Expanding catalyst discovery across material classes by machine learning

Junseok Moon[*1,2] **Seungwoo Yoo**[*1,2] **Taeghwan Hyeon**[1,2]

[*]*Equal contribution* [1]*Center for Nanoparticle Research, Institute for Basic Science (IBS), Seoul 08826, Republic of Korea* [2]*School of Chemical and Biological Engineering, and Institute of Chemical Processes, Seoul National University, Seoul 08826, Republic of Korea*

Correspondence to: Taeghwan Hyeon thyeon@snu.ac.kr

1. Introduction

The development of efficient catalysts is crucial for both traditional chemical processes and sustainable technologies such as water electrolysis. For the oxygen evolution reaction (OER), a key bottleneck in hydrogen production, multimetallic oxides and single-atom catalysts (SACs) have been widely studied. Although machine learning (ML) has accelerated discovery within individual material groups, expanding this search across all material classes requires integrating knowledge from diverse groups—an objective aligned with artificial general intelligence. However, combining datasets from different research articles remains challenging due to varying experimental conditions. This study addresses these limitations by introducing an ML system that unifies knowledge from multiple material groups to predict the catalytic performance of unseen materials. Our "crossbreeding" neural network (CBNN) simultaneously learns from experimental datasets on (1) SACs on carbon and (2) bulk perovskite oxides, enabling predictions for a new material class: SACs on perovskite oxides. The model not only predicted the overpotential (η) rankings of 14 SACs on oxides with unity precision but also identified a multimetallic SAC with an intrinsic overpotential of 349 mV at 10 mA cm^{-2}—surpassing all synthesized in both the original and new material groups. This demonstrates the potential of materials discovery through the integration of experimental data from diverse groups.

2. Results and Discussion

2.1 Extrapolated ML by integrating experimental datasets

Inorganic catalyst design typically emphasizes surface control and bulk structure optimization. To integrate experimental datasets across material groups, we selected two prior studies reporting over 20 OER catalysts each: 22 SACs on carbon [1] and 41 perovskite oxides [2], both evaluated in 0.1 M KOH. We merged these datasets to predict the activity of a new class—SACs on perovskite oxides (**Fig. 1a**). Our crossbreeding neural network (CBNN) incorporates two inputs (surface and bulk structures) and one output (OER overpotential) (**Fig. 1b**). SAC-on-carbon data were processed through the surface model and perovskite oxides through the bulk model; correspondingly, bulk layers deactivate for SACs on carbon, and surface layers deactivate for perovskite oxides, while both activate for SACs on perovskite oxides. Surface structures were represented as 2D images encoding elemental properties, and bulk oxides as graphs with elemental features assigned to nodes. The trained CBNN extrapolated catalytic activities for SACs on perovskite oxides (**Fig. 1c**), guiding the synthesis of 14 catalysts by depositing seven metals onto two representative perovskite oxides.

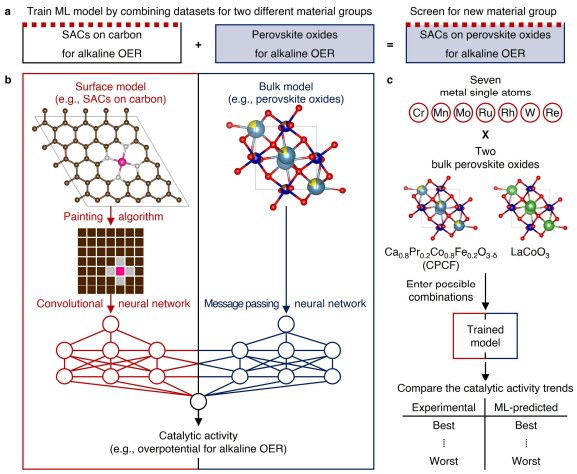

Fig. 1: Crossbreeding neural network. a, Schematic illustration of the catalyst discovery. b, Overview of the ML algorithm. c, Screening process of OER overpotentials.

2.2 Model training based on key co-descriptors selection

The CBNN surface (image) and bulk (graph) inputs were augmented with experimental and theoretical elemental properties. To identify the most relevant OER descriptors, we implemented an automated feature-selection workflow (**Fig. 2a**). Ten candidate features were first screened for statistical independence and correlation with overpotential, while a parallel NLP analysis of 3,634 research articles identified activity-related terms. Five key co-descriptors were consistently correlated with catalytic ac-

tivity across both material groups and were retained, whereas the remaining features were discarded. As shown in **Figs. 2b–c**, models trained with the selected co-descriptors outperformed those using discarded features, and joint training across both datasets further improved generalization, confirming effective cross-material learning.

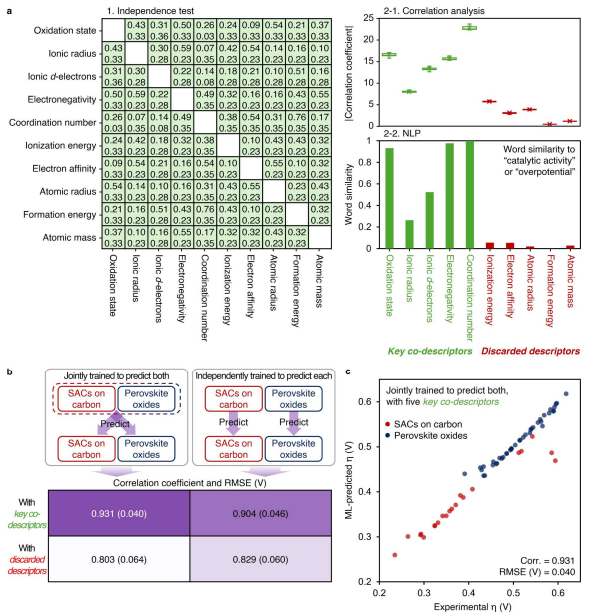

Fig. 2. Key co-descriptors selection and complementary effect on model training. a, Computation-driven key co-descriptors selection. c, Plot of experimental overpotentials versus ML-predicted values.

2.3 Predicting catalytic activity of new material group

We synthesized 14 monometallic SACs on perovskite oxides by depositing seven metal dopants onto CPCF and $LaCoO_3$ at 0.5 wt%. Structural characterization confirmed these catalysts were well-defined. To predict their catalytic performance, we input the structures into the CBNN trained with experimental data and key co-descriptors. Absolute overpotential predictions were impractical due to differences in measurement methods across datasets, thus we ranked catalytic activity based on ML-predicted overpotentials. Remarkably, ML-predicted rankings matched experimentally determined rankings with unity precision except two Re-containing cases (**Figs. 3a-b**). Building on the success of monometallic SACs, we explored multimetallic systems by combining seven dopants in different ratios, generating 8,008 SAC configurations on CPCF. The ML model identified the best catalyst: W, Mo, Ru, and Rh on CPCF (MM/CPCF). Experimental validation showed that MM/CPCF outperformed both monometallic SACs and all catalysts from the original two datasets (**Figs. 3c-e**).

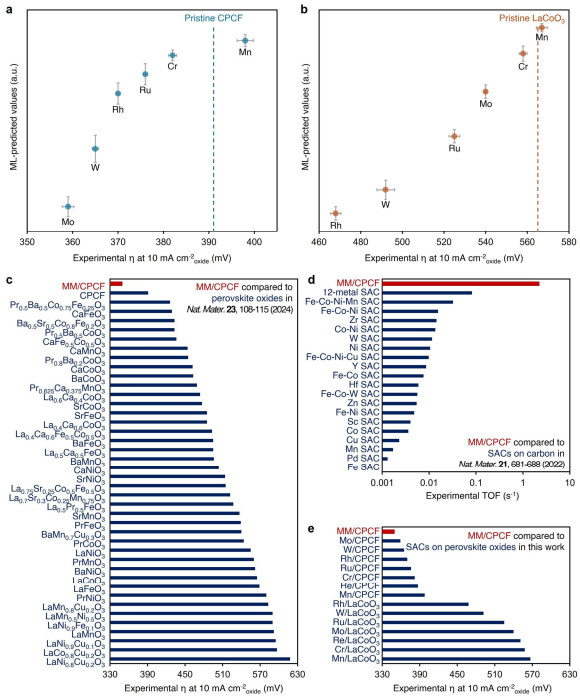

Fig. 3: Model prediction and experimental validation. a-b, Plot of experimental overpotential values versus ML-predicted values. c-e, Comparison of OER activity of MM/CPCF with oxides (c), SACs on carbon (d), and SACs on oxides (e).

3. Conclusion

This study shows that experimental datasets from different material groups can be unified through deep learning to predict performance in a new catalyst class, without additional experiments or theory. The framework combines automated descriptor selection with explainable ML and is readily extendable beyond catalysis to broader materials discovery.

Acknowledgments

The authors acknowledge the financial support by the Research Center Program of the IBS (IBS-R006-D1) in Korea.

References

[1] Han, L. et al. A single-atom library for guided monometallic and concentration-complex multimetallic designs. *Nature Materials* **21**, 681–688, 2022

[2] Moon, J. et al. Active learning guides discovery of a champion four-metal perovskite oxide for oxygen evolution electrocatalysis. *Nature Materials* **23**, 108–115, 2024