Co_3O_4 and (c), (d) $\text{La-Co}_3\text{O}_4$ at different magnifications. Both samples exhibit uniform nanorod structures. Scale bars are 5 μm in (a), 2 μm in (b) and (c), and 500 nm in (d).



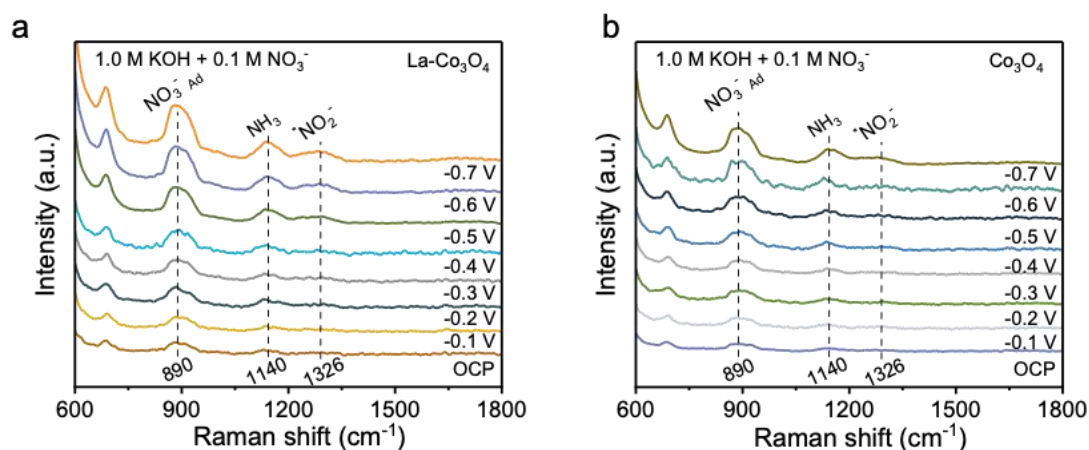
Supplementary Figure 14. O 1s XPS spectra reveal vacancy-related oxygen species in RE-Co₃O₄. O 1s X-ray photoelectron spectroscopy spectra of Co₃O₄ and RE-Co₃O₄ samples, where RE represents La, Ce, Pr, Nd, Gd, Sm, or Eu. The fitted components correspond to lattice oxygen, oxygen-vacancy-related oxygen, and adsorbed oxygen species. The open circles represent experimental spectra, the dark blue lines represent fitted envelope curves, and the light blue component peaks represent deconvoluted oxygen species. This figure was generated using Origin under a valid institutional license.



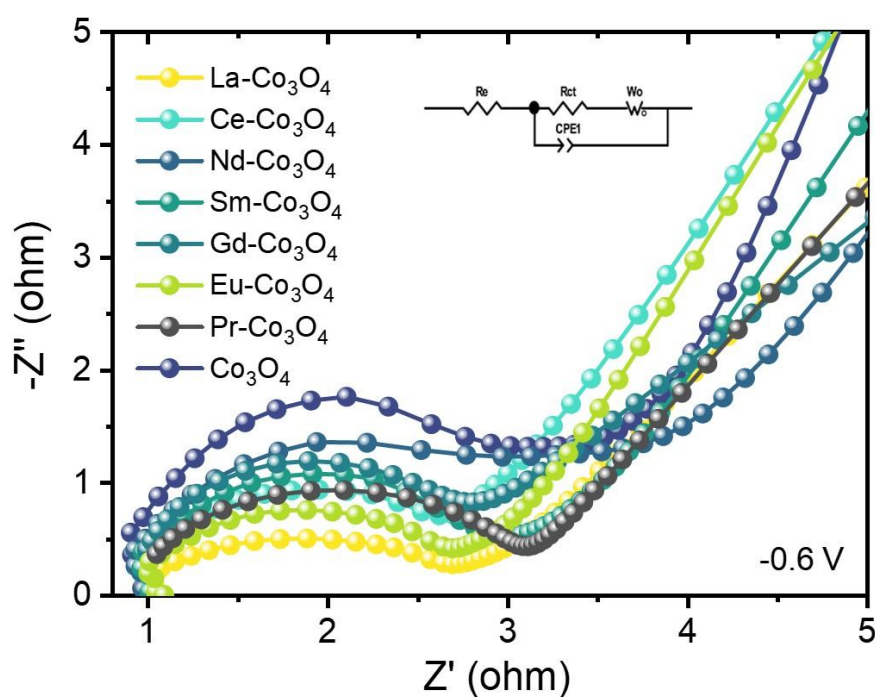
Supplementary Figure 15. UV-vis calibration for nitrite and ammonium quantification. (a) The corresponding UV-Vis adsorption spectra of NO_2^- . (b) A calibration curve for colorimetric nitrite assay using Griess reagent. (c) The corresponding UV-Vis adsorption spectra of NH_4^+ . (d) A calibration curve for colorimetric ammonium assay using the indophenol blue method. This figure was generated using Origin under a valid institutional license.



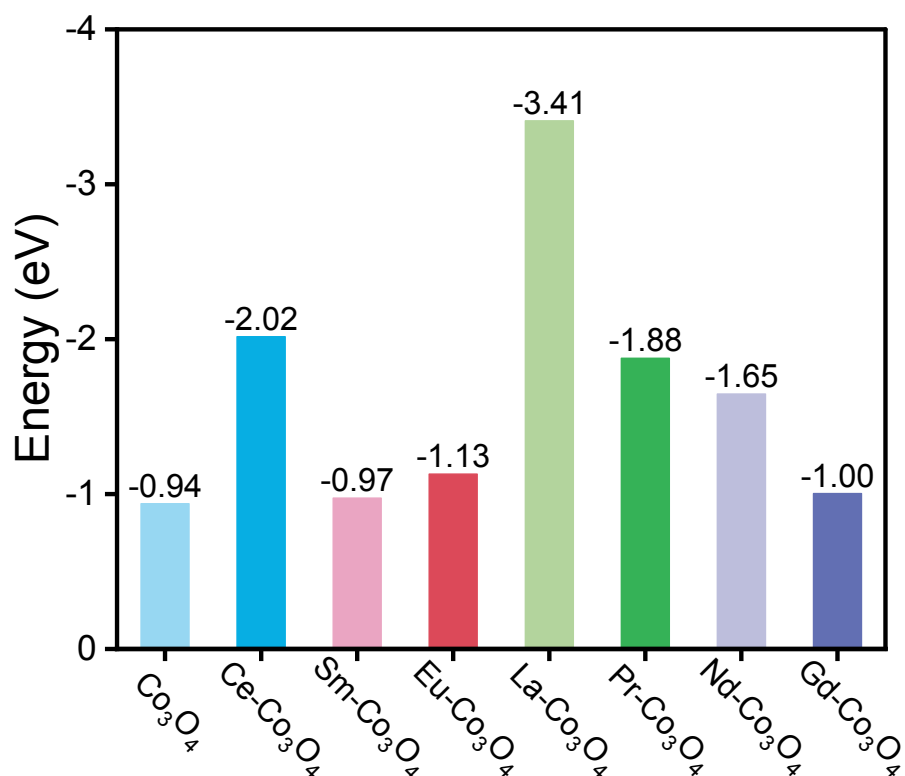
Supplementary Figure 16. Cyclic voltammetry responses of RE-Co₃O₄ in alkaline nitrite electrolyte. Cyclic voltammetry curves of RE-Co₃O₄ measured in 1.0 M KOH with 0.1 M NO₂⁻. (a) La-Co₃O₄, (b) Ce-Co₃O₄, (c) Pr-Co₃O₄, (d) Nd-Co₃O₄, (e) Gd-Co₃O₄, (f) Sm-Co₃O₄, and (g) Eu-Co₃O₄. This figure was generated using Origin under a valid institutional license.



Supplementary Figure 17. *In situ* Raman spectra of La-Co₃O₄ and Co₃O₄ during nitrate reduction. Electrochemical in situ Raman spectra recorded in 1.0 M KOH with 0.1 M NO₃⁻. (a) La-Co₃O₄ and (b) Co₃O₄ measured under different applied potentials. The colored spectra correspond to the potential sequence shown in each panel. This figure was generated using Origin under a valid institutional license.

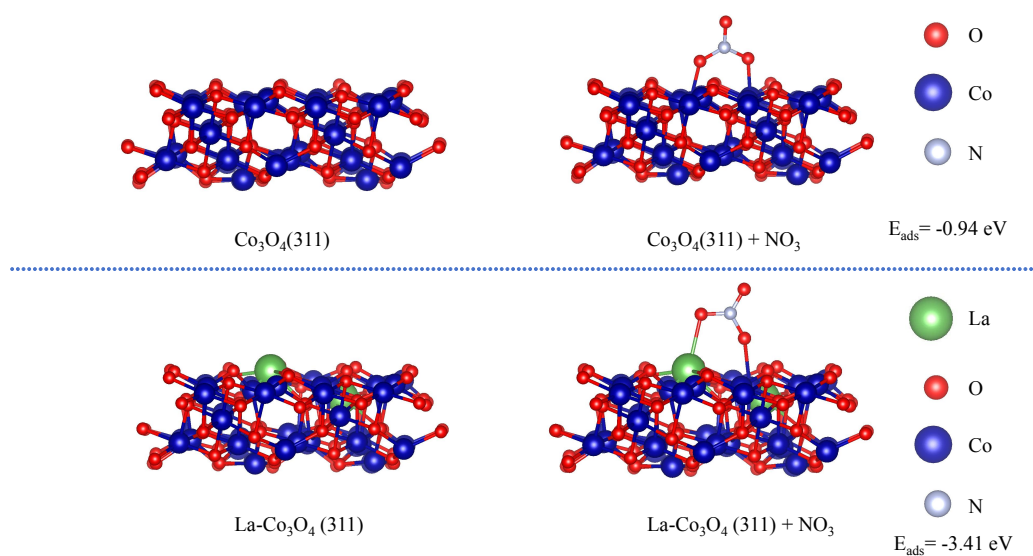


Supplementary Figure 18. Nyquist plots of RE-Co₃O₄ at -0.6 V versus RHE. Electrochemical impedance spectroscopy Nyquist plots of Co₃O₄ and RE-Co₃O₄ samples measured at -0.6 V versus the reversible hydrogen electrode. Colored circle-marked curves represent La-Co₃O₄, Ce-Co₃O₄, Nd-Co₃O₄, Sm-Co₃O₄, Gd-Co₃O₄, Eu-Co₃O₄, Pr-Co₃O₄, and Co₃O₄, as indicated in the legend. The inset shows the equivalent circuit used for fitting the impedance spectra. This figure was generated using Origin under a valid institutional license.

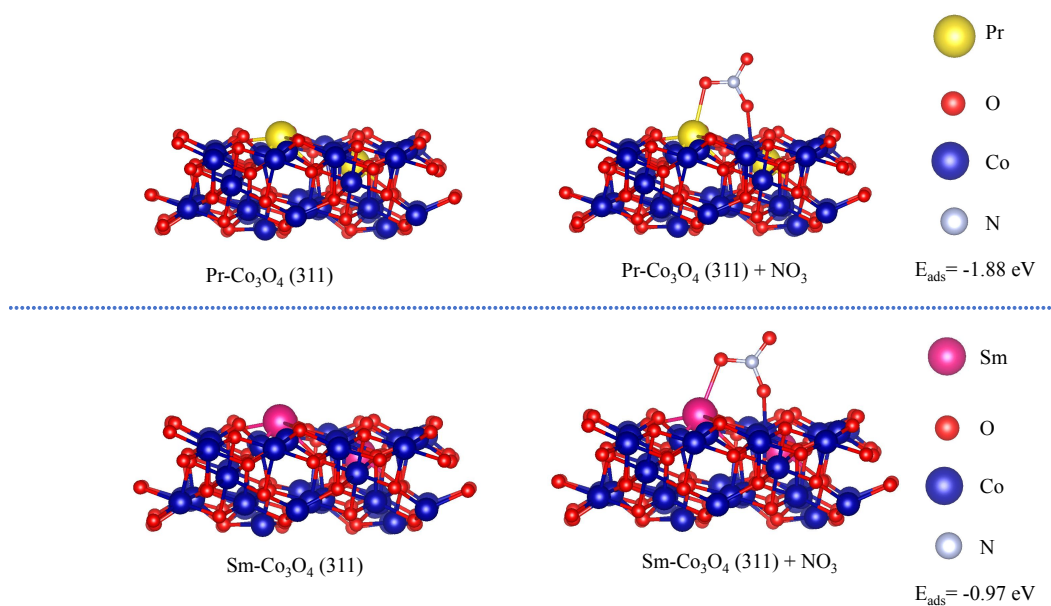


Supplementary Figure 19. RE doping modulates nitrate adsorption energies on Co_3O_4 models.

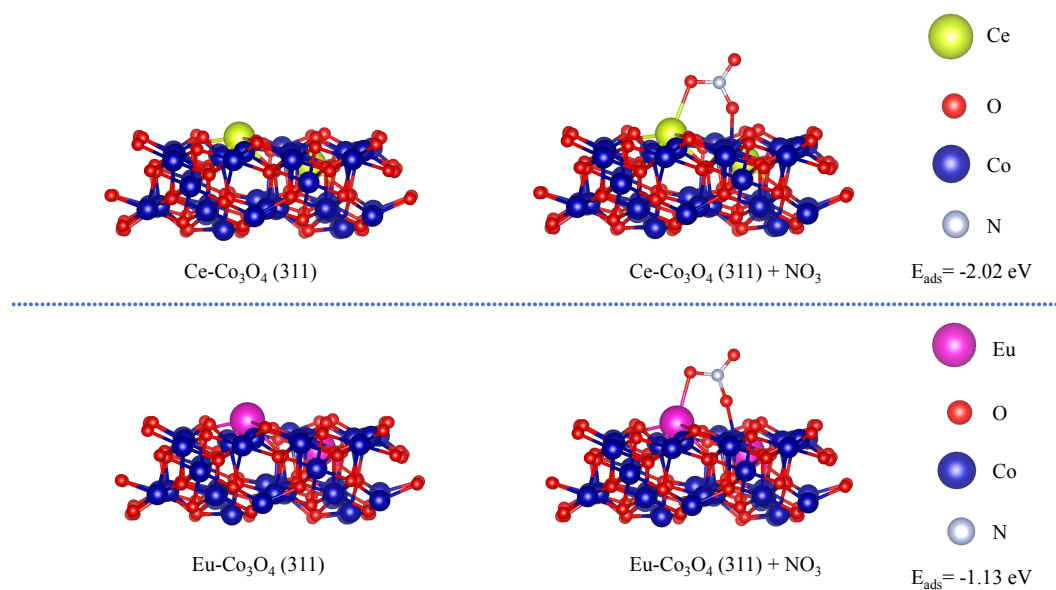
The bar chart compares the calculated NO_3^- adsorption energies on pristine Co_3O_4 and RE-doped Co_3O_4 models. Each colored bar represents one catalyst model, including pristine Co_3O_4 and Co_3O_4 doped with La, Ce, Pr, Nd, Sm, Eu, or Gd, as indicated on the x-axis. This figure was generated using Origin under a valid institutional license.



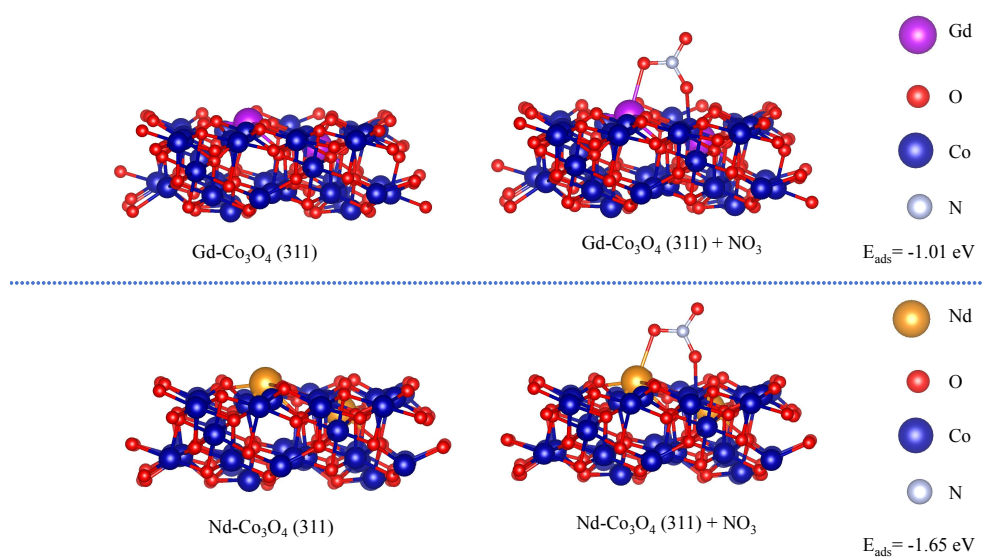
Supplementary Figure 20. Optimized nitrate adsorption configurations on pristine and La-doped Co₃O₄ surfaces. Optimized structural models showing NO₃⁻ adsorption on pristine Co₃O₄ (311) and La-Co₃O₄ (311) surfaces. The upper row shows pristine Co₃O₄ before and after NO₃⁻ adsorption, and the lower row shows La-Co₃O₄ before and after NO₃⁻ adsorption.



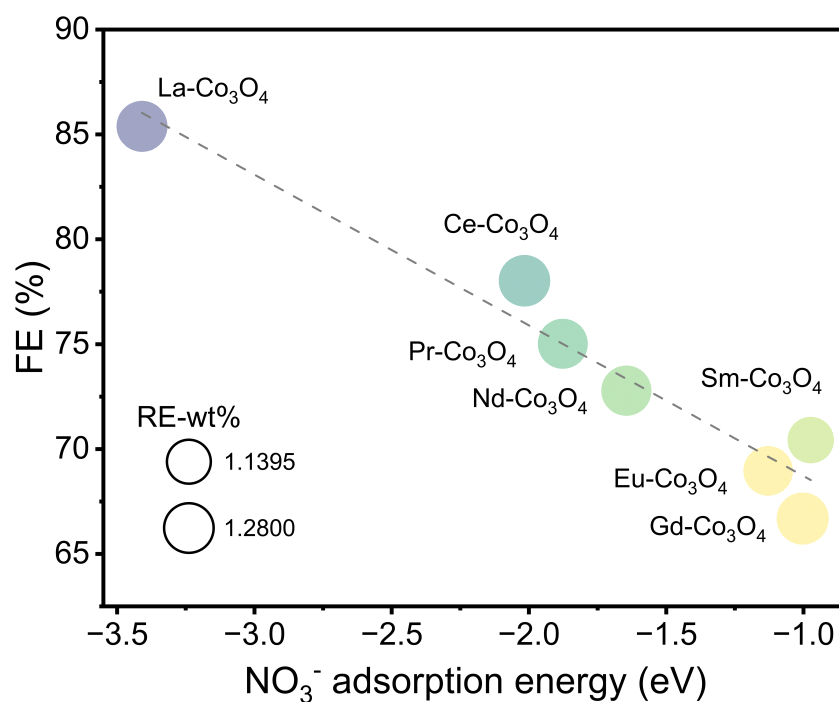
Supplementary Figure 21. Optimized nitrate adsorption configurations on Pr- and Sm-doped Co_3O_4 surfaces. Optimized structural models showing NO_3^- adsorption on Pr- Co_3O_4 (311) and Sm- Co_3O_4 (311) surfaces. The upper row shows Pr- Co_3O_4 before and after NO_3^- adsorption, and the lower row shows Sm- Co_3O_4 before and after NO_3^- adsorption.



Supplementary Figure 22. Optimized nitrate adsorption configurations on Ce- and Eu-doped Co_3O_4 surfaces. Optimized structural models showing NO_3^- adsorption on $\text{Ce-Co}_3\text{O}_4$ (311) and $\text{Eu-Co}_3\text{O}_4$ (311) surfaces. The upper row shows $\text{Ce-Co}_3\text{O}_4$ before and after NO_3^- adsorption, and the lower row shows $\text{Eu-Co}_3\text{O}_4$ before and after NO_3^- adsorption.



Supplementary Figure 23. Optimized nitrate adsorption configurations on Gd- and Nd-doped Co_3O_4 surfaces. Optimized structural models showing NO_3^- adsorption on $\text{Gd-Co}_3\text{O}_4$ (311) and $\text{Nd-Co}_3\text{O}_4$ (311) surfaces. The upper row shows $\text{Gd-Co}_3\text{O}_4$ before and after NO_3^- adsorption, and the lower row shows $\text{Nd-Co}_3\text{O}_4$ before and after NO_3^- adsorption.



Supplementary Figure 24. FE correlates with nitrate adsorption energy and RE content in RE- Co_3O_4 nanorods. The scatter plot shows the relationship between FE, calculated NO_3^- adsorption energy, and RE content for RE-doped Co_3O_4 nanorods. Each circle represents one RE- Co_3O_4 sample, with circle color indicating the RE content according to the color scale. The gray dashed line represents the fitted trend between faradaic efficiency and NO_3^- adsorption energy. This figure was generated using Origin under a valid institutional license.

Supplementary Tables

Supplementary Table 1. Comparison of feature encoding dimensions across different semantic encoders.

	Materialsbert	Matbert	Scibert	TextEmbedding
Catalyst Type Dimension	10	2	5	5
Architecture Dimension	10	1	20	20
Defect Engineering Dimension	10	1	20	25
Cell Type Dimension	5	1	5	10
Electrolyte Dimension	10	1	20	25
Nitrate Concentration Dimension	10	1	20	20
Adsorption Energy Dimension	10	1	20	25

Supplementary Table 2. Comparison of embedding and hidden dimensions of the ReactionLSTM Module across different encoder models.

	Materialsbert	Matbert	Scibert	TextEmbedding
Embedding Dimension	64	128	128	128
Hidden dimension	32	1	64	1

Supplementary Table 3. Input-Output dimensions and key parameter configurations of each layer in the deep neural network.

	Input Dimension	Output Dimension	Key Parameters
Layer 1	Input_dim	256	dropout_rate = 0.3
Layer 2	256	512	dropout_rate = 0.3
Layer 3	512	256	dropout_rate = 0.3
Layer 4	256	128	dropout_rate = 0.3
Layer 5	128	128	dropout_rate = 0.3
Layer 6	128	1	

Supplementary Table 4. Hyperparameter grid search configurations and optimal parameter sets for traditional machine learning baseline models.

Model	Parameter Space	Optimal Parameters
Random Forest	"n_estimators": [80, 120, 200, 350, 500, 700, 1000, 1400],	"n_estimators": 700,
	"criterion": ["squared_error", "absolute_error"],	"criterion": absolute_error,
	"max_depth": [None, 4, 8, 12, 16, 20, 26, 32, 40, 50, 65],	"max_depth": 40,
	"min_samples_split": [2, 3, 5, 8, 12, 18, 25, 35],	"min_samples_split": 5,
	"min_samples_leaf": [1, 2, 3, 4, 6, 8, 12, 16],	"min_samples_leaf": 3,
	"min_weight_fraction_leaf": [0.0, 0.01, 0.03, 0.06, 0.1],	"min_weight_fraction_leaf": 0.0,
	"max_features": ["sqrt", "log2", None, 0.12, 0.25, 0.4, 0.55, 0.7, 0.85, 1.0],	"max_features": 0.7,
	"max_leaf_nodes": [None, 31, 63, 127, 255, 511, 1023],	"max_leaf_nodes": 127,
	"min_impurity_decrease": [0.0, 1e-9, 1e-7, 1e-5, 1e-3],	"min_impurity_decrease": 1e-09,
	"bootstrap": [True],	"bootstrap": True,
	"max_samples": [None, 0.45, 0.6, 0.75, 0.88, 1.0],	"max_samples": 0.75,
	"ccp_alpha": [0.0, 1e-7, 5e-6, 1e-4, 5e-3, 0.02]	"ccp_alpha": 0.0
SVR	"kernel": ["rbf", "linear", "poly", "sigmoid"],	"kernel": sigmoid,
	"C": [0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0, 300.0, 1000.0],	"C": 0.3,
	"epsilon": [0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.15, 0.25, 0.4],	"epsilon": 0.002,
	"gamma": ["scale", "auto", 1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1.0, 10.0],	"gamma": scale,
	"degree": [2, 3, 4, 5],	"degree": 3,
	"coef0": [0.0, 0.01, 0.1, 0.5, 1.0],	"coef0": 0.01,
	"shrinking": [True, False],	"shrinking": True,
	"tol": [1e-5, 1e-4, 1e-3]	"tol": 1e-05
XGBoost	"n_estimators": [100, 200, 350, 500, 700, 1000, 1400],	"n_estimators": 200,
	"max_depth": [3, 4, 5, 6, 7, 8, 10, 12],	"max_depth": 6,
	"learning_rate": [0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.15, 0.2, 0.3],	"learning_rate": 0.02,
	"subsample": [0.55, 0.65, 0.75, 0.85, 0.95, 1.0],	"subsample": 0.65,
	"colsample_bytree": [0.55, 0.7, 0.85, 0.95, 1.0],	"colsample_bytree": 0.7,
	"colsample_bylevel": [0.6, 0.8, 1.0],	"colsample_bylevel": 1.0,
	"colsample_bynode": [0.6, 0.8, 1.0],	"colsample_bynode": 0.6,
	"min_child_weight": [1, 2, 3, 5, 7, 10, 15],	"min_child_weight": 1,
	"gamma": [0.0, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0],	"gamma": 0.1,
	"reg_alpha": [0.0, 1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1.0],	"reg_alpha": 0.1,
	"reg_lambda": [0.5, 1.0, 2.0, 5.0, 10.0, 20.0],	"reg_lambda": 1.0,
	"max_delta_step": [0, 1, 4],	"max_delta_step": 4,
GBDT	"tree_method": ["hist", "approx"]	"tree_method": approx
	"loss": ["squared_error", "absolute_error", "huber"],	"loss": huber,

LightGBM	"learning_rate": [0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.15, 0.2], "n_estimators": [100, 200, 350, 500, 700, 1000], "subsample": [0.65, 0.75, 0.85, 0.95, 1.0], "criterion": ["friedman_mse", "squared_error"], "max_depth": [3, 4, 5, 6, 7, 8, 10], "min_samples_split": [2, 4, 6, 10, 15, 20], "min_samples_leaf": [1, 2, 3, 4, 6, 8], "min_weight_fraction_leaf": [0.0, 0.01, 0.05, 0.1], "max_features": ["sqrt", "log2", None, 0.25, 0.5, 0.75, 1.0], "max_leaf_nodes": [None, 31, 63, 127, 255, 511], "min_impurity_decrease": [0.0, 1e-9, 1e-7, 1e-5], "ccp_alpha": [0.0, 1e-6, 1e-4, 1e-3], "alpha": [0.85, 0.9, 0.95] "boosting_type": ["gbdt", "dart"], "n_estimators": [100, 200, 350, 500, 700, 1000, 1400], "learning_rate": [0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.15, 0.2], "max_depth": [-1, 4, 6, 8, 10, 12], "num_leaves": [31, 63, 127, 255, 511], "min_child_samples": [5, 10, 15, 20, 40, 80], "subsample": [0.65, 0.8, 0.9, 1.0], "subsample_freq": [0, 1], "colsample_bytree": [0.65, 0.8, 0.9, 1.0], "reg_alpha": [0.0, 1e-4, 1e-3, 1e-2, 0.1, 1.0], "reg_lambda": [0.0, 1e-3, 1e-2, 0.1, 1.0, 10.0], "min_split_gain": [0.0, 0.01, 0.05, 0.1]	"learning_rate": 0.02, "n_estimators": 700, "subsample": 0.85, "criterion": squared_error, "max_depth": 3, "min_samples_split": 10, "min_samples_leaf": 2, "min_weight_fraction_leaf": 0.0, "max_features": 0.25, "max_leaf_nodes": 127, "min_impurity_decrease": 1e-09, "ccp_alpha": 0.0, "alpha": 0.85 "boosting_type": gbdt, "n_estimators": 1400, "learning_rate": 0.02, "max_depth": 6, "num_leaves": 31, "min_child_samples": 5, "subsample": 1.0, "subsample_freq": 1, "colsample_bytree": 1.0, "reg_alpha": 0.001, "reg_lambda": 0.1, "min_split_gain": 0.1
----------	---	---

Supplementary Table 5. Performance comparison of the MatSemNet model and traditional machine learning baseline models.

Model	Train R ²	Test R ²	Train MAE	Test MAE
MatSemNet	0.8203	0.7270	0.0225	0.0891
Random Forest	0.5680	0.3147	0.0701	0.0687
SVR	0.5358	0.2508	0.0540	0.0726
XGBoost	0.5808	0.2611	0.0836	0.0727
GBDT	0.5511	0.2536	0.0648	0.0672
LightGBM	0.4231	0.1820	0.0939	0.0766

Supplementary Table 6. Coverage of RE elements and exact RE-doped spinel Co₃O₄ compositions in the dataset.

RE element	Occurrence in training set	Simple RE-doped spinel Co ₃ O ₄ present
La	10	No
Ce	11	No
Pr	3	No
Nd	2	No
Sm	0	No
Eu	2	No
Gd	1	No

Supplementary Table 7. XPS results for Co contents in the obtained RE-Co₃O₄.

Atomic Co ²⁺ /Co ³⁺	
Co ₃ O ₄	1.62
La-Co ₃ O ₄	1.68
Ce-Co ₃ O ₄	1.71
Pr-Co ₃ O ₄	1.76
Nd-Co ₃ O ₄	1.79
Gd-Co ₃ O ₄	1.82
Sm-Co ₃ O ₄	1.84
Eu-Co ₃ O ₄	1.85

Supplementary Table 8. Comparison of the NO₃RR performance and other recently reported high-performance catalysts.

	Catalysts	Potential (V vs. RHE)	Electrolyte	Stability (cycles)	Electrode type	FE (%)	Yield rate (mg NH ₃ h ⁻¹ cm ⁻²)	Ref.
1	Co ₃ O ₄ /Cu NC	-1	1 M KOH 1 M NO ₃ ⁻	20	Carbon paper	97.7	114.0	1
2	La-Co ₃ O ₄ /C C	-0.6	0.1 M NaOH 0.1 M NO ₃ ⁻	12	Carbon paper	96.4	9.14	2
3	Pd-Co ₃ O ₄ /T M	-0.3	0.1 M NaOH 0.1 M NO ₃ ⁻	8	Ti mesh	98.7	12.68	3
4	Fe-Co ₃ O ₄ /T M	-0.7	0.1 M PBS 50 mM NO ₃ ⁻	5	Ti mesh	95.5	0.624	4
5	vCo-Co ₃ O ₄ / CC	-0.4	0.1 M NaOH 0.1 M NO ₃ ⁻	12	Carbon cloth	97.2	8.81	5
6	C/Co ₃ O ₄	-0.6	0.5 M K ₂ SO ₄ 50 mM NO ₂ ⁻	10	Carbon paper	100	4.18	6
7	Co ₂ B@Co ₃ O ₄ /TM	-0.7	0.1 M PBS 0.1 M NO ₃ ⁻	8	Ti mesh	97	8.57	7
8	Co ₃ O ₄ @Ni O	-0.7	0.5 M Na ₂ SO ₄ 3.2 mM NO ₃ ⁻	6	Carbon paper	63	6.46	8
9	(Cu _{0.6} Co _{0.4}) Co ₂ O ₄	-0.45	1.0 M KOH 0.1 M NO ₃ ⁻	/	Carbon paper	96.5	18.53	9
10	Ag-Co ₃ O ₄ /C C	-0.65	0.5 M K ₂ SO ₄ 200 ppm NO ₃ ⁻	9	Carbon cloth	96.1	3.86	10
11	Co ₃ O ₄ -Mnx	-1.2	0.5 M K ₂ SO ₄ 0.1 M NO ₃ ⁻	10	Cobalt foam	99.5	35	11
12	Co ₃ O ₄ /SiC	-0.79	1.0 M Na ₂ SO ₄ 0.1 M NO ₃ ⁻	10	Carbon paper	96.4	18.4	12
13	Co ₃ O ₄ -L	-0.9	1 M KOH 0.1 M NO ₃ ⁻	160 h	Carbon cloth	82.5	22.38	13
14	Co ₃ O ₄ -14	-0.7	0.1 M K ₂ SO ₄ 0.1 M NO ₃ ⁻	10	Carbon paper	94.1	25.28	14
15	Cu ₂ O+Co ₃ O 4	-0.3	0.1 M NaOH 0.1 M NO ₃ ⁻	10 h	Carbon paper	85.4	25.52	15

Supplementary Table 9. RE and Co contents of synthesized catalysts determined by ICP-OES (weight ratio, wt.%).

Catalysts	Elemental content (RE/wt.%)	Co/wt.%
La-Co ₃ O ₄	1.25	63.5
Ce-Co ₃ O ₄	1.27	62.7
Pr-Co ₃ O ₄	1.23	65.7
Nd-Co ₃ O ₄	1.24	62.4
Gd-Co ₃ O ₄	1.28	63.2
Sm-Co ₃ O ₄	1.14	63.5
Eu-Co ₃ O ₄	1.22	62.1

Supplementary References

1. Zhou, X.Y.; Gao, C.X.; Guan, M.H.; et al. Fe³⁺/Fe²⁺-Tuned iron oxides for enhanced interfacial nitrate activation toward efficient electrochemical ammonia production. *ChemSusChem*. **18**, e202501375, (2025).
2. He, X.; Xie, T.; Dong, K.; et al. Enhancing nitrate electroreduction for ammonia production over electron-deficient Co₃O₄ with La doping regulation. *Science China Materials*. **68**, 2756-2763, (2025).
3. Wu, Z.-Y.; Karamad, M.; Yong, X.; et al. Electrochemical ammonia synthesis via nitrate reduction on Fe single atom catalyst. *Nature Communications*. **12**, 2870, (2021).
4. Wei, P.; Liang, J.; Liu, Q.; et al. Iron-doped cobalt oxide nanoarray for efficient electrocatalytic nitrate-to-ammonia conversion. *Journal of Colloid and Interface Science*. **615**, 636-642, (2022).
5. Huang, P.; Fan, T.; Ma, X.; et al. 3D Flower-like zinc cobaltite for electrocatalytic reduction of nitrate to ammonia under ambient conditions. *ChemSusChem*. **15**, e202102049, (2022).
6. Li, P.; Jin, Z.; Fang, Z.; et al. A single-site iron catalyst with preoccupied active centers that achieves selective ammonia electrosynthesis from nitrate. *Energy & Environmental Science*. **14**, 3522-3531, (2021).
7. Hu, J.; Zhang, L. Hollow FeCo₂O₄@CuCo₂O₄ Z-Scheme heterostructure self-tandem photoelectro-reduction of nitrate to ammonia on Cu-O-Fe atomic bridge and energy supply in a Zn-nitrate cell system. *Applied Catalysis B: Environment and Energy*. **365**, 124899, (2025).
8. Li, J.; Zhao, D.; Zhang, L.; et al. A FeCo₂O₄ nanowire array enabled electrochemical nitrate conversion to ammonia. *Chemical Communications*. **58**, 4480-4483, (2022).
9. Zhao, Z.; Zhang, X.; Liu, X.; et al. Preparation of a branched hierarchical Co₃O₄@Fe₂O₃ core-shell structure with high-density catalytic sites for efficient catalytic conversion of nitrate in wastewater to ammonia. *Chemical Engineering Journal*. **499**, 156495, (2024).
10. Li, T.; Tang, C.; Guo, H.; et al. Understanding first electron transfer kinetic process of electrochemical nitrate reduction to ammonia on Fe₂O₃ nanorods array. *Chemical Engineering Journal*. **485**, 149560, (2024).
11. Li, T.; Tang, C.; Guo, H.; et al. In situ growth of Fe₂O₃ nanorod arrays on carbon cloth with rapid charge transfer for efficient nitrate electroreduction to ammonia. *ACS Applied Materials & Interfaces*. **14**, 49765-49773, (2022).
12. Zhang, S.; Li, M.; Li, J.; et al. High-ammonia selective metal-organic framework-derived Co-doped Fe/Fe₂O₃ catalysts for electrochemical nitrate reduction. *Proc Natl Acad Sci U S A*. **119**, (2022).
13. Sun, Y.; Shi, Y.; Gao, Y.; et al. Electroreduction of nitrate into ammonia on Co₃O₄: Mechanistic insights into Co²⁺-promoted NO₃RR performance. *Chemical Engineering Journal*. **512**, 162506, (2025).
14. Ying, Y.; Fan, K.; Luo, X.; et al. Unravelling the origin of bifunctional OER/ORR activity for single-atom catalysts supported on C₂N by DFT and machine learning. *Journal of Materials Chemistry A*. **9**, 16860-16867, (2021).
15. Du, Y.; Lu, H.; Wu, J.; et al. Selenium-deficient FeSe₂/Fe₃O₄ electrocatalyst for nitrate reduction to ammonia. *Angewandte Chemie International Edition*. **64**, e202420903, (2025).