Science Advances
AAAS

Supplementary Note 1
(Details of Domain Knowledge Dataset)
for

Leveraging Data Mining, Active Learning, and Domain Adaptation in a Multi-Stage, Machine Learning-Driven Approach for the Efficient Discovery of Advanced Acidic Oxygen Evolution Electrocatalysts

Rui Ding *et al.*

*Corresponding author. Email: junhongchen@uchicago.edu, chenyuxin@uchicago.edu

Supplementary Note Discussion SND 1-1

Details of Dataset Construction and Preprocessing

Data Collection and Literature Search Strategy

The domain knowledge dataset for this study was meticulously constructed from peer-reviewed, publicly available research papers that focus on experimental explorations of metal-alloy/oxide type electrocatalysts for OER in acidic conditions. While acknowledging the existence of research works based purely on theoretical DFT or MD simulation, we have decided to exclude these from our domain knowledge dataset due to the disparity in metrics, such as properties calculated by DFT, between theoretical and experimental studies. This exclusion ensures consistency in our study, especially since, except for the final stage where domain adaptation is applied for training ML surrogate models for DFT simulation, our data resources for unsupervised/supervised data mining and the active learning workflow rely solely on experimental studies and feedback. The candidate publications are searched and obtained based on the following search query on Web of Science:

(TS=(oxygen evolution catalyst) OR TS=(oxygen evolution electrocatalyst) OR TS=(oxygen evolution electrocatalyst proton exchange) OR TS=(water oxidation catalyst) OR TS=(water oxidation electrocatalyst) OR TS=(OER catalyst) OR TS=(OER electrocatalyst) OR TS=(oxygen evolution catalyst pH) OR TS=(water oxidation catalyst pH) OR TS=(OER catalyst pH)) AND (TS=("acid" OR "acidic" OR "proton" OR "H2SO4" OR "HClO4" OR "proton exchange" or "polymer electrolyte"))

Criteria for High-Quality Literature Selection

Regarding the criteria for segregating a separate "high-quality dataset" from our original full dataset, the selection conditions are as follows: a journal impact factor >10 or an average daily citation rate ≥0.025 (according to the Web of Science dataset records as of January 20, 2023) or a publication date within 365 days as of January 20, 2023. As mentioned in the main text, this high-quality dataset reduces the number of records related to electrocatalytic activity to 1,358 and durability records to 345.

Data Preprocessing and Missing Value Imputation

To ensure data quality, we also excluded publications with incomplete records of key variable parameters. Nevertheless, the dataset still contains a small percentage of records missing information on precursor mixing (or hydrothermal process) duration (11.6%) or, in very few cases, calcination duration (0.8%). This is likely because these variables were not considered critical in most previous studies based on empirical observations. To address this, we adopted a common method of filling missing values using the median value from the dataset, a technique proven effective in handling minor missing values in our past research and in the ML field, as training ML algorithms cannot be performed with an input matrix containing missing values. We ensured that all critical parameters for performance and stability that are usually reported in the majority of publications, such as the types and amounts of precursors, processing temperatures, and catalyst loadings, are based on actual values recorded in the literature. **Tables SN 1-1** & **SN 1-2** (**Table S1** and **Table S2** in the major **Supplementary Materials** document respectively) display all the input features related to activity and durability in the domain knowledge dataset, along with their corresponding ranges of independent variables.

Standardization of Experimental Parameters

To manage variability in experimental conditions and testing standards, we standardized key experimental parameters across studies. Catalyst loadings, precursor compositions, annealing temperatures, and testing conditions were normalized to common units where applicable. For

instance, catalyst loading amounts were converted to a standard unit (e.g., mg/cm²), and testing conditions like electrolyte concentration and pH were standardized where possible. Moreover, when studies reported catalyst performance at different current densities, we interpolated or extrapolated data to a standard current density of 10 mA/cm² where feasible. Often, we would also extract manually from polarization curve figures the η_{10} values for secondary samples when the text is not providing, or the polarization curve is provided as Tafel plots. Conversion would also be conducted when current density by geometric area is not provided but mass. For the element composition and percentages, the percentage values are calculated manually from the synthesis paragraphs of publications by the amount and chemical types. And further, we would separate supporting materials in the precursors. These approaches ensured comparability across different datasets.

Integration of Chemical Domain Knowledge for Encoding ML Model Input

Finally, it is important to note that while the direct visualization of the domain knowledge dataset and the unsupervised data mining process do not necessitate the digitization of the chemical and elemental information of the metal elements, it is essential to transform these element types into fundamental intrinsic atomic properties data for supervised learning in ML model training as input encodings. To enrich the dataset with extended chemistry domain knowledge, we have incorporated properties such as relative atomic mass, atomic number, period, group, ionization potential, electronegativity, number of d electrons, and atomic radius. Consequently, for ML model training, this results in an increase in the actual dimensions of input features from 26 and 27 to 54 and 55, respectively, for the training of ML models aimed at predicting catalytic activity and stability. For a more direct visualization and comprehension of how the domain knowledge dataset is processed and utilized in different parts of the data mining works, we provide **Fig. SN 1-1** for a vivid illustration.

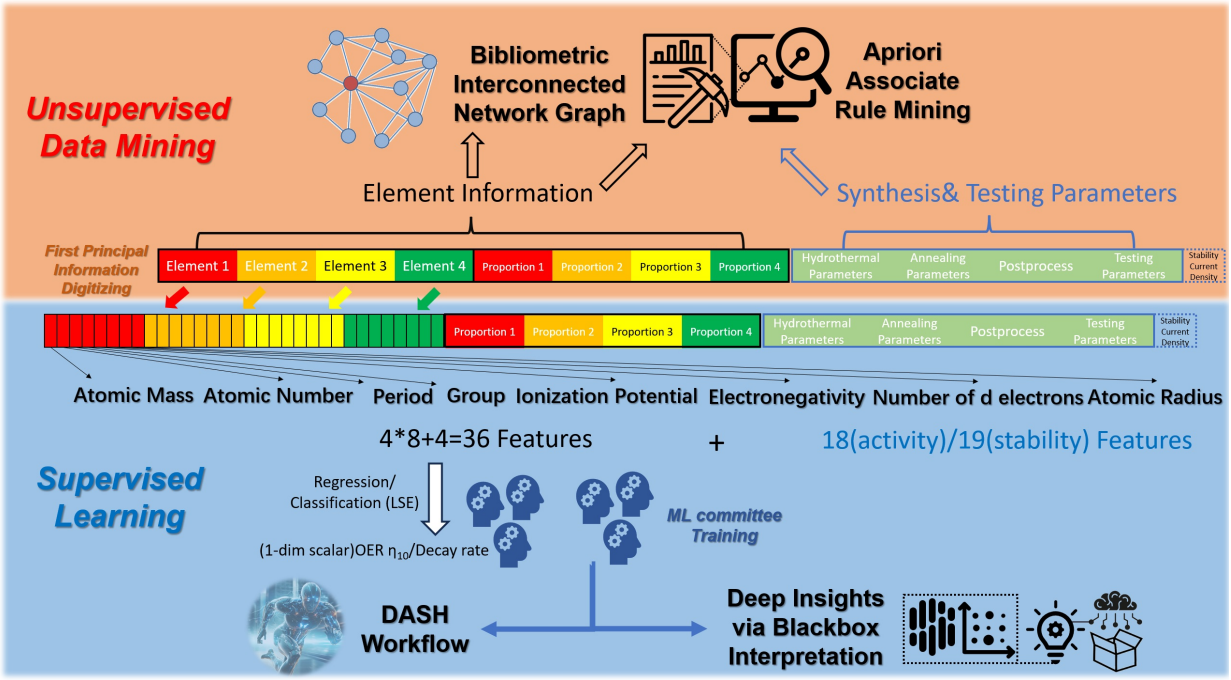

Fig. SN 1-1 Illustration of the digitization process for constructing the domain knowledge dataset for different data science work parts.

Table SN 1-1 The features in the domain knowledge dataset (the full dataset, 1,847 entries) for catalytic activity and corresponding variable range

Feature Name (Unit)	Variable Range
Metal_Dopant_1[a] （1~4 represent the proportion in precursor from high to low）	49 different metal elements[a,b]
Metal_Dopant_2	55 different metal elements or none
Metal_Dopant_3	39 different metal elements or none
Metal_Dopant_4	7 different metal elements or none
Metal_Dopant_1 Proportion in Precursor （at. %; refers to that in total 4 types of metal）	28.57~100
Metal_Dopant_2 Proportion in Precursor （at. %）	0~50
Metal_Dopant_3 Proportion in Precursor （at. %）	0~33.33
Metal_Dopant_4 Proportion in Precursor （at. %）	0~17.67
Hydrothermal Temperature (℃) (or precursor mixing)	-196~320
Hydrothermal Time (min) (or precursor mixing)	0 ~86,400
Hydrothermal Still/Stirring (0/1) (or precursor mixing)	0: still incubation; 1: stirring or sonication
Hydrothermal Strong Reductant in Liquid (0/1) (or precursor mixing)	0: False; 1: True
Hydrothermal Weak Reductant in Liquid (0/1) (or precursor mixing)	0: False; 1: True
Mixed in Solid or Liquid (0/1)	0: False (liquid mixing or hydrothermal); 1: True (ball milling)
Annealing Temperature (℃)	25~1,400
Annealing Time (min)	0~20,160
Annealing Still/Stirring (0/1)	0: False; 1: True
Annealing Atmosphere Inert (0/1)	0: False; 1: True
Annealing Atmosphere Reducing (0/1)	0: False; 1: True
Post-processing Acid Wash etc. (after annealing; 0/1)	0: False; 1: True
Catalyst Loading (mg cm^{-2})	0.000714~490
Support Material Loading (mg cm^{-2})	0~18
Support Material is not Carbon (support material, TiO_x etc.; 0/1)	0: False; 1: True
Electrode Type_Glassy Carbon/Carbon Paper or Ti Mesh (0/1)	0: Glassy Carbon;1: Carbon Paper or Ti Mesh
LSV Scanning Speed (mV s^{-1})	0.1~100
Electrolyte Proton Concentration (M)	0.01~6

Note:
a) For clarity and consistency in terminology, all metal elements in the precursor are referred to as “dopants” even though the 1st “dopant” is actually the primary metal. These are arranged in descending order of their proportions, from the 1st to the 4th.
b) When conducting non-ML method data mining, the dataset directly treats elements as frequent items. However, for ML modeling and fitting, the elemental information is digitized and represented by properties such as Relative Atomic Mass, Atomic Number, Period, Group, Ionization Energy, Electronegativity, Number of Outermost d Electrons, and Atomic Radius, as we have illustrated in detail in previous **Fig. SN 1-1**.
c) Additionally, in some studies, such as those using TiN or TiC as catalyst supports, the elements N and C do not exist in the final mixed oxide product in the same way they do in the precursor organic compounds. Therefore, despite the term "metal elements" being used in this paper, a few records in the dataset will include N and C. This is in consideration of their stable presence during the catalytic process and to account for their potential synergistic effects in catalysis.

Table SN 1-2 The features in the domain knowledge dataset (the full dataset, 453 entries) for stability and corresponding variable range

Feature Name (Unit)	Variable Range
Metal_Dopant_1 （1~4 represent the proportion in precursor from high to low）	41 different metal elements
Metal_Dopant_2	53 different metal elements or none
Metal_Dopant_3	32 different metal elements or none
Metal_Dopant_4	3 different metal elements or none
Metal_Dopant_1 Proportion in Precursor （at. %; refers to that in total 4 types of metal）	28.57~100
Metal_Dopant_2 Proportion in Precursor （at. %）	0~50
Metal_Dopant_3 Proportion in Precursor （at. %）	0~32.34
Metal_Dopant_4 Proportion in Precursor （at. %）	0~2.06
Hydrothermal Temperature (℃) (or precursor mixing)	-77~320
Hydrothermal Time (min) (or precursor mixing)	0 ~10,080
Hydrothermal Still/Stirring (0/1) (or precursor mixing)	0: still incubation; 1: stirring or sonication
Hydrothermal Strong Reductant in Liquid (0/1) (or precursor mixing)	0: False; 1: True
Hydrothermal Weak Reductant in Liquid (0/1) (or precursor mixing)	0: False; 1: True
Mixed in Solid or Liquid (0/1)	0: False (liquid mixing or hydrothermal); 1: True (ball milling)
Annealing Temperature (℃)	25~1,200
Annealing Time (min)	0~20,160
Annealing Still/Stirring (0/1)	0: False; 1: True
Annealing Atmosphere Inert (0/1)	0: False; 1: True
Annealing Atmosphere Reducing (0/1)	0: False; 1: True
Post-processing Acid Wash etc. (after annealing; 0/1)	0: False; 1: True
Catalyst Loading (mg cm^{-2})	0.000714~490
Support Material Loading (mg cm^{-2})	0~5.58
Support is not Carbon (support material, TiO_x etc.; 0/1)	0: False; 1: True
Electrode Type_Glassy Carbon/Carbon Paper or Ti Mesh (0/1)	0: Glassy Carbon;1: Carbon Paper or Ti Mesh
Electrolyte Proton Concentration (M)	0.01~6
Stability Constant Current Density (mA cm^{-2})	0.1~1,000
Stability Test Time (h)	0.28~8,000

Note:
After integrating quantum chemistry fundamental intrinsic atomic properties information into the dataset, we further examined the correlation matrices for the initial dataset using Kendall, Spearman, and Pearson methods. The corresponding results could be checked in the online GitHub repository (https://github.com/ruiding-uchicago/DASH) for readers with interests ("/Online Repository Figs./"): **Fig. OR1~OR12**. While there is a high degree of inter-correlation among features that inevitably express fundamental intrinsic atomic properties information of elements, such as relative atomic mass, atomic number, period, group, ionization potential, electronegativity, number of d electrons, and atomic radius, the correlations are significantly lower among other features. Specifically, this lower correlation is observed both within features related to material synthesis and testing, and between these features and the fundamental intrinsic atomic properties of elements (except for reasonable correlation between annealing time and temperature). This finding is particularly encouraging for several reasons. Firstly, the lower degree of correlation among these sets of features suggests a diverse and rich dataset, which is crucial for the robustness of ML models. It indicates that our dataset encompasses a wide range of independent variables, enhancing the potential for our models to capture and learn from a broad spectrum of material behaviors and properties. Secondly, the low correlation between material synthesis/testing features and elemental fundamental intrinsic atomic properties implies that our dataset is not dominated by any single type of information. This diversity ensures that our ML models are not biased towards fundamental intrinsic atomic properties features alone but are also informed by practical, experimental data. In summary, this characteristic of our dataset is advantageous for developing nuanced and comprehensive ML models. It allows for the exploration of complex interactions within materials, potentially leading to novel insights and breakthroughs in the field of electrocatalysis.
Corresponding unprocessed domain knowledge dataset .csv file (that is readable in Excel software), python script files and generated pkl dataset files (for supervised learning, training first iteration ML committee based on the previously mentioned domain knowledge dataset.) are stored and publicly available in the: "/ML Databases and Scripts/Domain Knowledge Database Preprocessing" directory of the DASH online repository.

Supplementary Discussion SND 1-2
Visualization of the Domain Knowledge Datasets

In **Fig. SN 1-2**, we use scatter plot to directly visualize the distribution of the data points in the full domain knowledge dataset to provide a comprehensive overall picture of the current research progress. In **Fig. SN 1-2a** for η_{10}, it can be observed with the following patterns:

1. The majority of the η_{10} reported are above 200 mV (93.0%).
2. The annealing temperatures are usually under 700 °C (81.2%)
3. The catalyst loading values have been in a wide range, but generally between 0.01 and 1 mg cm^{-2} (89.2%)

Fig. SN 1-2b shows the time averaged voltage decay rate in constant current density stability test, which is computed by dividing the voltage decay with the testing time. And a very compelling trend could be observed: the logarithmic decay rate exhibits a near-linear relationship with the logarithmic test time. Intriguingly, this suggests that with prolonged test durations, the average decay rate diminishes. This observation aligns with the consensus in existing literature[1-4], indicating potential underlying mechanisms such as catalyst stabilization phenomena, the formation of protective layers, and a self-optimization of active sites over extended operational periods. These findings not only corroborate previous research but also provide deeper insights into the durability of electrocatalysts. Moreover, they underscore the importance of long-term testing for a more accurate assessment of catalyst performance, reinforcing the value of extensive operational studies in understanding and improving electrocatalyst longevity.

In our further analysis, as depicted in **Fig. SN 1-3**, we delved into another crucial aspect of our dataset: the metal elements, focusing particularly on the two most commonly used elements, Ru and Ir. We meticulously quantified the proportions of these elements in the precursors across all data entries. Our contour map comparisons of these proportions with η_{10} and decay rate metrics revealed a trend consistent with existing consensus: a higher ratio of Ru tends to yield superior catalytic performance, whereas Ir appears to offer greater long-term service stability. This observation becomes particularly significant when considering the current market prices, where Iridium is approximately 11.8 times more expensive than Ruthenium (source: https://pmm.umicore.com/en/prices/iridium/; https://pmm.umicore.com/en/prices/ruthenium/; date:2023-11-21). Balancing the excellent catalytic activity of ruthenium oxides with their stability could, therefore, emerge as a more cost-effective approach, potentially serving as a pragmatic starting point for further exploration. Furthermore, it is noteworthy that current research and applications in PEM electrolyzer single cells predominantly utilize Iridium oxides[5], with ruthenium oxides and their derivatives being less common in practical scenarios. Consequently, in this study, along with the results obtained through data mining process, we have chosen to maintain Ru as an indispensable element among the four elements considered in our DASH workflow. This decision is aimed at exploring this relatively less trodden path, which not only presents greater challenges but also offers ample opportunities for discovery in the realm of cost-effective and efficient electrocatalysts.

In general, through a superficial glimpse in this supplementary discussion, we could get consistent first impression with what we have obtained from unsupervised and supervised data mining sections.

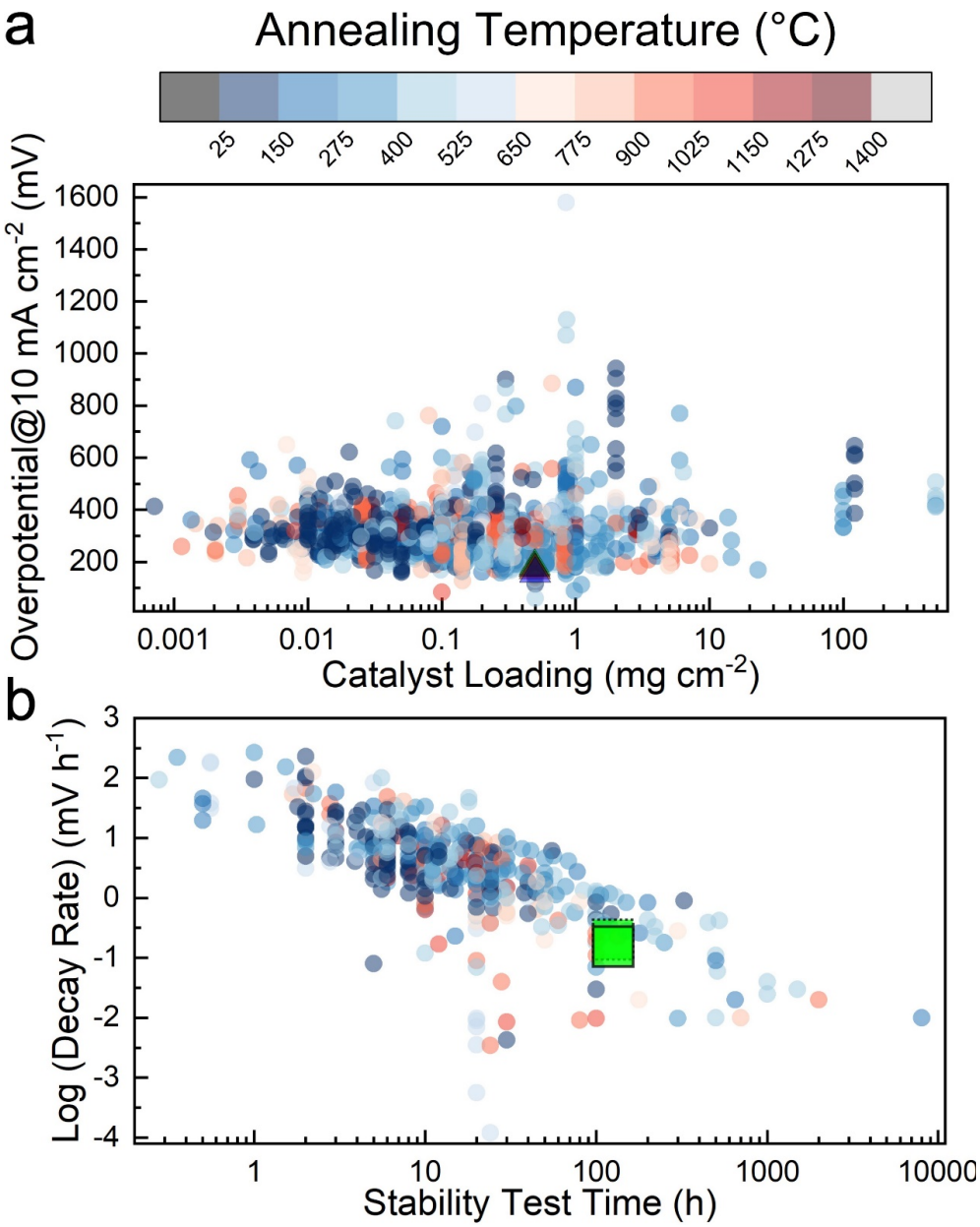

Fig. SN 1-2 Visualizations of the distribution of the data points in the full domain knowledge dataset, with a color bar representing annealing temperature. For samples that did not undergo the annealing step during data collection, their annealing temperature is defaulted to 25 °C, and the corresponding annealing time is recorded as 0. (a) illustrates the distribution of η_{10} as a function of catalyst loading, with the triangles representing the final best samples obtained in the 5th batch by the DASH workflow (**Fig. 4F**); (b) displays the distribution of the Decay Rate (logarithmically transformed) as a function of stability test time, with the dashed and solid edge squares representing the stability performance of the final sample C's stability metrics measured in the single cell electrolyzer test corresponding to 20 mA cm^{-2} and 10 mA cm^{-2}, respectively (**Fig. 4I**).

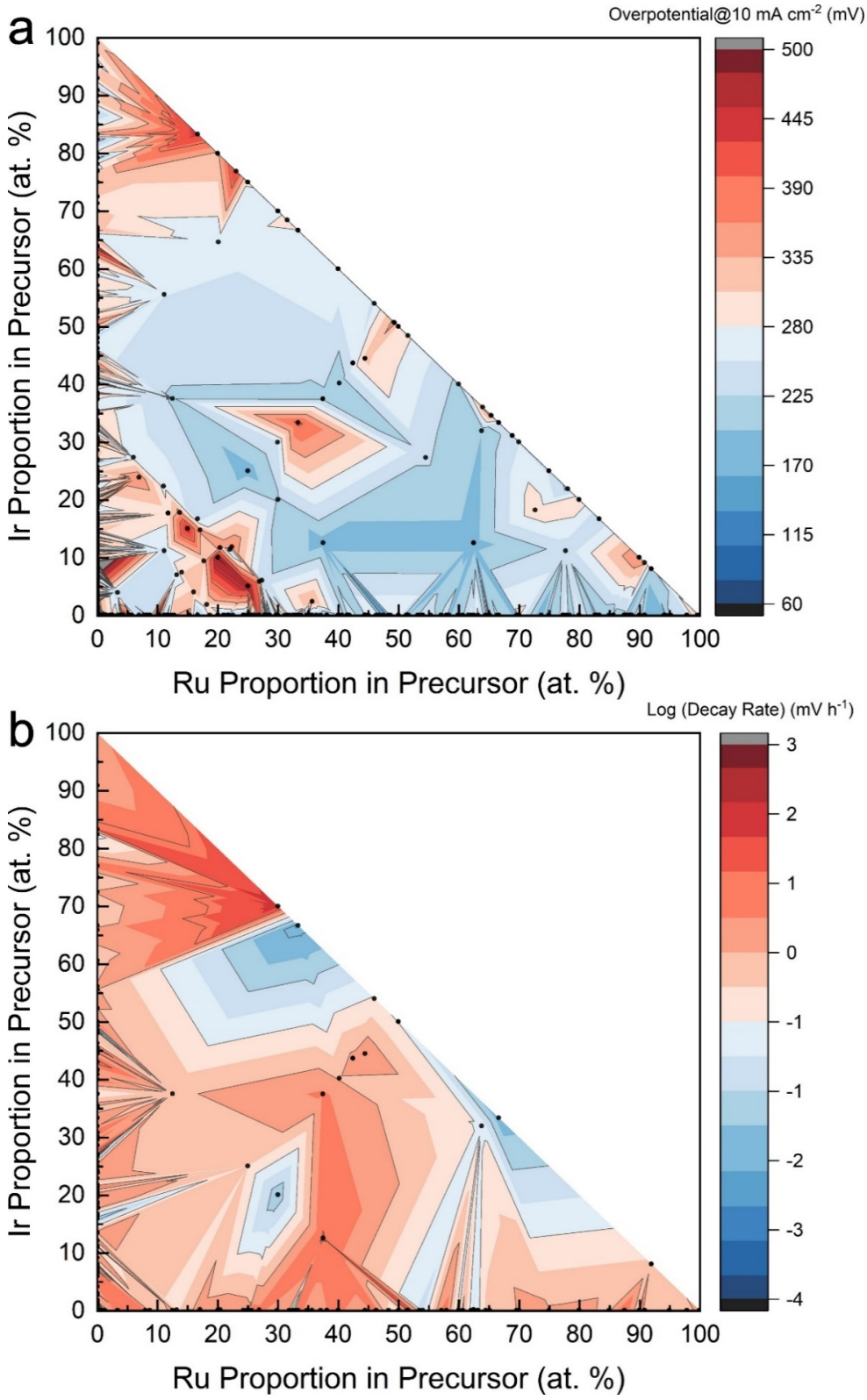

Fig. SN 1-3 Contour heat map of the distribution of OER activity and stability's changing pattern with Ir and Ru proportion in precursors. (a) η_{10}; (b) decay rate, in the full domain knowledge dataset.

Supplementary Note 1 References:

1 Lai, Q., Vediyappan, V., Aguey-Zinsou, K.-F. & Matsumoto, H. One-Step Synthesis of Carbon-Protected Co3O4 Nanoparticles toward Long-Term Water Oxidation in Acidic Media. *Advanced Energy and Sustainability Research* **2** (2021). https://doi.org:10.1002/aesr.202100086

2 He, H. *et al.* Dual Metal-Loaded Porous Carbon Materials Derived from Silk Fibroin as Bifunctional Electrocatalysts for Hydrogen Evolution Reaction and Oxygen Evolution Reaction. *ACS Appl Mater Interfaces* **13**, 30678-30692 (2021). https://doi.org:10.1021/acsami.1c07058

3 Li, G., Li, S., Ge, J., Liu, C. & Xing, W. Discontinuously covered IrO2–RuO2@Ru electrocatalysts for the oxygen evolution reaction: how high activity and long-term durability can be simultaneously realized in the synergistic and hybrid nano-structure. *Journal of Materials Chemistry A* **5**, 17221-17229 (2017). https://doi.org:10.1039/c7ta05126c

4 Bai, J. *et al.* Surface Engineering of Carbon-Supported Platinum as a Route to Electrocatalysts with Superior Durability and Activity for PEMFC Cathodes. *ACS Appl Mater Interfaces* **14**, 5287-5297 (2022). https://doi.org:10.1021/acsami.1c20823

5 Ding, R. *et al.* Guiding the Optimization of Membrane Electrode Assembly in a Proton Exchange Membrane Water Electrolyzer by Machine Learning Modeling and Black-Box Interpretation. *ACS Sustainable Chemistry & Engineering* **10**, 4561-4578 (2022). https://doi.org:10.1021/acssuschemeng.1c08522