Science Advances
AAAS

Supplementary Note 2
(Details of Unsupervised Data Mining)
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 2-1

Details of Bibliometric Interconnected Network Graph and Apriori Associate Rule Mining

Both methods employed in our research, the Bibliometric Interconnected Network Graph and Apriori Associate Rule Mining, fundamentally rely on the association of patterns or parameter judgment statements, conceptualized as "itemsets" in the associative rule mining process[1,2]. These itemsets are pivotal in identifying and categorizing elements, element combinations, and synthesis parameter values that contribute positively or negatively to the Oxygen Evolution Reaction (OER) activity and stability. The associative rule-based methods draw from a rich theoretical foundation, initially developed for market basket analysis in transactional data. This foundation is adapted in our research to the scientific investigation of catalytic materials, where each "transaction" or dataset entry can be viewed as a combination of elemental and processing conditions akin to a shopping list. Through this analogy, we systematically explore how different combinations of elements and synthesis conditions associate with desirable or undesirable catalytic properties. By discretizing input features and values (as the initial domain knowledge datasets, part of the features is continuous variables), we create a structured framework that allows for the systematic identification of frequent itemsets, which in turn reveals the underlying patterns and associations that define effective catalytic systems. This methodological approach not only elucidates the direct correlations but also enables the discovery of complex multi-dimensional relationships that govern the behavior of catalytic materials under various operational conditions. And in the results, two concepts are most important:

Lift Value: In Apriori data mining, the 'Lift' value indicates the strength of an association between itemsets. A high 'Lift' means the presence of one item significantly increases the likelihood of the other, suggesting a strong association that could indicate effective performance in the context analyzed.

Support Value: The 'Support' value in Apriori data mining represents the prevalence of an itemset within the entire dataset. A high 'Support' value denotes that the itemset frequently occurs together, thus underlining its commonality and potential significance in the analyzed context.

Since elements might be most decisive and are naturally discretized, we are able to present the lift values and supporting values in a very vivid way by a Bibliometric Interconnected Network Graph approach. In this simplified stage, we focus only on element types and their combinations, visualization plays a crucial role in illustrating the complex relationships within the data. The network graph serves as a vivid representation, using colors and geometrical dimensions to convey significant data properties. Here, the color intensity of nodes actually represents the lift value, indicating the strength of association or relevance of particular elements to high-quality OER performance. The size of the nodes and the width of the lines connecting them denote the support values, reflecting the frequency of occurrence of individual elements or element pairs across the dataset. This graphical representation aids in intuitively understanding the distribution and significance of various chemical elements in the context of OER catalysis.

Apriori Associate Rule Mining which comes next, as the standard method extends the associative analysis by encompassing a broader range of features within the dataset, facilitating a more detailed exploration of patterns that govern material properties and process parameters. In our study, we applied the Apriori algorithm to extract and analyze associative rules, again using lift and support as key metrics. As we mentioned, the lift value serves as an indicator of the strength of association between itemsets, where a higher lift value signifies a stronger likelihood of the itemset contributing to optimal OER performance (activity or stability). Conversely, the support value measures the prevalence of an itemset within the dataset, helping identify the most common and potentially significant associations. Our analysis extends to the synthesis and testing parameters, integrating physical process conditions such as annealing temperature, hydrothermal conditions, and precursor mixing techniques. By correlating these parameters with OER activity and stability, we uncover valuable insights into the synthesis process's influence on the electrocatalytic properties of materials.

In summary, the integration of Bibliometric Interconnected Network Graph and Apriori Associate Rule Mining into our research methodology provides a comprehensive analysis of the data from statistical perspective over the whole domain. This analytical synergy facilitates a deeper understanding of the

material properties and synthesis parameters, ultimately guiding the development of more efficient and stable OER catalysts. Through these unsupervised data mining techniques, we leverage the vast data landscape to extract meaningful insights. Compared with traditional subjective approach in determining initial exploration space, our approach would be considered data analytic-based and hence more rational. The results under different thresholds exhibit a consistent trend, with the exception of the most suitable visualization results at the moderate threshold, which can be observed in **Fig. S3. Fig. S4, Fig. 2 A-B** and **Fig. S5 A-B**. In this **Supplementary Note 2**, we provide the network results for other thresholds to the readers, illustrated in **Figs. SN 2-1** to **Fig. SN 2-4**.

The scripts for all of the unsupervised data mining analysis are publicly available in our GitHub repository in the directory "/Unsupervised Data Mining".

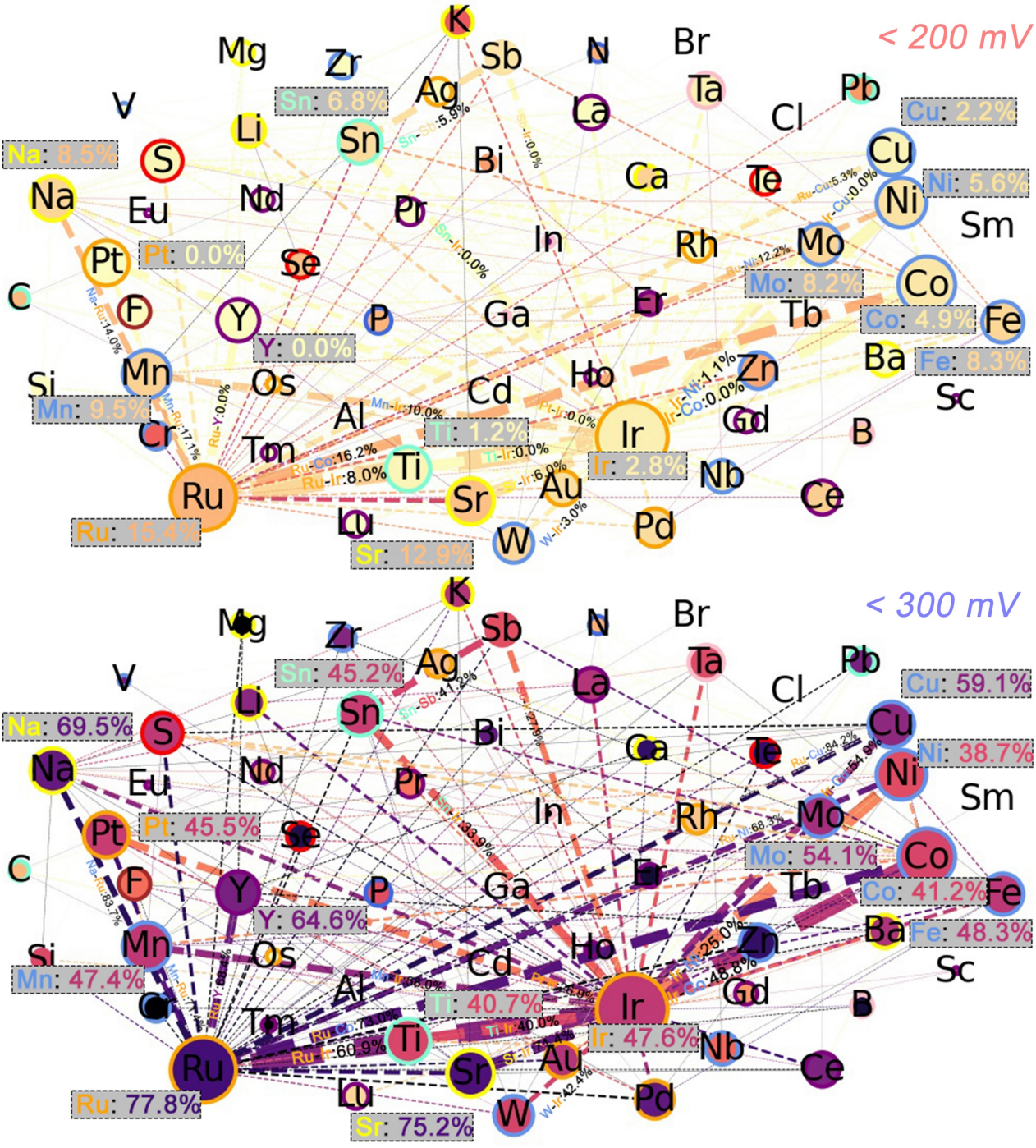

Fig. SN 2-1 Upper: network graph based on the full domain knowledge dataset of chemical elements with a qualified overpotential boundary set at 200 mV; Bottom: network graph based on the full domain knowledge dataset of chemical elements with a qualified overpotential boundary set at 300 mV.

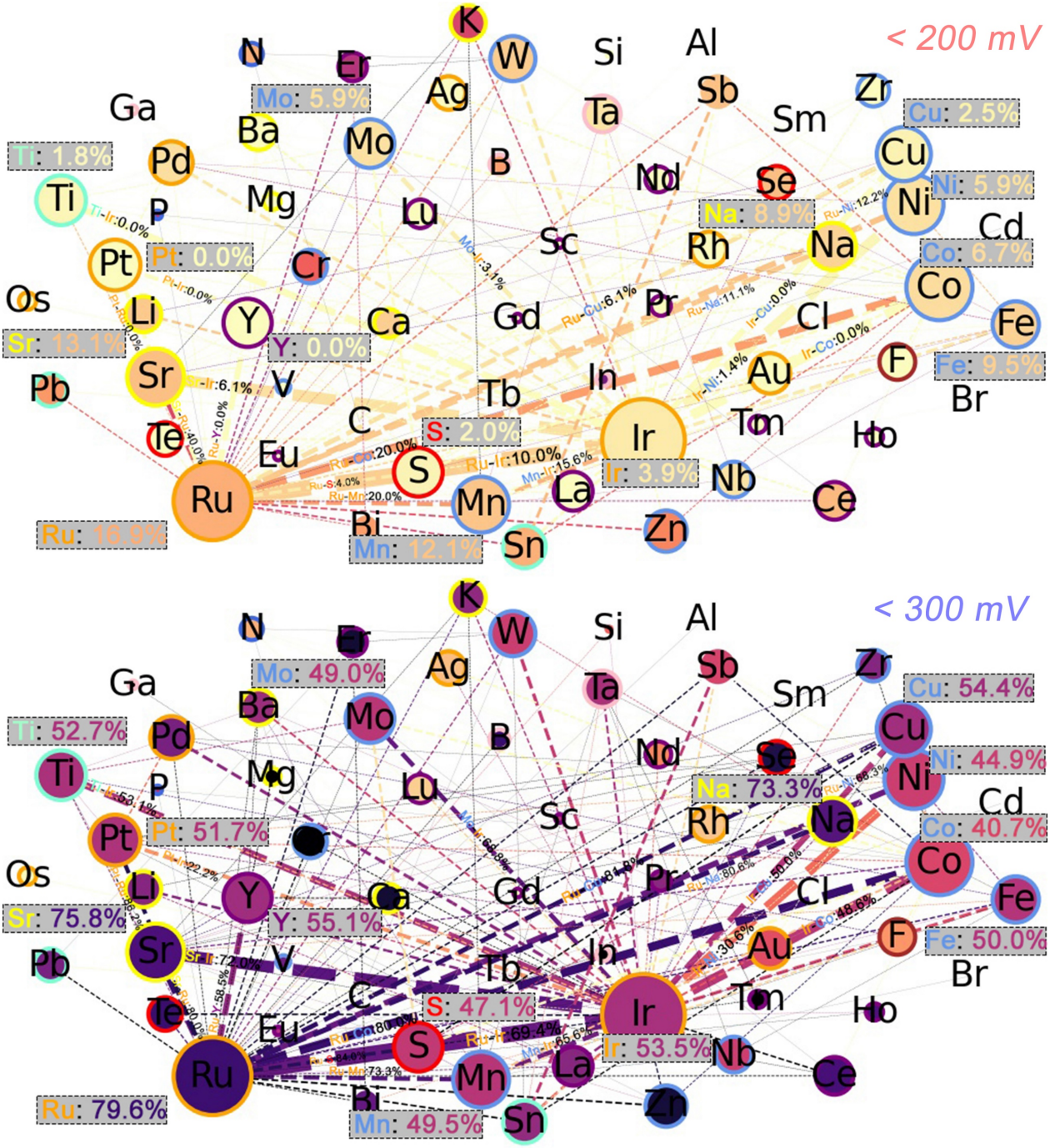

Fig. SN 2-2 Upper: network graph based on the high-quality domain knowledge dataset of chemical elements with a qualified overpotential boundary set at 200 mV; Bottom: network graph based on high-quality domain knowledge dataset of chemical elements with a qualified overpotential boundary set at 300 mV.

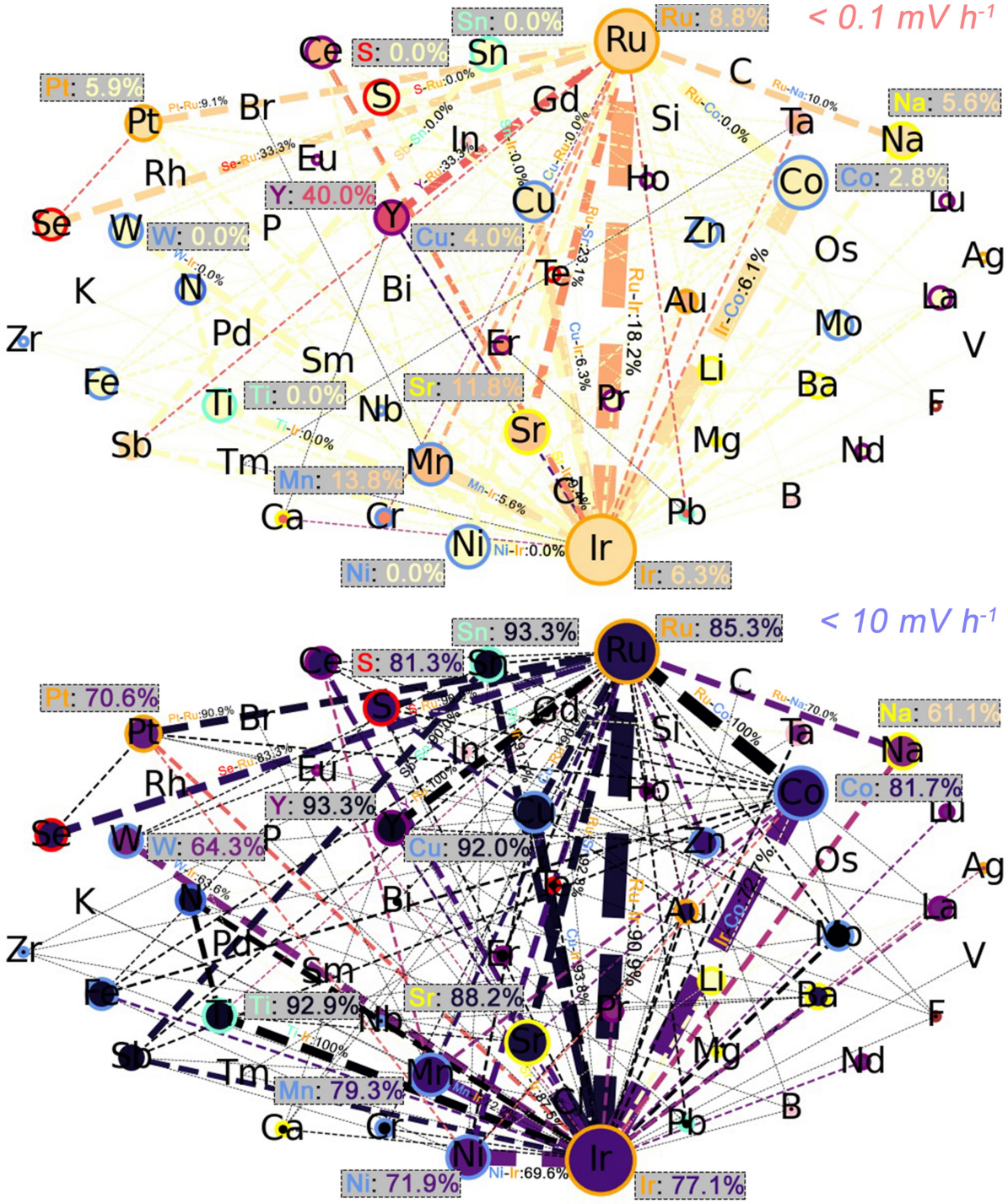

Fig. SN 2-3 Upper: network graph based on the full domain knowledge dataset of chemical elements with a qualified voltage decay rate boundary set at 0.1 mV h^{-1}; Bottom: network graph based on the full domain knowledge dataset of chemical elements with a qualified voltage decay rate boundary set at 10 mV h^{-1}.

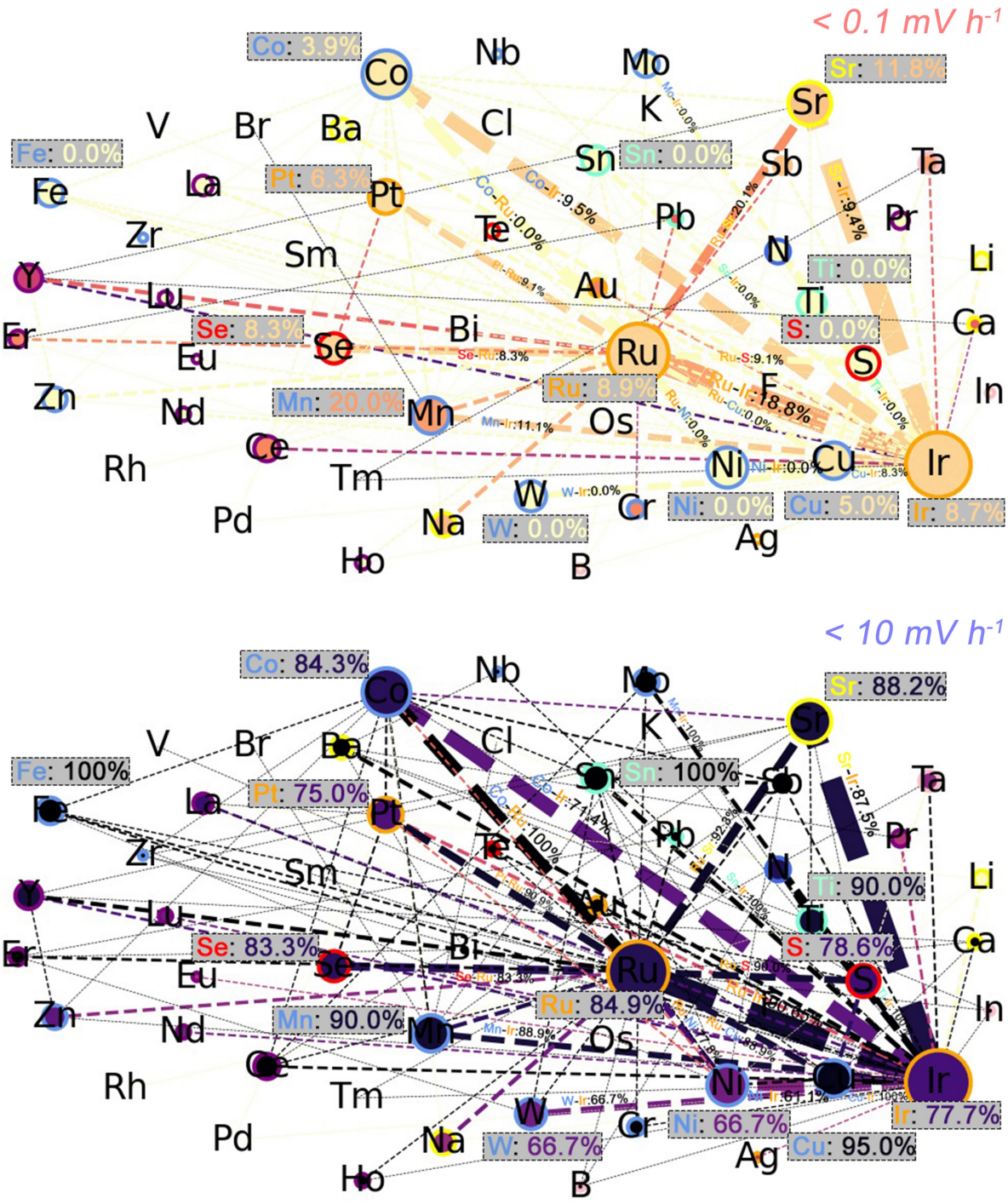

Fig. SN 2-4 Upper: network graph based on the high-quality domain knowledge dataset of chemical elements with a qualified voltage decay rate boundary set at 0.1 mV h^{-1}; Bottom: network graph based on the high-quality domain knowledge dataset of chemical elements with a qualified voltage decay rate boundary set at 10 mV h^{-1}.

Supplementary Note Discussion SND 2-2

Unsupervised data mining for based on domain knowledge dataset.

We have presented the main conclusions, drawn from unsupervised data mining process that directly help us with narrowing down and determining the initial exploration space for later experimental attempts. However, in this supplementary discussion, we also highlight the details of how we obtain them, along with other interesting discoveries from the data mining.

Activity-Fig. 2, **Fig. S3**, **Fig. SN 2-1**, **Fig. SN 2-2**

We start with discussing the element network graphs (**Fig. S3**, **Fig. SN 2-1**, **Fig. SN 2-2**) summarized in **Fig. 2A-B**. Firstly, focusing on individual elements, Ir is most frequently reported, followed by Ru, Co, and Ni in both full initial and high-quality datasets. However, in terms of qualification rate, Ru stands out as expected. Intriguingly, in the high-quality dataset, Na slightly surpasses Ru with a qualification rate of 55.6%. This highlights that the first and second group elements like Na and Sr, with their impressive qualification percentages, are only slightly behind Ru in potential. Secondly, when examining the coexistence of two different elements, the frequency trends align with the individual appearance rates of each element. The coexistence of Ir and Ru is the most common, likely reflecting past efforts in the field to balance activity and stability. However, a closer look at the qualification percentages reveals distinct trends between Ir and Ru oxides when doped with other elements. For instance, in the full dataset, if we use the single element qualification percentage as baselines (51.1% and 12.8% for Ir and Ru, respectively), the presence of another transition metal alongside Ru more frequently results in higher qualification percentages compared to Ir (6 out of 13 vs. 4 out of 13). This phenomenon becomes more obvious in the high-quality dataset (7 out of 13 vs. 4 out of 13). This propensity of Ru to effectively interact with other metal dopants to enhance the OER activity, underscores its potential for synergistic effects, making it a valuable avenue for exploration. This synergy, coupled with Ru's cost-effectiveness and underrepresentation in current PEM electrolyzer applications (**Supplementary Discussion SND 1-2** in **Supplementary Note 1**), solidifies our decision to select Ru as the indispensable element in our next module in active learning of the DASH workflow. Furthermore, in the last DFT-domain adaptation section of our study, we also chose RuO_2 (110) as the doping host matrix, due to its above-mentioned potential in offering a rich landscape for discovery and optimization in electrocatalyst design.

As for Apriori associate rule mining in **Fig. 2C-D**, our findings highlight Ru as a standout element as expected, both as a primary and as a secondary metal. Additionally, the preference of second group alkaline earth metals like Sr and Zn are found. Also, Mn as a secondary element, is evidently preferred in the high-quality dataset. The consistency observed in catalyst loading, annealing conditions, and precursor mixing parameters between the full and high-quality datasets offers a robust, data-driven foundation for our subsequent experimental research to define the candidate space for parameters.

Stability-Fig. S5, **Fig. S4**, **Fig. SN 2-3**, **Fig. SN 2-4**

And as mentioned, beyond OER catalytic activity, we also conducted a similar analysis of stability data (**Fig. S4**, **Fig. SN 2-3**, **Fig. SN 2-4**), an often overlooked but crucial aspect in practical applications. In **Fig. S5** in major **Supplementary Materials** document, paralleling the methodology of **Fig. 2** in the main text, we similarly present a nuanced unsupervised data mining analysis centered on the stability of OER electrocatalysts. Our initial observations regarding the frequency of earth-abundant elements such as Co, Ni, Mn, Cu, and first/second group metals like Sr and Na, indicate a subtle shift in their order of appearance. Upon closer examination of individual elements, it becomes apparent that Ir, while in consensus to be regarded as more durable in acidic OER than Ru, shows a slightly lower qualification rate compared to Ru in both the full and high-quality domain knowledge datasets (18.7% vs. 27.1% in the full dataset; 23.8% vs. 26.0% in the high-quality dataset). Nevertheless, the narrative slightly changes when we consider the interplay with other metal elements. Here, Ir matches Ru in its ability to exceed the single-element baseline in our datasets (7 out of 13 cases in both datasets). This phenomenon is particularly noteworthy when Ir is doped with other elements, consistently showing an enhanced qualification rate. This trend can be attributed to the intrinsic stability of Ir as an OER catalyst, which, while well-recognized, may be somewhat limited in isolation due to factors like restricted catalytic sites or suboptimal electronic structures.

But all in all, we could intriguingly via this domain knowledge perspective, conclude that Ru is actually comparable to Ir, if using decay rate and from the whole domain publication perspective. Another point that should be noted is that the introduction of other elements seems to indeed create synergistic effects, potentially enhancing the electronic structure or increasing the number of active sites. This synergy is especially pronounced for Ir, whose properties appear to be more effectively complemented or augmented by other elements as validated by previous research works[3,4], leading to a superior performance when used in combination. These findings underscore the multifaceted nature of catalyst stability and highlight the potential of element combinations in optimizing OER catalyst performance. Finally, another point that we could obtain from a broad literature review perspective, actually Ru-based candidates could show evenly matched performance in terms of stability, which further solidify our confidence in choosing Ru as the base element in the initial exploration space, by constraining following active learning modules to include it in the four element candidates.

In our Apriori associate rule mining analysis for stability, as depicted in **Fig. S5C-D**, a striking consistency emerged between the itemsets derived from both the full and high-quality domain knowledge datasets. A key observation was the association of prolonged annealing times with a reduced average decay rate in performance. This trend likely stems from the role of extended processing in fostering more ordered and defect-free crystal structures in OER electrocatalysts. Longer annealing times allow for the gradual and thorough formation of these structures, which are inherently more stable and resilient to long-term operational stresses. This enhanced stability is crucial for maintaining consistent catalytic activity over extended periods. Additionally, our analysis highlighted a preference for higher annealing temperatures, suggesting at least 500 °C as optimal. This preference aligns with the thermodynamic principle that higher temperatures can drive reactions towards more thermodynamically stable phases. Elevated temperatures during annealing can promote the formation of more crystalline and phase-pure materials, which are typically more stable and exhibit enhanced catalytic properties. Such materials are better equipped to withstand the harsh conditions of OER processes, where stability is paramount. Furthermore, the analysis revealed an interesting trend regarding the choice of substrate for stability testing. Stability tests conducted on carbon paper or titanium mesh were found to yield more reliable results compared to those on glassy carbon electrodes. This could be attributed to the mechanical robustness and enhanced electron transfer capabilities of carbon paper and titanium mesh. These substrates offer a more stable and uniform platform for electrochemical reactions, which is particularly important in OER applications. The aggressive nature of oxygen evolution in OER can be detrimental to less robust electrode materials like glassy carbon, leading to skewed or unreliable stability assessments. Therefore, the use of carbon paper or titanium mesh as substrates not only provides a more accurate reflection of the catalyst's stability but also contributes to the overall reliability of the testing process.

In our expanded analysis, where we increased the itemset length to three, as shown in **Figs. OR13-OR16** for readers with interests to reach out in online repository, we observed a general consistency with the results obtained from itemsets of length two, as previously discussed in **Figs. 2C-D** and **Fig. S5C-D**. This consistency was particularly evident in terms of preferred elemental choices and optimal conditions for annealing and hydrothermal processing. However, the analysis also revealed new insights. For catalytic activity, cobalt (Co) emerged as a promising dopant, a conclusion supported by both the full and high-quality domain knowledge datasets. In terms of stability, extending the itemset length to three brought additional focus to catalyst loading, with a recommendation favoring a loading of over 1 mg cm^{-2}. This recommendation aligns well with chemical intuition, as higher catalyst loadings can provide a greater number of active sites for the electrochemical reactions, potentially leading to improved stability. This is particularly relevant in OER processes, where the efficiency and longevity of the catalyst are closely tied to the availability of active sites.

Supplementary Note 2 References:

1 Can, E. & Yildirim, R. Data mining in photocatalytic water splitting over perovskites literature for higher hydrogen production. *Applied Catalysis B: Environmental* **242**, 267-283 (2019). https://doi.org:10.1016/j.apcatb.2018.09.104

2 Tapan, N. A., Günay, M. E. & Yildirim, R. Constructing global models from past publications to improve design and operating conditions for direct alcohol fuel cells. *Chemical Engineering Research and Design* **105**, 162-170 (2016). https://doi.org:10.1016/j.cherd.2015.11.018

3 Escalera-López, D., Jensen, K. D., Rees, N. V. & Escudero-Escribano, M. Electrochemically Decorated Iridium Electrodes with WS3−x Toward Improved Oxygen Evolution Electrocatalyst Stability in Acidic Electrolytes. *Advanced Sustainable Systems* **5** (2021). https://doi.org:10.1002/adsu.202000284

4 Spori, C. *et al.* Molecular Analysis of the Unusual Stability of an IrNbO(x) Catalyst for the Electrochemical Water Oxidation to Molecular Oxygen (OER). *ACS Appl Mater Interfaces* **13**, 3748-3761 (2021). https://doi.org:10.1021/acsami.0c12609