

A multimodal robotic platform for multi-element electrocatalyst discovery

<https://doi.org/10.1038/s41586-025-09640-5>

Received: 3 January 2024

Accepted: 16 September 2025

Published online: 23 September 2025



Zhen Zhang^{1,9}, Zhichu Ren^{1,9}, Chia-Wei Hsu^{1,9}, Weibin Chen^{2,9}, Zhang-Wei Hong³, Chi-Feng Lee⁴, Aubrey Penn⁵, Hongbin Xu¹, Daniel J. Zheng¹, Shuhan Miao⁶, Yimeng Huang², Yifan Gao², Weiyin Chen², Hugh Smith¹, Yaoshen Niu¹, Yunsheng Tian³, Ying-Rui Lu⁷, Yu-Cheng Shao⁷, Sipei Li², Hsiao-Tsu Wang⁴, Iwnetim I. Abate¹, Pulkit Agrawal¹, Yang Shao-Horn^{1,8} & Ju Li^{1,2}✉

One of the goals of ‘AI for Science’ is to discover customized materials through real-world experiments. Pioneering advances have been made in computational predictions and the automation of materials synthesis^{1–7}. Yet most materials experimentation remains constrained to using unimodal active learning approaches, relying on a single data stream. The potential of artificial intelligence to interpret experimental complexity remains largely untapped^{8,9}. Here we present Copilot for Real-world Experimental Scientists (CRESt), a platform that integrates large multimodal models (incorporating chemical compositions, text embeddings and microstructural images) with knowledge-assisted Bayesian optimization and robotic automation. CRESt uses knowledge-embedding-based search space reduction and adaptive exploration–exploitation strategy to accelerate materials design, high-throughput synthesis and characterization, and electrochemical performance optimization. CRESt enables monitoring with cameras and the generation of vision-language-model-driven hypotheses to diagnose and correct experimental anomalies. Applied to electrochemical formate oxidation, CRESt explored more than 900 catalyst chemistries and 3,500 electrochemical tests within 3 months, identifying a state-of-the-art catalyst in the octonary chemical space (Pd–Pt–Cu–Au–Ir–Ce–Nb–Cr) that exhibits a 9.3-fold improvement in cost-specific performance.

Design of experiments (DOE) for targeted materials discovery remains a fundamental challenge. Even subtle variations in chemistry or processing could yield markedly different properties, and experimental optimization is often constrained by human and machine errors, cost and limited throughput¹⁰. Active learning (AL), in the form of Gaussian process-based Bayesian optimization (GP-BO), provides a DOE framework by balancing exploration of uncertain parametric regions and exploitation of already-promising regions¹¹. However, conventional unimodal AL operates on a single stream of data of fixed dimensions—for example, mapping the ratios of a particular set of quinary elements directly to alloy properties—ignoring the diverse knowledge sources routinely used by human experts, such as literature, previous experience, microstructural features, intuition and hypothesis making.

Here we introduce Copilot for Real-world Experimental Scientists (CRESt), a multimodal platform that integrates large multimodal models (LMMs) with robotic automation¹². Chemical compositions are precisely controlled by robotic preparation, whereas large language models (LLMs) embed literature knowledge to guide design and explorations in reduced-dimensional spaces. Beyond chemical descriptors,

CRESt incorporates microstructural features from high-throughput scanning electron microscopy (SEM) imaging and computer vision analysis, enriching the surrogate model and accelerating optimization. Dimensionality reduction using principal component analysis (PCA) preserves variance before AL is performed in the latent space, with candidate solutions mapped back to compositions by sequential least squares programming (SLSQP).

Real-world experiments also face reproducibility challenges arising from mechanical, electrical, thermal, magnetic and even organizational and cognitive factors that remain invisible to standard algorithms. To address this, CRESt enables voice and text interaction, as well as camera-based monitoring of experiments. Leveraging vision language models (VLMs), CRESt analyses irreproducibility issues and proposes corrective procedures, offering a breadth of expertise that surpasses any individual researcher.

We demonstrate CRESt for the electrochemical formate oxidation reaction. Within 3 months, the platform explored more than 900 chemistries and 3,500 electrochemical tests, leading to the discovery of an octonary multi-element catalyst (Pd–Pt–Cu–Au–Ir–Ce–Nb–Cr). This catalyst achieved a 9.3-fold improvement in cost-specific performance

¹Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁴Department of Physics, Tamkang University, New Taipei City, Taiwan. ⁵MIT.nano, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁶Department of Architecture, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁷National Synchrotron Radiation Research Center, Hsinchu, Taiwan. ⁸Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁹These authors contributed equally: Zhen Zhang, Zhichu Ren, Chia-Wei Hsu, Weibin Chen. ✉e-mail: liju@mit.edu

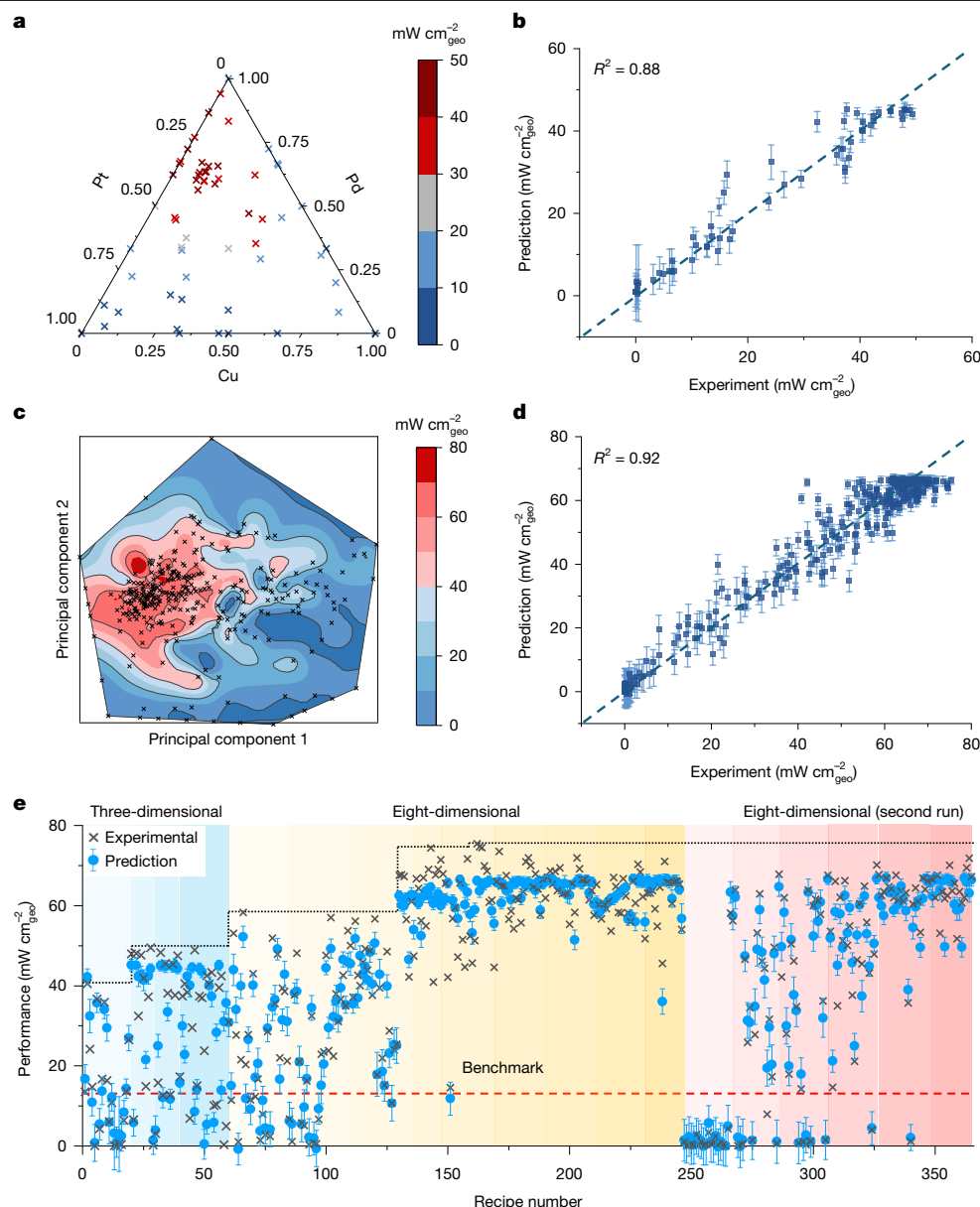


Fig. 1 | AL discovery of multi-element electrocatalysts for formate oxidation reactions. **a**, The trajectory plot of the AL optimization results for ternary electrocatalysts (Pd–Pt–Cu). The triangular axes indicate the composition ratios of the three metallic elements. **b**, Cross-validation plot for the ternary AL, with prediction versus experiments. **c**, PCA plot with contour lines for octonary alloys (Pd–Pt–Cu–Au–Ir–Ce–Nb–Cr). **d**, Cross-validation plot for the octonary AL,

with prediction versus experiments. **e**, The complete AL campaign plot with the baseline AL model, with the performance value compared with the recipe number. Blue shading indicates the ternary optimization region. The yellow and red shadings indicate the octonary optimization. The colour bars in **a** and **c** indicate the maximum power density ($\text{mW cm}_{\text{geo}}^{-2}$) of a catalyst. Error bars in **b**, **d**, **e** represent the predictive uncertainty estimated by the GP-BO model.

(power density normalized by catalyst cost) over the pure Pd benchmark (Fig. 1 and Supplementary Table 1). Mechanistic investigations showed enhanced tolerance to CO and hydrogen poisoning, consistent with electronic tuning in rational catalyst design. Finally, in a direct formate fuel cell, the optimized recipe delivered record power density at just one-quarter the precious-metal loading of previous devices. Together, these results demonstrate the potential of LMM- and robotics-driven materials experimentation.

Copilot for Real-world Experimental Scientists

We introduce our LMM-assisted robotic platform named CRES^t¹² (Extended Data Figs. 1 and 2 and Supplementary Videos 1 and 2). CRES^t consists of three main components: a user interface; the LMM-powered back end; and a range of actuators. The user interface supports both

voice and text interactions, connecting the LMM back end and allowing researchers to control the robotic platform without computer programming experience. The actuators encompass various robotic equipment, including sample preparation tools, such as the liquid-handling robot and the carbothermal shock system¹³; sample testing tools, such as an automated electrochemical workstation; characterization equipment, including X-ray diffraction (XRD) and software-driven SEM; and auxiliary devices, such as automated peristaltic pumps and gas valves. To facilitate coordinated electrochemical research, we customized these actuators extensively through coding, three-dimensional (3D) printing, laser cutting, electrical modifications, and so on. All actuators operate remotely through our customized Python codes. Apart from implementing instructions, CRES^t actively converses with human scientists in natural languages, making observations and hypotheses¹⁴ along the way.

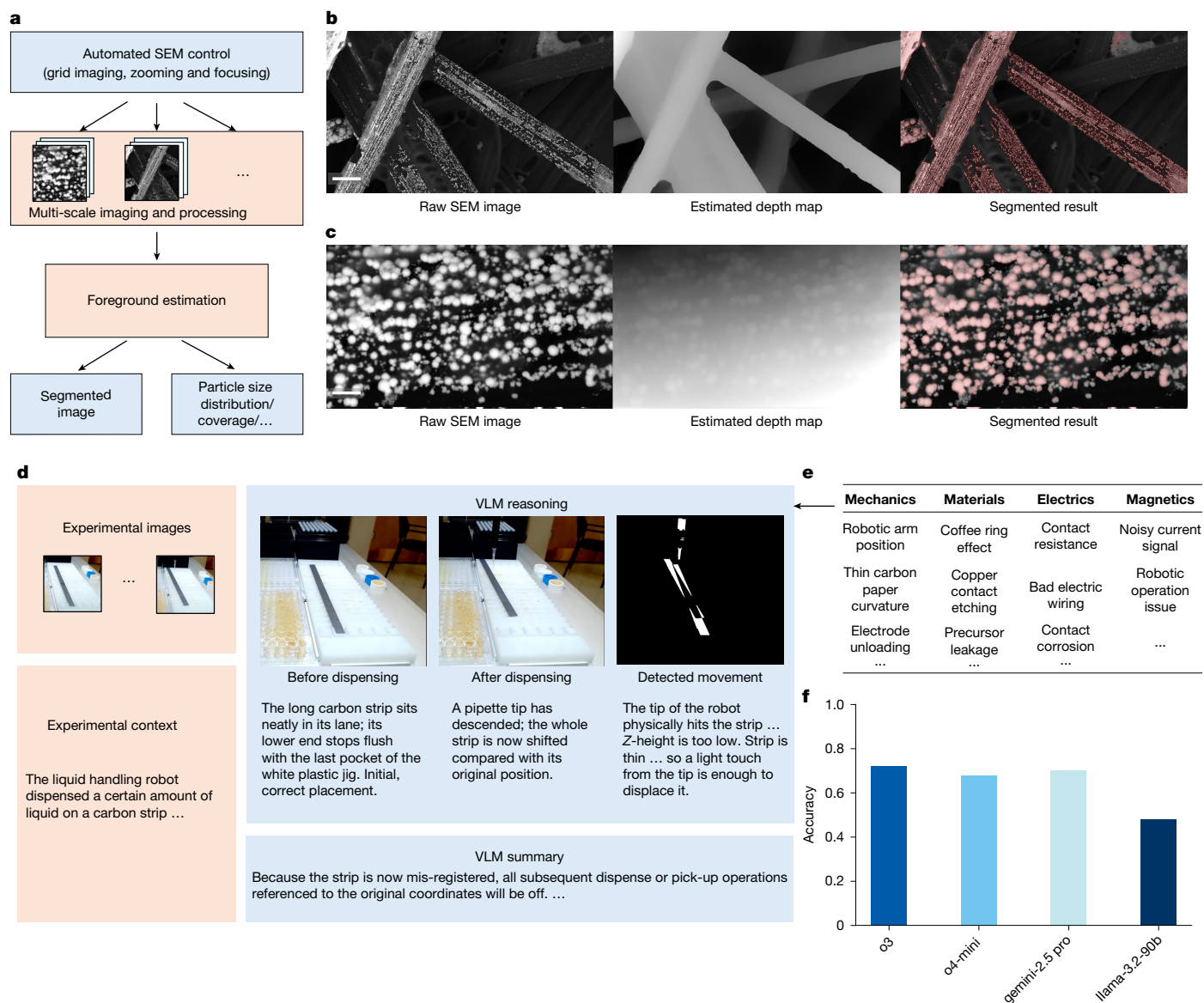


Fig. 2 | Vision analysis modules of the CREST platform. a–c, Computer vision analysis of microstructures by the automated SEM control. Pipeline of the multi-scale image feature acquisition and analysis (**a**) and visualization of microstructure analysis (**b,c**). The first column shows the original SEM images. The second column shows the predicted depth maps, highlighting the separation of foreground and background. The third column shows particles successfully identified by our algorithm, marked in red. The recipe shown is for $\text{Pd}_{0.232}\text{Pt}_{0.133}\text{Ir}_{0.330}\text{Nb}_{0.176}\text{Cr}_{0.128}$. Scale bar, 8 μm (**b**); 1 μm (**c**). **d–f,** Experimental

diagnosis by the VLM. An illustrative example of leveraging a VLM is shown in **d** to diagnose issues during the automated liquid pipetting process. Provided with a minimal experimental context by the human operator, the VLM performed reasoning, invoked tools such as movement detection to identify shifts in the carbon paper, and subsequently proposed relevant solutions. Diverse sources of experimental failure in the CREST pipeline are shown in **e**. Overall success rate of debugging experimental issues with different VLMs are shown in **f**.

Knowledge-assisted AL

Conventional unimodal AL relies on a single data stream of small, fixed dimensions, overlooking the richness and complexity of real-world materials and experiments. Embedding methods have proven effective across a range of downstream tasks in integrating diverse information sources^{15,16}, such as 3D molecular and crystal structures¹⁷, property databases¹⁸, lab logs and discussion notes, into unified vector representations. Here, modifying the standard Bayesian optimization method (Supplementary Note 1), we introduce a knowledge-assisted Bayesian optimization (KABO) method that leverages LMMs to bridge this gap (Supplementary Note 2). For the text-embedding-based dimensionality reduction, CREST first searches the literature for comprehensive descriptions of metallic elements relevant to formate oxidation

reactions and then embeds these responses into alloy-specific vectors using word embedding models¹⁹. For the image embedding, we developed a new workflow to perform high-throughput SEM imaging (Fig. 2a–c, Extended Data Fig. 2b and Supplementary Fig. 1) for our samples with different imaging resolutions (Extended Data Figs. 2b and 3). Taking advantage of the near-spherical morphology of our catalyst particles in most cases, we elected to extract four morphological features (slope of the particle distribution function, bias of the particle distribution function, particle coverage ratio and the particle count per unit area) for the AL training (Supplementary Note 4). Next, KABO applies PCA to update the latent space, reducing dimensionality while explaining at least 80% of the variability in the data, thus effectively representing both literature-derived knowledge and known experimental features. Using the knowledge gradient acquisition function²⁰, KABO

optimizes in this latent space, mapping candidate solutions back to practical materials compositions by SLSQP.

We also address the challenge of adaptively balancing exploration and exploitation in Bayesian optimization, which conventionally depends on a manually tuned weight parameter (κ) in the acquisition function. Here, we incorporate a dynamic tuning mechanism inspired by the Bayesian optimization with policy improvement constraints (BOPIC) framework in reinforcement learning²¹, introducing a Lagrange multiplier (λ) that adaptively adjusts the trade-off between exploration and exploitation without manual intervention (Supplementary Note 3).

We used CRES to discover catalysts for the electrochemical oxidation of formate. We elected to optimize the maximum power density of a fuel cell device using the synthesized catalyst for formate oxidation, with oxygen reduction reaction at the counter electrode. To simplify, we assumed that the current and potential data from the three-electrode linear sweep voltammetry reflect the performance of the formate oxidation half-cell and a constant overpotential for oxygen reduction reaction²² (Supplementary Note 1). These choices provided a relatively consistent ranking of the catalyst performance when compared with fuel cell experiments (Supplementary Figs. 2 and 3). We independently evaluated the performance of KABO, BOPIC and standard Bayesian optimization through experimentation (Extended Data Fig. 4). Notably, both KABO and BOPIC achieved continuous improvements by identifying higher-performing recipes batch after batch, whereas standard Bayesian optimization required several batches before substantial gains were obtained. This delay could be detrimental in high-throughput settings, in which limited optimization may cause premature termination of experiments. By pre-screening the design space, KABO provided substantial early-stage benefits and reached convergence first. BOPIC proved most effective in sampling high-performing recipes, thereby generating a richer candidate pool for downstream selection. We further benchmarked the convergence efficiency of the three algorithms using a trained gradient boosting model to simulate the optimization process (Supplementary Fig. 4). It is shown that, on average, there is a 36% (KABO) and 27% (BOPIC) performance improvement compared with Bayesian optimization under the same number of experimental trials. In the quest for high-performance samples (here, more than five times the performance of the benchmark), it only required 25% (KABO) and 47% (BOPIC) of the recipes. Notably, KABO algorithm shows equally good performance when using open-source LLMs under various conditions, demonstrating the generalizability of our method (Supplementary Note 2 and Supplementary Fig. 5). For the image-embedding method, to further validate the use of the selected morphological features for AL, we designed various experimental settings and ablation studies. We found that microstructural features, when optimized jointly with elemental ratios, led to faster convergence (Supplementary Fig. 6).

Owing to the complexity of real-world experiments with Bayesian optimization, we first attempted an optimization process in a ternary chemical space (Pd–Pt–Cu), a subset of the octonary chemical space (Pd–Pt–Cu–Au–Ir–Ce–Nb–Cr). As a result, in the ternary space, we identified an optimal catalyst composition with less than 60 experimental recipes, from around 5×10^5 potential recipes (Fig. 1a and Supplementary Note 5), with cross-validation confirming model accuracy against experiments (Fig. 1b). The optimized catalyst, $\text{Pd}_{0.635}\text{Pt}_{0.258}\text{Cu}_{0.107}$, delivered around 3.5 times the power density of pure Pd with the same molarity of metallic loading, which is commonly used as the benchmark catalyst²³. We then progressed to the octonary space, which introduces exponentially larger search space (around 2×10^{17} potential recipes). We achieved optimized performance, with PCA indicating a primary peak with smaller satellites (Fig. 1c; for further analysis of the physical importance of the two primary axes from PCA, see Supplementary Fig. 7), and cross-validation showing a good alignment with experimental data (Fig. 1d). To validate global optimization, we started a second trial of AL, but favoring exploration. The optimization again converged to a similar value, potentially confirming global optimization (Fig. 1e).

The optimized catalyst, $\text{Pd}_{0.487}\text{Pt}_{0.185}\text{Cu}_{0.018}\text{Ir}_{0.037}\text{Ce}_{0.106}\text{Nb}_{0.168}$, showed 5.6 times the power density, whereas the catalyst $\text{Pd}_{0.381}\text{Pt}_{0.080}\text{Cu}_{0.009}\text{Au}_{0.004}\text{Ir}_{0.02}\text{Ce}_{0.086}\text{Nb}_{0.338}\text{Cr}_{0.082}$ showed 9.3 times the power density normalized by catalyst cost, compared with the pure Pd catalyst benchmark (Supplementary Table 1).

Experimental error diagnosis by VLM

At the early stages of our electrochemical experiments, poor reproducibility emerged as the main obstacle and time sink. Although large datasets were generated, inconsistencies across trials rendered them unreliable for AL training (Supplementary Fig. 8a and Supplementary Data 1 and 2). Careful inspection showed numerous hidden errors—mechanical, thermal, electrical, magnetic and even organizational and cognitive—that subtly altered conditions despite seemingly identical steps, initially requiring extensive human oversight to debug (Fig. 2d,e). The advent of VLMs provided a new approach: by coupling computer vision with broad domain knowledge, AI proposed sources of irreproducibility and plausible corrective procedures (Supplementary Video 1). For example, the VLM can invoke tools such as coordinate alignment and movement detection to diagnose issues in which the pipette displaces the carbon paper, even from micrometre-scale height changes (Fig. 2d). In another case, a 1-mm deviation in a 1-cm sample geometry introduces an error of approximately 10%, substantially degrading AL performance. Although laser-cut wooden stages were designed to ensure uniformity, the VLM identified charring artefacts that caused dimensional variation, attributing them to surface non-uniformity in laser absorption (Supplementary Fig. 9). This insight prompted a switch to stainless steel stages, yielding improved consistency, with minimal human guidance during the debugging process. To systematically evaluate this ability, we curated a troubleshooting log ('Criminals in the AI City') and constructed a question and answer set from recorded failures (Appendix in the Supplementary Information). Benchmarking several VLMs demonstrated promising accuracies—72% (OpenAI o3), 70% (Gemini-2.5 Pro), 68% (OpenAI o4-mini) and 48% (Llama-3.2-90B)—highlighting their potential as practical experimental assistants (Fig. 2f and Extended Data Fig. 5). After hypothesis making¹⁴, checking and correcting the root causes, reproducibility was markedly improved (Supplementary Fig. 8b).

Performance testing

We leveraged in situ electrochemical deposition for the purpose of high-throughput testing. The synthesized nanoparticles showed uniform distribution across the entire electrode surface and on single carbon fibres (Fig. 3a). The optimized catalyst demonstrated a homogeneous mix of elements (Fig. 3b) and a single face-centred cubic (FCC) phase in the XRD analysis (Fig. 4a and Supplementary Fig. 10a,b). In the cyclic voltammetry analysis for formate oxidation, a pronounced peak appears during the forward scan, followed by a marked decrease in the current density as oxidation continues (Fig. 3c). This decrease is due to the formation of metal oxides, such as PdO_x , which are inactive for formate oxidation^{24,25}. A pronounced peak is observed in the reverse scan, which corresponds to a surge in formate oxidation activity when metal oxides are reduced back to their metallic form²⁶. In three-electrode testing, the optimized catalysts demonstrated a much higher current density, despite having an equivalent molar metallic catalyst loading, compared with the benchmark catalyst. The enhanced activity allowed for a reduction in the usage of precious metals. In the fuel cell tests, our catalyst (with 2.0 mg cm^{-2} of entire metal loading, and thus around 1.2 mg cm^{-2} of precious-metal loading) demonstrated a peak power density of 325 mW cm^{-2} , higher than the benchmark Pd catalyst with various loadings ($0.5\text{--}4.0 \text{ mg cm}^{-2}$), either tested by us (Fig. 3d and Supplementary Fig. 11) or reported in the literature^{23,27–35} (Fig. 3e). Furthermore, we found that the difference of catalyst activity could

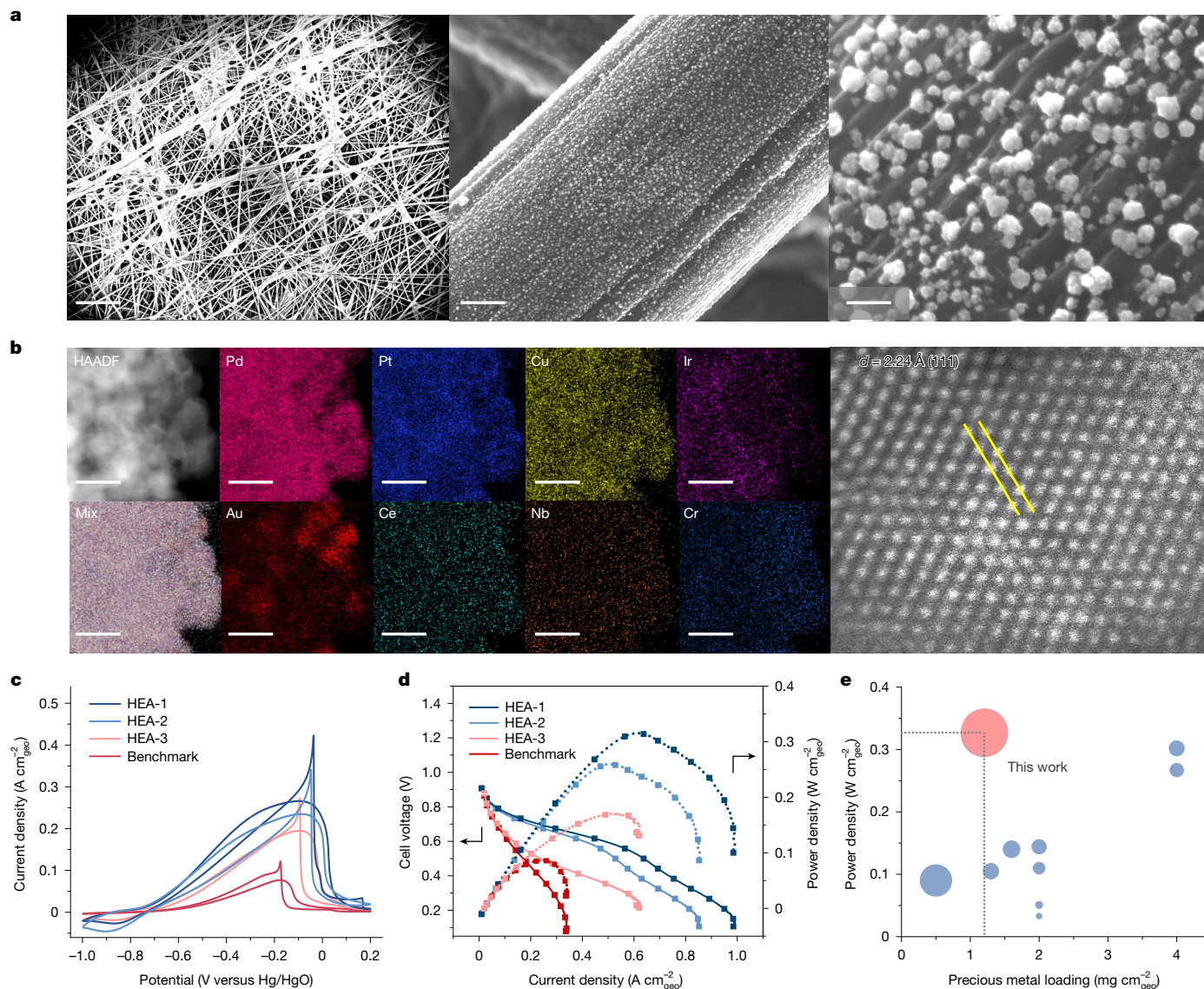


Fig. 3 | Morphology characterization and electrochemical testing of the optimized catalysts for the formate oxidation reaction. **a**, SEM images of the electrochemically prepared catalysts on the carbon substrate. Uniform catalyst distribution was observed on a single carbon fibre. Scale bar, 200 μm , 1 μm and 40 nm (left to right). **b**, High-angle annular dark-field scanning transmission electron microscopy and EDS of the optimized catalysts. Scale bar, 10 nm. **c**, The cyclic voltammetry plot of selected chemistries in a three-electrode setup under nitrogen sparging at ambient conditions. A low catalyst precursor solution loading of 10 $\mu\text{l cm}^{-2}$ (corresponding to about 0.05 mg cm^{-2} for pure Pd) was used for fast catalyst screening. The scan rate was 50 mV s^{-1} .

at least partly come from the distinct chemical nature of these catalysts, as indicated by a consistent trend of voltage/current slope in the activation region in both the three-electrode and fuel cell testing (Supplementary Fig. 12).

Mechanistic analysis

For the octonary catalyst (HEA-8D), Pd and Pt serve as primary active elements for formate oxidation reactions²³. Small amounts of Au and Ir, included in the optimized composition, have also been previously used in direct formate fuel cells²⁶. The remaining elements (Cu, Ce, Nb and Cr) could alloy with the primary elements to provide tailored coordination environments and reduce catalyst cost^{13,36,37}. The optimized catalyst,

d, The polarization curve of selected chemistries tested in the direct formate fuel cells at 60 °C in an electrolyte of 2.0 M KOH and 1.0 M HCOOK. To enable comparison with literature, all samples had a total catalyst loading of 2.0 mg cm^{-2} . Details are in the Methods. **e**, Comparison plot of our optimized catalysts with results in the literature under similar testing conditions (typically at 60 °C in an electrolyte of 2.0 M KOH and 1.0 M HCOOK)^{23,27–35}. The catalyst HEA-1, HEA-2, HEA-3 and Benchmark represent the composition of $\text{Pd}_{0.487}\text{Pt}_{0.185}\text{Cu}_{0.018}\text{Ir}_{0.037}\text{Ce}_{0.106}\text{Nb}_{0.168}$, $\text{Pd}_{0.381}\text{Pt}_{0.080}\text{Cu}_{0.009}\text{Au}_{0.004}\text{Ir}_{0.02}\text{Ce}_{0.086}\text{Nb}_{0.338}\text{Cr}_{0.082}$, $\text{Pd}_{0.635}\text{Pt}_{0.258}\text{Cu}_{0.107}$ and Pd, respectively. HEA, high-entropy alloy.

$\text{Pd}_{0.381}\text{Pt}_{0.080}\text{Cu}_{0.009}\text{Au}_{0.004}\text{Ir}_{0.02}\text{Ce}_{0.086}\text{Nb}_{0.338}\text{Cr}_{0.082}$, demonstrated a single FCC phase in the XRD analysis with Rietveld refinement (Fig. 4a). The lattice parameters for Pd and HEA-8D were determined as 3.896 Å and 3.899 Å, respectively, indicating that here alloying does not lead to huge lattice deformation, potentially ensuring structural integrity (Supplementary Table 2).

To probe the oxidation states and local coordination environments of the optimized catalyst, in situ XAS was used. Owing to the complexity of the composition, we elected to investigate the primary catalytic elements Pd and Pt. X-ray absorption near-edge structure analysis showed that both Pd and Pt retained their metallic states during formate oxidation (Supplementary Fig. 13), an important factor given that both PdO and PtO₂ exhibit negligible catalytic activity²⁶. A slight shift

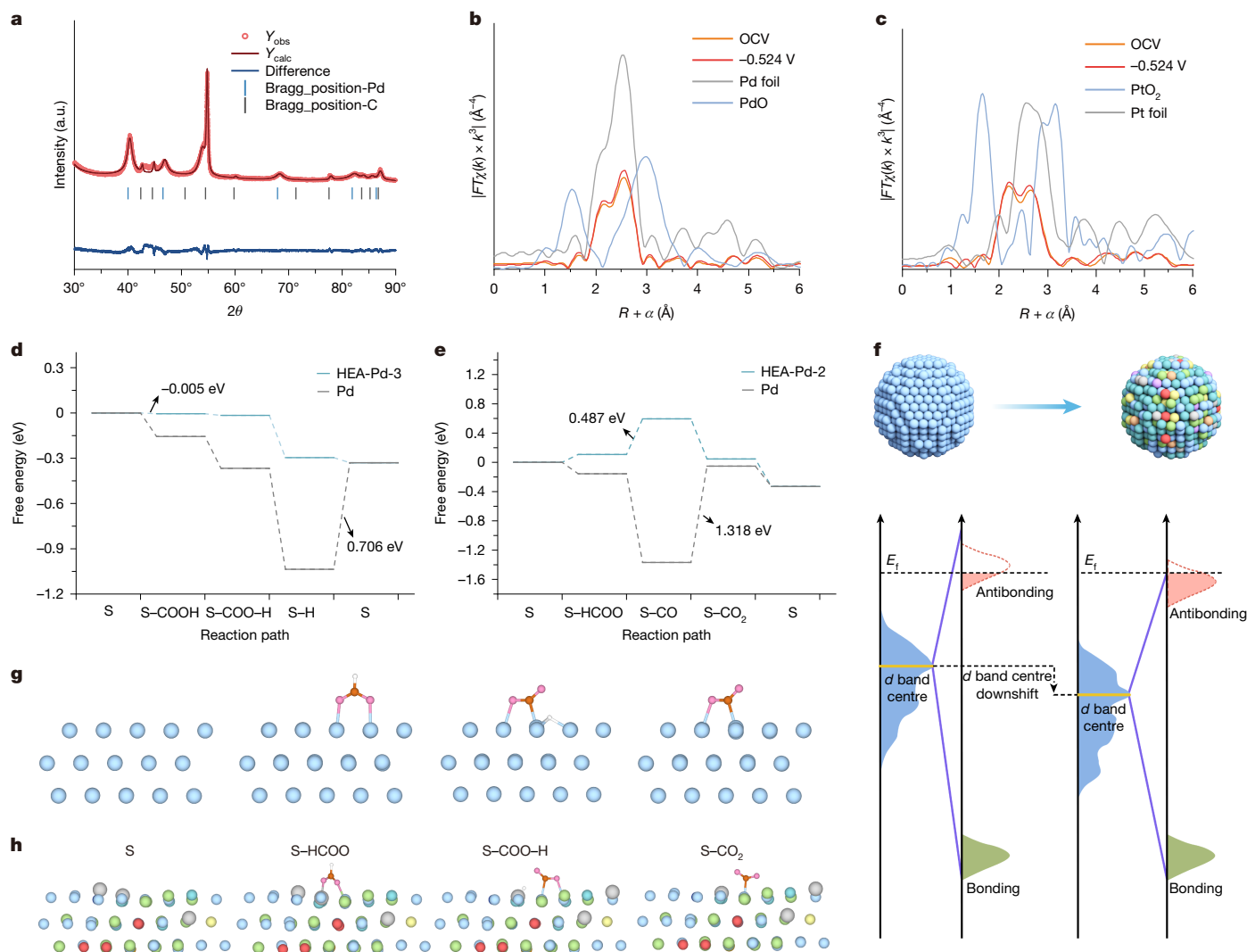


Fig. 4 | Mechanistic analysis of the optimized multi-element electrocatalysts.

a, XRD patterns with Rietveld refinement for the octonary catalyst, showing a singular FCC solid solution phase. **b,c**, In situ Fourier transform EXAFS spectra of the octonary catalyst collected at the Pd edge: the open circuit voltage spectrum, the potential-applied spectrum (at -0.524 V compared with Hg/HgO in 1.0 M KOH), as compared with reference samples, including metallic Pd and PdO (**b**) and spectra collected at the Pt edge: the open circuit voltage spectrum, the potential-applied spectrum (at -0.524 V compared with Hg/HgO in 1.0 M KOH), as compared with reference samples, including metallic Pt and PtO₂ (**c**). **d,e**, Reaction pathways computed by the DFT calculations. Benchmark Pd catalysts show a

reaction barrier of 0.706 eV in the direct pathway (**d**) and 1.318 eV in the indirect pathway (**e**), whereas the high-entropy catalysts showed a reduced activation barrier of -0.005 eV in the direct pathway (**d**) and 0.487 eV in the indirect pathway (**e**). **f**, Schematic of the standard Pd particle and the octonary alloy particle, and the position of the d-band centre. **g,h**, Schematic of the model and key intermediates and adsorbates on the Pd (**g**) and the optimized octonary catalyst (**h**). From top to bottom: the pure Pd reaction model and the HEA-Pd-3 reaction model. From left to right: the clean model surface, the model with HCOO adsorbed, the model with COO-H adsorbed and the model with CO₂ adsorbed. OCV, open circuit voltage.

in the Pd absorption edge relative to Pd foil indicated local structural perturbations arising from alloying. Furthermore, extended X-ray absorption fine structure (EXAFS) spectra showed distinct Pd-Pd and Pd-Pt coordination peaks, indicative of strong electronic interactions associated with tuned catalytic activity (Fig. 4b,c). Quantitative fitting of the EXAFS data showed minor variations in bond lengths and coordination numbers, suggesting a robust atomic structure in HEA-8D during the operation (Supplementary Note 6, Supplementary Figs. 14 and 15 and Supplementary Table 3).

A structural model was constructed based on refined XRD results, and high-throughput calculations were performed to screen and identify stable structures for subsequent simulations (Supplementary Fig. 16 and Supplementary Note 7). Reaction pathways and free energy profiles for both direct and indirect pathways were calculated for multiple possible adsorption sites (Supplementary Figs. 17–27 and Supplementary Notes 8 and 9) and benchmarked against a conventional Pd catalyst.

The octonary catalyst demonstrates strong resistance to hydrogen poisoning (direct pathway, Fig. 4d,g,h) and CO poisoning (indirect pathway, Fig. 4e) relative to pure Pd. Notably, the Pd site in the HEA exhibits a substantially lower potential-determining step barrier for the indirect pathway than that of pure Pd. To further explain the mechanism, projected density of states analyses were performed for the Pd sites in HEA-8D and pure Pd in DFT calculations (Supplementary Fig. 28). The *d*-band centres of Pd-1, Pd-2 and Pd-3 are all much lower than that of pure Pd, indicating weakened hydrogen binding due to reduced orbital overlap (Fig. 4f), which promotes desorption. Moreover, isotope study (with electrolytes containing HCOONa or DCOONa) and CO stripping experiments were performed to experimentally validate the tolerance of the catalyst to surface H_{ads} and CO_{ads}, respectively (Supplementary Figs. 29 and 30). These findings demonstrate that tailored alloying and atomic-level structural tuning effectively regulate surface electronic properties and reaction energetics.

Conclusion

In summary, we developed CReSt, an LMM- and robotic-driven platform for experimental materials discovery. Our method incorporates previous literature and database knowledge, human and machine hypothesis making, composition tuning and process design, and microstructural features into the materials experimentation framework. We further leverage VLMs to analyse experimental processes to find and correct the root causes of anomalies. Mechanistic studies confirm that the optimized multi-element catalyst exhibits enhanced tolerance to hydrogen and CO poisoning, attributed to tailored atomic and electronic structures. This demonstrates that LMM-based approach could more effectively explore the rich and complex real-world materials and experiments, uniting automation with intelligence.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-025-09640-5>.

- Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
- Dai, T. et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature* **635**, 890–897 (2024).
- Joung, J. F. et al. Electron flow matching for generative reaction mechanism prediction. *Nature* **645**, 115–123 (2025).
- Koscher, B. A. et al. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science* **382**, eadi1407 (2023).
- Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
- Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
- Tom, G. et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633–9732 (2024).
- Ren, Z., Ren, Z., Zhang, Z., Buonassisi, T. & Li, J. Autonomous experiments using active learning and AI. *Nat. Rev. Mater.* **8**, 563–564 (2023).
- Noorden, R. V. & Perkel, J. M. AI and science: what 1,600 researchers think. *Nature* **621**, 672–675 (2023).
- Attia, P. M. et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **578**, 397–402 (2020).
- Balandat, M. et al. BoTorch: a framework for efficient Monte-Carlo Bayesian optimization. *NeurIPS* <https://doi.org/10.48550/arxiv.1910.06403> (2020).
- Ren, Z., Zhang, Z., Tian, Y. & Li, J. CReSt – copilot for real-world experimental scientist. Preprint at <https://doi.org/10.26434/chemrxiv-2023-tnz1x-v4> (2023).
- Yao, Y. et al. Carbothermal shock synthesis of high-entropy-alloy nanoparticles. *Science* **359**, 1489–1494 (2018).
- Park, Y. J. et al. Can ChatGPT be used to generate scientific hypotheses? *J. Materiom.* **10**, 578–584 (2024).
- Nussbaum, Z., Morris, J. X., Duderstadt, B. & Mulyar, A. nomic embed: training a reproducible long context text embedder. Preprint at <https://doi.org/10.48550/arxiv.2402.01613> (2024).
- Ramos, M. C., Michtav, S. S., Porosoff, M. D. & White, A. D. Bayesian optimization of catalysis with in-context learning. Preprint at <https://doi.org/10.48550/arxiv.2304.05341> (2025).
- Park, Y. J., Kumaran, M., Hsu, C.-W., Olivetti, E. & Li, J. Contrastive learning of English language and crystal graphs for multimodal representation of materials knowledge. Preprint at <https://doi.org/10.48550/arxiv.2502.16451> (2025).
- Park, Y. J., Jerng, S. E., Yoon, S. & Li, J. 1.5 million materials narratives generated by chatbots. *Sci. Data* **11**, 1060 (2024).
- Neelakantan, A. et al. Text and code embeddings by contrastive pre-training. Preprint at <https://doi.org/10.48550/arxiv.2201.10005> (2022).
- Frazier, P. I., Powell, W. B. & Dayanik, S. A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optim.* **47**, 2410–2439 (2008).
- Chen, E., Hong, Z.-W., Pajarinen, J. & Agrawal, P. Redeeming intrinsic rewards via constrained optimization. *NeurIPS* <https://doi.org/10.48550/arxiv.2211.07627> (2022).
- Xu, H. et al. An actor–critic algorithm to maximize the power delivered from direct methanol fuel cells. *Nat. Energy* **10**, 951–961 (2025).
- An, L. & Chen, R. Direct formate fuel cells: a review. *J. Power Sources* **320**, 127–139 (2016).
- Choun, M., Hong, S. & Lee, J. Adsorbed hydrogen as a site-occupying species in the electrocatalytic oxidation of formate on Pd/C in alkaline medium. *J. Electrochem. Soc.* **165**, J3266–J3270 (2018).
- Wang, Q. et al. Transition from core-shell to janus segregation pattern in AgPd nanoalloy by Ni doping for the formate oxidation. *Appl. Catal. B Environ.* **270**, 118861 (2020).
- Folkman, S. J., González-Cobos, J., Giancola, S., Sánchez-Molina, I. & Galán-Mascarós, J. R. Benchmarking catalysts for formic acid/formate electrooxidation. *Molecules* **26**, 4756 (2021).
- Tran, K., Nguyen, T. Q., Bartrom, A. M., Sadiki, A. & Haan, J. L. A fuel-flexible alkaline direct liquid fuel cell. *Fuel Cells* **14**, 834–841 (2014).
- Bartrom, A. M. & Haan, J. L. The direct formate fuel cell with an alkaline anion exchange membrane. *J. Power Sources* **214**, 68–74 (2012).
- Bartrom, A. M. et al. Optimization of an anode fabrication method for the alkaline direct formate fuel cell. *J. Power Sources* **229**, 234–238 (2013).
- Lan, L. et al. Membrane-less direct formate fuel cell using an Fe–N-doped bamboo internode as the binder-free and monolithic air-breathing cathode. *ACS Appl. Mater. Inter.* **12**, 27095–27103 (2020).
- Bartrom, A. M., Ognibene, G., Ta, J., Tran, J. & Haan, J. L. Catalysts for alkaline direct ethanol and direct formate fuel cells. *Electrochem. Soc. Trans.* **50**, 1913–1918 (2013).
- Zeng, L., Tang, Z. K. & Zhao, T. S. A high-performance alkaline exchange membrane direct formate fuel cell. *Appl. Energy* **115**, 405–410 (2014).
- Li, Y., Feng, Y., Sun, X. & He, Y. A sodium-ion-conducting direct formate fuel cell: generating electricity and producing base. *Angew. Chem. Int. Ed.* **56**, 5734–5737 (2017).
- Castello, C. et al. Direct formate anion exchange membrane fuel cells with a PdAu bimetallic nanoparticle anode electrocatalyst obtained by metal vapor synthesis. *Energy Adv.* **3**, 2520–2529 (2024).
- Sun, X., Li, Y., An, L. & Lv, X. Comparative performance evaluation of self-basifying direct formate fuel cells. *J. Electrochem. Soc.* **166**, F768–F773 (2019).
- Abdelhafiz, A., Wang, B., Harutyunyan, A. R. & Li, J. Carbothermal shock synthesis of high entropy oxide catalysts: dynamic structural and chemical reconstruction boosting the catalytic activity and stability toward oxygen evolution reaction. *Adv. Energy Mater.* **12**, 2200742 (2022).
- Tajuddin, A. A. H. et al. Corrosion-resistant and high-entropic non-noble-metal electrodes for oxygen evolution in acidic media. *Adv. Mater.* **35**, e2207466 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2025

Methods

Implementation of the CRES platform

The system supports a unified workflow in which the user engages with our AI platform, CRES, a large language-model-driven layer, through natural language prompts (Extended Data Fig. 2a). CRES interprets these requests and orchestrates the execution of relevant routines, here collectively referred to as 'end-effectors'. Each end-effector is accessed through standardized HTTP requests or Python calls. Some routines query local and remote databases (for example, Materials Project) for data retrieval, whereas others directly manipulate physical laboratory hardware, such as a liquid-handling robot (Opentrons), multi-axis robotic arm (xArm) or specialized lab components for pumping, gas flow control or illumination.

In the present implementation, many devices rely on PyAutoGUI for emulating keyboard and mouse interactions, enabling a universal approach to instrument control irrespective of the availability of hardware-specific application programming interfaces. This is expected to evolve as laboratory equipment increasingly offers dedicated endpoints compatible with AI-ready protocols. Although not limited to these, commonly invoked routines involve automatic experimental preparation (for example, pipetting liquids with Opentrons), advanced sample handling (xArm) or environment manipulation (pump and gas valves). Thus, the integration of CRES with both digital data repositories and physical lab instrumentation consolidates a broad range of abilities within a single AI-governed function pool, reducing operator overhead while ensuring reproducibility and streamlined experimental workflows. Specific AI approaches are discussed in detail in Supplementary Notes 1–3.

The autonomous materials characterization framework integrates three GPT (generative pre-trained transformer)-based agents with a Phenom Pharos scanning electron microscope using Python (Extended Data Fig. 2b). At the top layer, CRES serves as the user interface, capturing natural language descriptions of desired imaging objectives. Rather than issuing low-level instructions, CRES converts these requests into structured goals and forwards them to the SEM actuation agent, which is also powered by GPT. The scanning electron microscope actuation agent then interprets the goals and translates them into direct scanning electron microscope commands (for example, zoom in, zoom out and stage navigation). Autofocus, brightness-contrast and image acquisition routines proceed automatically at each iteration. Following each round of data capture, the scanning electron microscope actuation agent sends the newly obtained micrographs and metadata to the vision agent, built on GPT-4V. The vision agent assesses whether the micrographs fulfil the predefined objective and, if necessary, recommends subsequent zoom or stage manoeuvres back to the scanning electron microscope actuation agent. A set-of-mark layer augments the raw images with reference markers before they are processed by GPT-4V, allowing explicit spatial referencing of specific features and ensuring more precise stage navigation. This cycle repeats until the vision agent confirms that the target electron micrograph has been successfully acquired, at which point the scanning electron microscope actuation agent returns the final image and summary report to CRES. Detailed methods for the computer vision analysis on catalyst particles are provided in Supplementary Note 4.

Materials and reagents

Palladium(II) chloride ($\geq 99.9\%$), chloroplatinic acid hexahydrate (ACS reagent, $\geq 37.50\%$ Pt basis), platinum(IV) chloride ($\geq 99.99\%$ trace metals basis), copper(II) nitrate trihydrate (puriss. p.a., 99–104%), gold(III) chloride ($\geq 99.99\%$ trace metals basis), gold(III) chloride trihydrate ($\geq 99.9\%$ trace metals basis), hydrogen hexachloroiridate(IV) hydrate (99.9% trace metals basis), cerium(III) nitrate hexahydrate (99% trace metals basis), niobium(V) chloride ($\geq 99.9\%$ trace metals basis), chromium(III) chloride (anhydrous, 99.99% trace metals

basis), bismuth(III) chloride (99.99% trace metals basis), tin(II) chloride ($\geq 99.99\%$ trace metals basis), yttrium(III) chloride (anhydrous, powder, 99.99% trace metals basis), iron(III) chloride (anhydrous for synthesis), zinc(II) chloride (reagent grade, $\geq 98\%$), indium(III) chloride (99.999% trace metals basis), praseodymium(III) chloride (anhydrous, powder, 99.99% trace metals basis), lanthanum(III) chloride heptahydrate (ACS reagent), ruthenium(III) chloride hydrate ($\geq 99.9\%$ trace metals basis), vanadium(III) chloride (99+), Tungsten(VI) chloride ($\geq 99.9\%$ trace metals basis), nickel(II) chloride hexahydrate (99.9% trace metals basis), cobalt(II) chloride hexahydrate (ACS reagent, 98%), ethyl alcohol (pure, 200 proof, anhydrous, $\geq 99.5\%$), hydrochloric acid (ACS reagent, 37%), Nafion 115 ionomer solution (5 wt%), potassium formate (ReagentPlus, 99%), potassium hydroxide (ACS reagent, $\geq 85\%$, pellets), oleylamine (technical grade, 70%), hexadecyltrimethylammonium chloride (CTAC, $\geq 98.0\%$), palladium(II) acetylacetonate ($\text{Pd}(\text{acac})_2$, Umicore, 99%), platinum(II) acetylacetonate ($\text{Pt}(\text{acac})_2$, $\geq 99.98\%$ trace metals basis), copper(II) acetylacetonate ($\text{Cu}(\text{acac})_2$, $\geq 99.9\%$ trace metals basis), iridium(III) acetylacetonate ($\text{Ir}(\text{acac})_3$, 97%), cerium(III) acetylacetonate hydrate ($\text{Ce}(\text{acac})_3 \cdot x\text{H}_2\text{O}$), chromium(III) acetylacetonate ($\text{Cr}(\text{acac})_3$, 97%), were purchased from Sigma Aldrich. Hexacarbonylmolybdenum ($\text{Mo}(\text{CO})_6$, 98%) was purchased from Thermo Scientific. Avcarb MGL 370 carbon paper was used as the loading substrate and gas diffusion layer in the fuel cell, and a commercial Pt cathode (2 mg cm^{-2}) was used as the counterelectrode, and both were directly purchased from Fuel Cell Store. Sustainion XA-9 anion exchange membrane and Sustainion XA-9 ionomer were purchased from Dioxide Materials. Palladium black (high surface area) was purchased from Fuel Cell Store.

Solutions, each containing a single metallic salt, were prepared as precursors for the robotic system. Most metallic salts were dissolved in ethanol to prepare 50 mM solutions. Salts that are insoluble or with low solubility in pure ethanol (palladium chloride, bismuth chloride and niobium chloride) were dissolved in aqueous hydrochloric acid solution (37%) to final concentrations of 20 vol% HCl(aq) and 50 mM metallic element concentration in ethanol. Both AuCl_3 and HAuCl_4 would be slowly reduced to metallic gold in the presence of ethanol, even when the solution is acidified with 20 vol% HCl(aq). Thus, we dissolve the HAuCl_4 in deionized water to prepare a 50 mM solution.

Liquid handling with Opentrons

The OT-2 liquid-handling robot was purchased from Opentrons. Two pipette channels were installed: P20 Single Channel Gen2 (20 μL tip) and P300 Single Channel Gen2 (300 μL tip). The liquid dispensing rate was optimized to be $0.378 \mu\text{L s}^{-1}$. A customized 3D-printed 36-slot sample stage was used to place carbon strips of size $1.0 \text{ cm} \times 1.3 \text{ cm}$. A 96-well plate (300 μL volume for each well) was used for mixing different solutions. The mixing protocol was executed by the 300 μL tip with a fast mixing rate of $7.56 \mu\text{L s}^{-1}$ to ensure a turbulent, homogeneous mix.

In situ electrodeposition synthesis

For AL, because the goal was to select the best recipe, only 10 μL of the precursor mixture was dropcast by Opentrons on each sample (Avcarb MGL 370 strip) of size $1.0 \text{ cm} \times 1.3 \text{ cm}$, which was pre-cut by laser to ensure high size consistency. After dropcasting, the samples were naturally dried in air for at least 2 h for the ethanol solvent to evaporate. During the later robotic electrochemical treatment, the 0.3 cm fringe was clamped by the sample holder and did not take part in reactions. Each sample was first immersed in the electrolyte for 1 min, and then activated with cyclic voltammetry for 20 scans (50 mV s^{-1}), and a linear voltammetry scan for once (10 mV s^{-1}). The potential range was from -1.0 V to 0.2 V compared with Hg/HgO . The electrolyte was a mixture of 1.0 M potassium hydroxide (KOH) and 1.0 M potassium formate (HCOOK). The treatment occurred at ambient conditions. To process samples with a higher loading density, more activation cycles would be required. After this process, nanostructures would be generated in situ on the carbon fibre for further electrochemical testing.

Nanoparticle synthesis

This method was adapted from a previous report³⁸. Specifically, CTAC (50 mg) and oleylamine (5 ml) were mixed and sonicated for about 15 min in a glass vial. After that, Pd(acac)₂, Pt(acac)₂, Cu(acac)₂, HAuCl₄, Ir(acac)₃, NbCl₅, Ce(acac)₃·xH₂O, Cr(acac)₃, glucose (60 mg) and Mo(CO)₆ (33 mg) were added into the vial with designated ratios. The total amount of the metal precursor added was controlled to be 0.125 mmol, and the exact mass of metal precursor added depends on the specific recipe. The mixture was then subjected to sonication for 1 h. The vial was then heated to 220 °C and kept at this temperature for 2 h under vigorous magnetic stirring. The black colloidal product was collected by centrifugation and washed at least twice with a mixture of ethanol and cyclohexane (1:1 in volume ratio). Then, the product was subjected to ultrasonication in 0.5 M acetic acid (in ethanol) for 2 h and then centrifuged to further remove organic impurities. The final powder product was collected after washing with ethanol solution twice.

Electrochemical testing with the 7-axis robot

The 7-axis xArm robotic arm with gripper was purchased from UFactory. A customized Cu–Au connection plate was fabricated to enable the connection of the electrode sample holder with the BioLogic Potentiostat (SP-150e). The electrolyte for the three-electrode setup test was a mixture of 150 ml 1.0 M KOH and 1.0 M HCOOK. Hg/HgO electrode (filled with 1.0 M KOH) was used as the reference, and the Pt foil electrode of size 1.0 cm × 1.0 cm was used as the counter. Pure N₂ gas (Airgas, Ultra High Purity) was continuously sparged into the electrolyte during tests. All the three-electrode tests occurred at ambient temperature. For each sample, a typical test time is around 20 min. The electrolyte would be changed, and the cell would be cleaned thoroughly every 10 samples. Alkaline solutions such as KOH may etch glass and introduce impurities into the electrolyte, but considering the relatively short amount of time for the testing, such an effect was considered minimal in this work.

Membrane electrode assembly fuel cell testing

The membrane electrode assembly flow electrolyser was purchased from the Fuel Cell Store. It has two compartments: an anolyte chamber with a titanium anode flow field and a catholyte chamber with 904-L stainless steel flow field. It also has a PID (proportional–integral–derivative) temperature controller and two customized heating pads attached to the two metallic blocks. Sustainion X37-50 was used as the anion exchange membrane for ion conduction across the inner circuit, and a commercial platinum black catalyst (2.0 mg cm⁻²) was used as the standard cathode for the oxygen reduction reactions. The commercial palladium black anode was fabricated by air-spraying 2.0 mg cm⁻² of palladium black (with 30 wt% of Sustainion XA-9 ionomer) on the carbon paper. The formate anode size was 1.0 cm × 1.0 cm. To better compare the performance of the formate oxidation, we oversized our oxygen reduction cathode to be 1.4 cm × 1.6 cm. All the current and power densities were reported against the formate anode size (1.0 cm²), which was the research topic of this work. O₂ gas (Airgas, Ultra High Purity) was constantly flowing in and out at 10 sccm through silicone tubes with a mass flow controller (Alicat) at ambient pressure. The anolyte (2.0 M KOH and 1.0 M HCOOK) was also flowing in and out at a constant flow rate of 20 ml min⁻¹ with a peristaltic pump. The membrane electrode assembly block was heated to 60 °C, and the electrolyte was heated to 70 °C on a hot plate with a thermometer immersed in the solution. All the parameters (including flow rate and temperature) are the optimized values with our setup for this work. For the polarization curve test, the linear sweep voltammetry method at a scan rate of 10 mV s⁻¹ was applied, similar to that in a previous literature report³⁹. CO₂ and carbonate are generated during the formate oxidation reaction, leading to a change in the local pH. But for the flow cell test, considering that a fresh solution was pumped in and out of the flow field constantly, this effect was insignificant.

We used both catalysts synthesized from the in situ electrochemical deposition method and the multi-element nanoparticle bulk synthesis method. The in situ electrochemical method provides a high-throughput way for sample preparation and performance optimization but suffers from low mass loading (typically below 0.2 mg cm⁻²). AL was conducted using catalysts synthesized by in situ electrochemical deposition for rapid screening. Promising or representative recipes were subsequently validated in fuel cell testing (Supplementary Fig. 3) using HEA powder prepared by conventional synthesis. For the device testing, we mixed the HEA nanoparticles with 30 wt% of Sustainion XA-9 ionomer and air-sprayed them onto the carbon paper, which improves the mass loading (controlled to be 2.0 mg cm⁻²) of our catalyst, and thus the device performance. However, such a workflow remains mostly manual for us. The key point of this paper is to leverage the robotic platform to quickly screen catalyst recipes using the three-electrode testing method, which is a more widely used electrochemical testing method in the literature.

Structural characterization

XRD was conducted on the X-ray diffractometer (Aeris Research edition) using a copper target at a voltage of 40 kV and a current of 15 mA. Deionized water was used to rinse the remaining KOH and HCOOK off the electrode surface before XRD testing. SEM was performed with a Zeiss Merlin High-resolution scanning electron microscope at the MIT Materials Research Laboratory. Characterization with transmission electron microscopy (TEM), energy-dispersive spectroscopy (EDS) and scanning transmission electron microscopy were performed with the Thermo Fisher Scientific Themis Z G3 aberration-corrected scanning transmission electron microscope with a resolution of <0.6 Å at MIT.nano. In situ XAS experiments were performed at the SPring-8 BL12B2 Taiwan Beamline, using a custom-designed electrochemical cell tailored for in situ XAS measurements. Fluorescence signals were acquired using a Lytle detector. A custom-made in situ XAS cell was used for this experiment. The electrochemical setup used a three-electrode system, consisting of a working electrode, a counterelectrode (Pt wire) and a reference electrode (Hg/HgO), immersed in an electrolyte solution composed of 1.0 M KOH and 1.0 M HCOOK.

DFT calculations

First-principles calculations were performed using spin-polarized DFT^{40,41} implemented in the Vienna ab initio simulation package⁴² with the Perdew, Burke and Ernzerhof⁴³ exchange-correlation potential within the generalized gradient approximation. The projector augmented wave pseudo-potential⁴⁴ was used to describe core electrons. For all optimization calculations, the cutoff energy was set at 450 eV and *k*-space was sampled by 3 × 3 × 1 for all models. The convergence criteria for energy and force were set at 10⁻⁴ eV and 0.02 eV Å⁻¹, respectively. The van der Waals interaction has been considered using the Grimme dispersion scheme⁴⁵.

For a certain reaction (A → B), the reaction free energy Δ*G* (including the DFT total energy, zero-point energy, vibrational enthalpy (thermal corrections) and vibrational entropy (at *T* = 298.15 K)) is defined as follows:

$$\Delta G \equiv G_B - G_A$$

Further details on the modelling approaches and model selection criteria can be found in Supplementary Note 6, whereas the energy calculations and treatment methods for formate fuel cells are described in Supplementary Note 7.

Data availability

The data that support the findings of this study are included in the main text, and the source files are available from the corresponding author upon request. Source data are provided with this paper.

Code availability

The code supporting the findings of this study is available at GitHub (<https://github.com/zhang21mit/CRESt>) and can be obtained from the corresponding author upon request.

38. Li, H. et al. Fast site-to-site electron transfer of high-entropy alloy nanocatalyst driving redox electrocatalysis. *Nat. Commun.* **11**, 5437 (2020).
39. Ni, W. et al. Synergistic interactions between PtRu catalyst and nitrogen-doped carbon support boost hydrogen oxidation. *Nat. Catal.* **6**, 773–783 (2023).
40. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
41. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
42. Kresse, G. & Hafner, J. *Ab initio* molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
43. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985 (1996).
44. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
45. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* **132**, 154104 (2010).

Acknowledgements We thank A. Quinn and F. Brushett from the Department of Chemical Engineering at MIT for discussions. The in situ XAS spectra obtained from beamline TPS 32A at

the National Synchrotron Radiation Research Center and Taiwan Beamline 12B2 at SPring-8 are appreciated. The TEM images obtained from MIT.nano are appreciated. This work was supported by the Defense Advanced Research Projects Agency under agreement no. HR00112490369 and the Defense Threat Reduction Agency (award no. HDTRA1-20-2-0002) Interaction of Ionizing Radiation with Matter University Research Alliance. H.-T.W. was supported by the National Science and Technology Council, Taiwan (grant no. 112-2112-M-Q32-017-MY2).

Author contributions Z.Z., Z.R. and J.L. conceptualized the study. Z.Z., Z.R., C.-W.H., Weibin C., Z.-W.H., C.-F.L., Y.-R.L., Y.-C.S., H.-T.W., I.I.A., P.A., Y.S.-H. and J.L. devised the methodology. Z.Z., Z.R., C.-W.H., Weibin C., Z.-W.H., C.-F.L., A.P., S.M., Y.S.-H., H.X., D.J.Z., Y.G., Weiyin C., H.S., Y.N., Y.T., Y.-R.L., Y.-C.S., S.L., H.-T.W., I.I.A., P.A., Y.S.-H. and J.L. investigated the study. Z.Z., Z.R., C.W.H., Weibin C., C.-F.L., S.M., Y.S.-H. and H.-T.W. visualized the projects. J.L. helped with funding acquisition and supervision. Z.Z., Z.R. and J.L. wrote the original draft. Z.Z., Z.R., C.W.H., Weibin C., Z.-W.H., C.-F.L., A.P., S.M., Y.S.-H., H.X., D.J.Z., Y.G., Weiyin C., H.S., Y.N., Y.T., Y.-R.L., Y.-C.S., S.L., H.-T.W., I.I.A., P.A., Y.S.-H. and J.L. reviewed and edited the paper.

Competing interests The authors declare no competing interests.

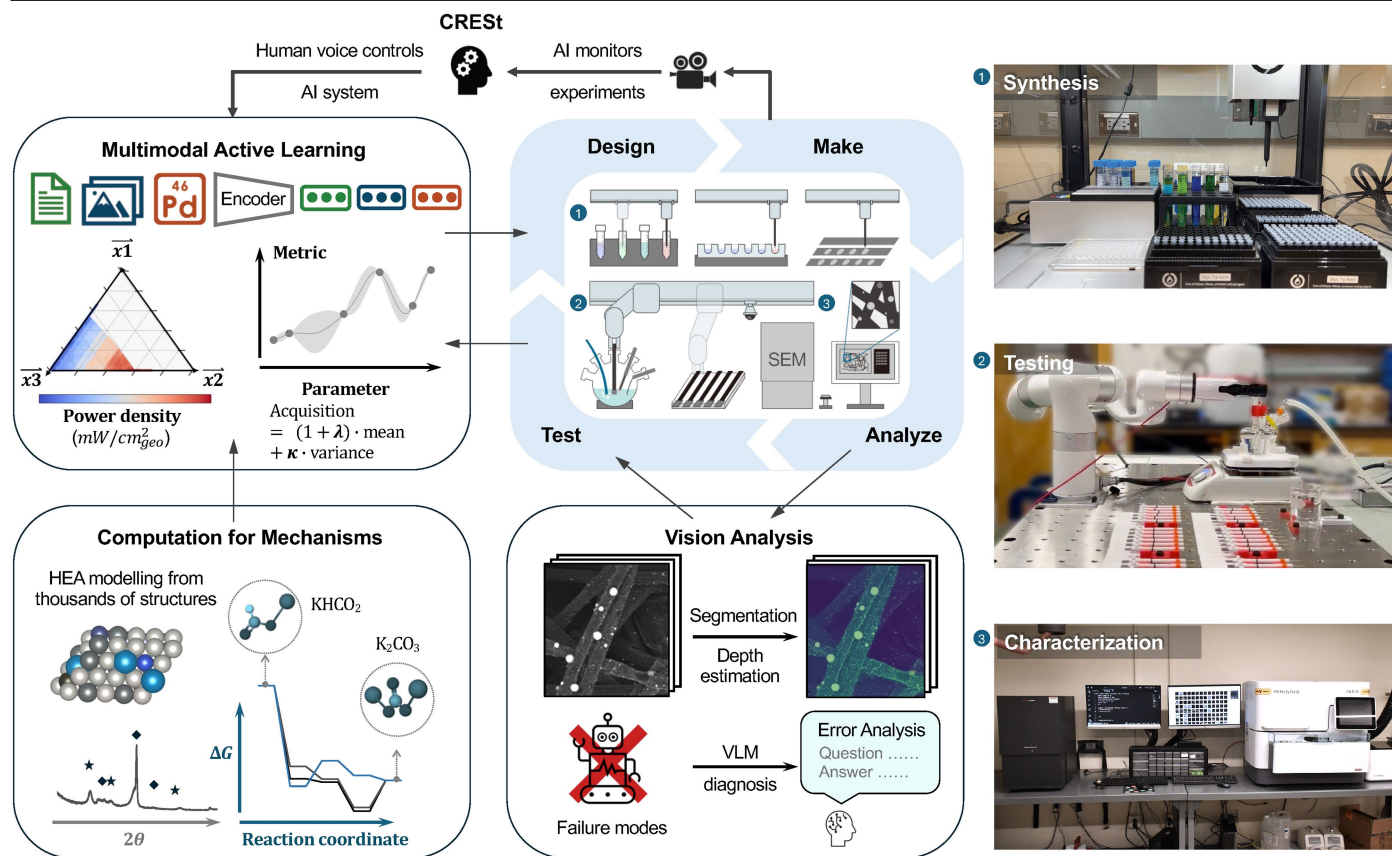
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09640-5>.

Correspondence and requests for materials should be addressed to Ju Li.

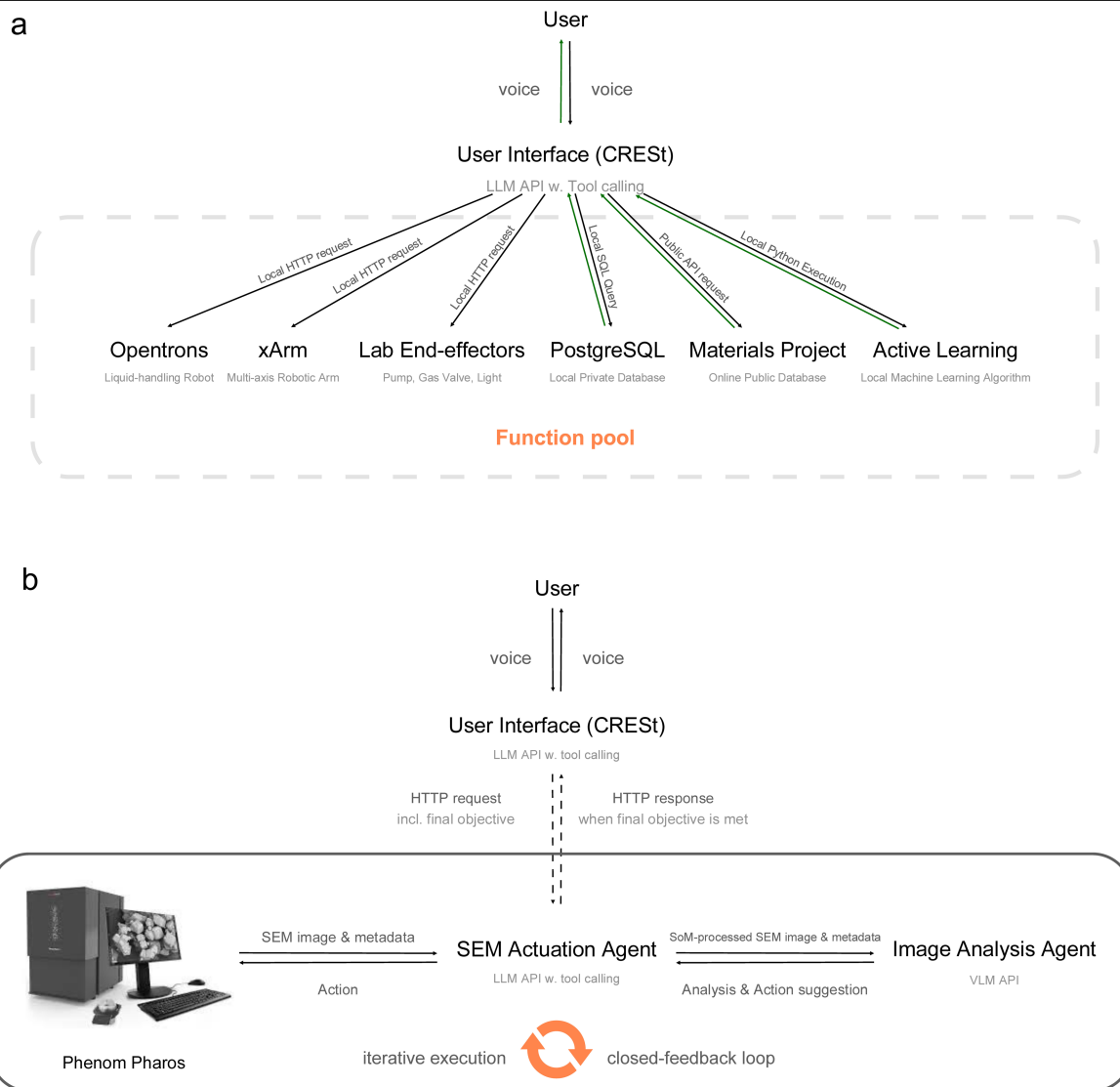
Peer review information *Nature* thanks Liang Zhang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



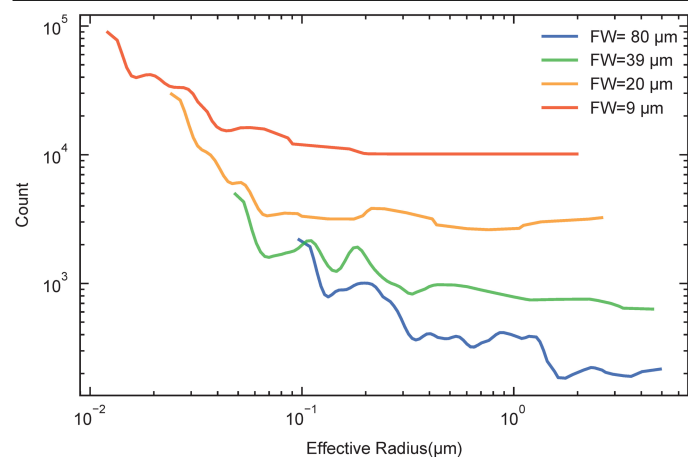
Extended Data Fig. 1 | The workflow of electrocatalyst discovery guided by CREST. A large multimodal model interfaces with researchers through text and voice, monitors experiments via visual analysis, and autonomously controls robotic instruments for electrochemical materials research. The multimodal active learning (MAL) module integrates literature and human knowledge, microstructural images, and composition and process tuning, operating in a compressed latent space to guide candidate selection. Robotic subsystems

carry out sample synthesis, testing, and characterization, while the vision module applies computer vision techniques such as segmentation and depth estimation to analyze microstructures. In parallel, a vision language model diagnoses experimental error modes to improve reproducibility. Experimental results are iteratively incorporated into the active learning loop, and optimized materials are validated through mechanistic studies and device-level testing.

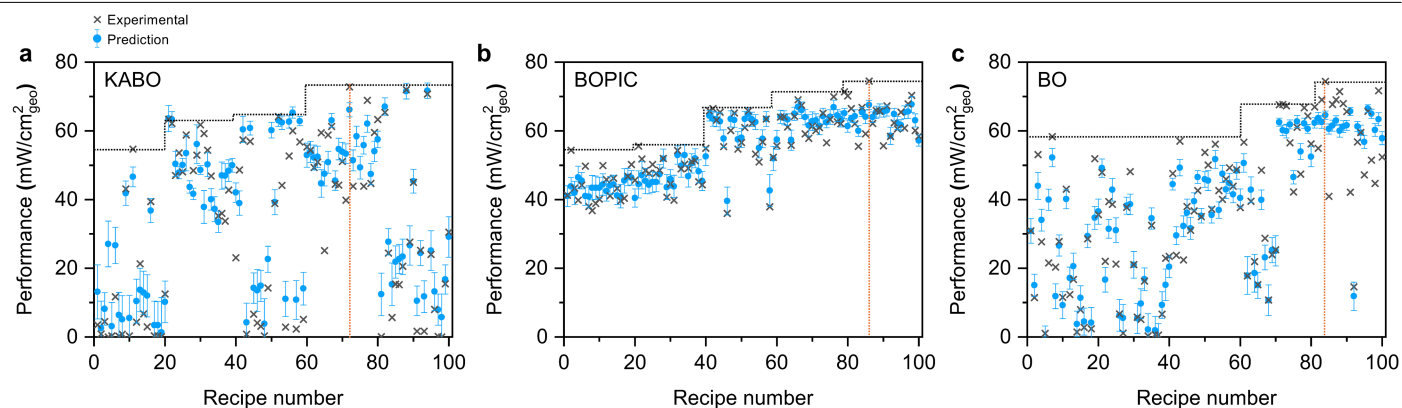


Extended Data Fig. 2 | The CREST coding architecture. (a) Architecture of the CREST main experimental platform. CREST acts as a laboratory orchestrator, dynamically invoking routines from a shared function library that includes instrument-control systems and Python workflows to execute real-world experiments. (b) Architecture of the autonomous scanning electron microscope (SEM) module. A three-agent loop coordinates (i) a dialogue/orchestration agent at the CREST layer that interfaces with users and sets goals, (ii) an SEM control

agent that performs navigation, focusing and zooming, and (iii) a vision agent that analyzes images and recommends the next action. The loop iterates until the imaging objective is met, after which the SEM agent returns the final micrograph and an execution report to CREST. A separate computer vision analysis module is also utilized to analyze SEM images for statistical features (Supplementary Note 4).

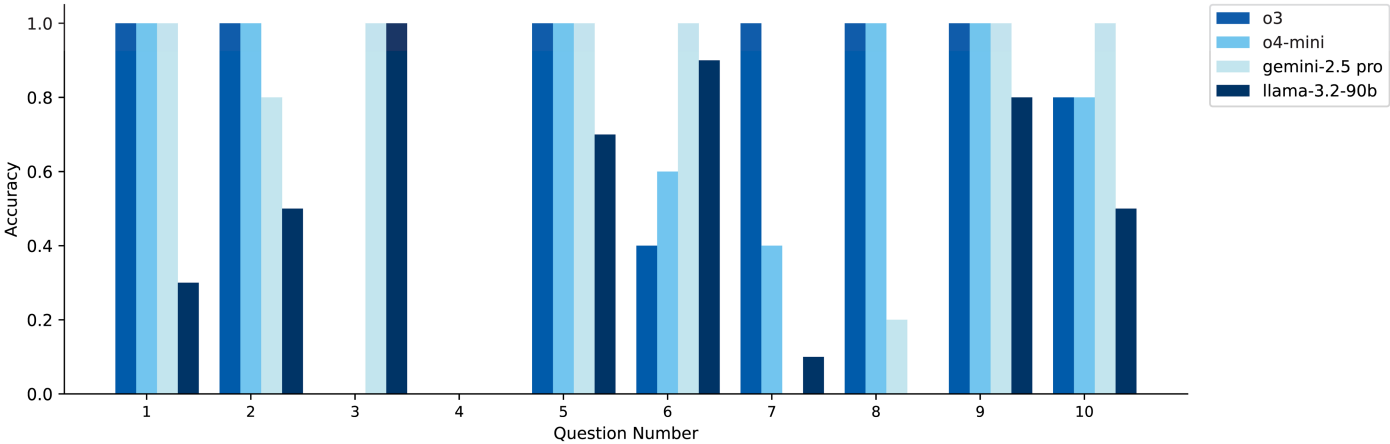


Extended Data Fig. 3 | Example of the particle number distribution function in the logarithm scale. The feature statistics were obtained from SEM images captured at varying field widths across our dataset. The effective radius is calculated by converting the measured particle area into an equivalent spherical radius. Our analysis shows that obtaining accurate feature statistics, especially those related to particle size, requires selecting a field of view that is appropriately scaled to the size of the particles being analyzed.



Extended Data Fig. 4 | Comparison plot of the active learning campaign for the three algorithms. (a) Knowledge-Assisted Bayesian Optimization (KABO). (b) Bayesian Optimization with Policy Improvement Constraints (BOPIC). (c) Standard Bayesian Optimization using the Upper Confidence Bound (UCB) acquisition function. Both KABO and BOPIC achieved continuous improvements batch after batch, whereas standard BO required ~70 samples before making progress—a critical limitation when each experiment is costly. KABO, likely due

to its integration of domain knowledge, showed the strongest alignment between predictions and experimental results, particularly in high-performance regions, and also identified the global optimum first (indicated by the orange vertical line). BOPIC proved especially effective in sampling high-performance candidates relative to the other two methods. Ultimately, all three algorithms converged to a similar high-performance value, consistent with the fixed ground truth in the chemistry space.



Extended Data Fig. 5 | Comparison of performance for the state-of-the-art vision language model (VLM) on a real-world materials science experimental question set. Considering the stochasticity of the vision language models, each question was tested for five times for each model to compute the average accuracy. Examples could be found in the Appendix of the Supplementary file.

Supplementary information

A multimodal robotic platform for multi-element electrocatalyst discovery

In the format provided by the
authors and unedited

Supplementary Materials for

A multimodal robotic platform guides the discovery of a champion high-entropy electrocatalyst

Zhen Zhang^{1†}, Zhichu Ren^{1†}, Chia-Wei Hsu^{1†}, Weibin Chen^{2†}, Zhang-Wei Hong³, Chi-Feng Lee⁴, Aubrey Penn⁵, Hongbin Xu¹, Daniel J. Zheng¹, Shuhan Miao⁶, Yimeng Huang², Yifan Gao², Weiyin Chen², Hugh Smith¹, Yaoshen Niu¹, Yunsheng Tian³, Ying-Rui Lu⁷, Yu-Cheng Shao⁷, Sipei Li², Hsiao-Tsu Wang⁴, Iwnetim I. Abate¹, Pulkit Agrawal³, Yang Shao-Horn^{1,8} and Ju Li^{1,2*}

†These authors contributed equally to this work.

*Corresponding author: liju@mit.edu

The PDF file includes:

Supplementary Note S1 to S9
Supplementary Figure S1 to S31
Appendix
References

Other Supplementary Materials for this manuscript include the following:

Supplementary Movie S1, S2
Supplementary Data S1, S2

Supplementary Note 1

Methods for standard Bayesian optimization (BO) with the upper confidence bound (UCB) acquisition function in the electrochemical settings

The active learning architecture we used in this work was a standard Gaussian Process plus Bayesian Optimization approach, with Matern 5/2 as the Gaussian kernel and Upper Confidence Bound (UCB) as the acquisition function¹. Ax & BoTorch, an active learning Python package developed by Meta, was deployed as the backend engine². A sequential batch Bayesian Optimization method was utilized to select candidates in each iteration.

$$a_{UCB}(x; \kappa) = \mu(x) + \kappa\sigma(x)$$

Where a_{UCB} is the value from the acquisition function, $\mu(x)$ is the function mean and $\sigma(x)$ is the function standard deviation. The exploration-exploitation coefficient κ (with a default value of 0.1 in the BoTorch package) was manually adjusted along the active learning campaign. After the 14th batch (arriving at the red shade in Fig. 1e in the main text), the algorithm may arrive at the local maxima, and thus the κ was intentionally adjusted to higher values to allow more exploration. In specific, we adjusted the beta to be 100 on batch 14, highly biasing the algorithm towards exploration, and as a result, most suggested recipes only consist of two elements and showed enormous uncertainty. We then reduced beta value to 10 in batches 15 to 17, biasing less towards exploration. For the last two batches, we further reduced the κ to 1.

One other thing to note here is that the equality constraint in this work (all compositions adding up to 1) is challenging to achieve technically. This is because rejection sampling is used in Ax & BoTorch by default, and equality constraint will render a feasible set with zero volume. As a workaround, we excluded one component (herein chromium, Cr) and relaxed the constraint to an inequality stating that the summation of atomic percentages of all non-chromium components is less than 1. The composition of Cr will then be individually calculated after each recipe recommendation is generated (simply subtracting all the other components from unity).

All the active learning procedures were based on electrochemical tests in a three-electrode setup at ambient temperature in this work. After numerous iterations, the selected materials were incorporated into the fuel cell device for the final tests.

The objective function to be optimized by the active learning algorithm is as follows:

$$P_{\max} = \arg \max (iE_{\text{cell}}(i, x_1, x_2, \dots, x_{i-1}))$$

Where P_{\max} (mW/cm²) is the maximized geometric power density, which is a function of the geometric current density (i , mA/cm²) and the full cell potential (E_{cell} , V). Further, E_{cell} is a function of current density and the catalyst composition (represented by the mass fractions of component

x_i). Here, i represents the geometric current density at the electrode where formate is oxidized (note, this distinction is necessary because our counter electrode is oversized).

E_{cell} is both a function of the current density and the catalyst composition, but not a function of the ORR reaction. Under the assumptions that the ohmic resistance is captured by a single area-specific resistance term (e.g., that of the membrane) and that the counter/working electrode processes are not influenced by the reagent/products crossing over the membrane, E_{cell} can be decomposed into the following:

$$E_{cell} = U - iR_{ohmic} - \eta_{ORR} - \eta_{FOR}(i, x_1, x_2, \dots, x_{i-1})$$

Here, U (V) is the full cell open circuit potential, R_{ohmic} ($\Omega \text{ cm}^2$) is the area-specific resistance of the membrane, η_{ORR} (V) is the overpotential of the oxygen reduction reaction (ORR), and η_{FOR} is the overpotential of the formate oxidation reaction (which here depends on i and catalyst recipe/synthesis), all at specified conditions (1 M KOH and 1 M KHCO_2 on the working electrode, 1 atm O_2 on the counter, 60 °C). The open circuit potential can be expressed with the formal reduction potentials of:

$$U = (E_{ORR}^{0'} - E_{FOR}^{0'})$$

Here, $E_{ORR}^{0'}$ is the formal oxygen reduction potential at the counter electrode vs RHE. $E_{FOR}^{0'}$ is the formal reduction potential of the working electrode (presumably bicarbonate reduction to formate) vs RHE. Substituting into the expression for E_{cell} and rearranging to group ORR and FOR terms yields:

$$E_{cell} = (E_{ORR}^{0'} - \eta_{ORR}) - (E_{FOR}^{0'} + \eta_{FOR}(i, x_1, x_2, \dots, x_{i-1})) - iR_{ohmic}$$

The two terms in parentheses are grouped to yield measurable or estimable potential values for the ORR half-cell (E_{ORR}) and for the formate half-cell (E_{FOR}), vs the same reference electrode. These values reflect the potential measurements at a given current density, including kinetic/mass transport overpotential effects, but neglecting effects such as distributed current/potential distributions.

$$E_{cell} = E_{ORR} - E_{FOR}(i, x_1, x_2, \dots, x_{i-1}) - iR_{ohmic}$$

And

$$P_{cell} = iE_{ORR} - iE_{FOR}(i, x_1, x_2, \dots, x_{i-1}) - i^2R_{ohmic}$$

φ_{ORR} is assumed to be the standard reduction potential of ORR (1.23 V vs. RHE), given the use of 1 atm O_2 feed. In fuel cell tests, we observed that most of polarization came from the formate oxidation half-cell and that the ORR and membrane overpotential was relatively constant (ranging between 0.2 ~ 0.3 V) over the current density range of 0.2 ~ 0.4 A. For simplicity, thus, we approximated $\eta_{\text{ORR}} + \eta_{\text{membrane}}$ to be constant at 0.23 V in our active learning algorithm such that $E_{\text{ORR}} = 1$ V at all studied conditions³. E_{FOR} is the measured potential of the working electrode. The potential of the working electrode is originally measured against a Hg/HgO reference electrode ($E_{\text{FOR, Hg/HgO}}$, V), which is converted to the RHE scale by $E_{\text{FOR}} = E_{\text{FOR, Hg/HgO}} - 0.924$ ($E_{\text{Hg/HgO}} = 0.924$ V vs. RHE at pH 14).

To further simplify the problem, we neglect the ohmic loss term (setting $R_{\text{ohmic}} = 0$):

$$P'_{\text{max}} = \arg \max(iE_{\text{ORR}} - iE_{\text{FOR}})$$

This objective function was used for our studies. While it states that the system is kinetically/mass-transport controlled, its use can also likely be justified even when R_{ohmic} is not 0, depending on the $E_{\text{FOR}} - i$ relationship. Further, R_{ohmic} can often be easily estimated in a full cell system and subtracted from the measured power. Objective functions should adequately capture features of the full-system for the desired application, to properly evaluate the catalyst role in that system.

The goal of the model is to contextualize the catalyst performance in a hypothetical device utilizing the oxygen reduction reaction (ORR) at the counter without the time- and equipment-intensive deploying of fuel cell experiments for every catalyst formulation. The simple electrochemical model shown above seems fit for this purpose: $E_{\text{FOR}} - i$ data automatically extracted with Python scripts from LSV experiments are used to calculate the maximum power and logged into a database. As we showed in this work, this approximate model gives the same relative ranking of the catalyst performance in the three-electrode setup as in a fuel cell device under the conditions studied.

Supplementary Note 2

Methods for the knowledge-assisted active learning

In the process of Knowledge-Assisted Bayesian Optimization, we replace the common orthogonal searching space, with the latent space of LLM embeddings. By providing prompts for all elements to be searched, the Text Embedding Model (such as text-embedding-ada-002 by OpenAI) or various other open-source model returns the representation vectors for these elements. These vectors illustrate the correlations between the elements. Utilizing this information, we perform a rational dimensionality reduction using PCA, ensuring that the explained variance ratio remains above 0.8. This method produces a 'new dimension' that reflects the properties of the elements. Following this, the BO algorithm searches for the optimal solution within this space, and the optimization algorithm (SLSQP) remaps the solutions back to the original atomic ratio search space. While some parts of the search space may be restricted, the efficient redistribution of samples in the latent space allows for the effective discovery of high-value samples.

Algorithm 1 Bayesian Optimization with Text embeddings assist

Require: Metal Element List \mathbf{E} , Bayesian Optimization Dimension d

- 1: Initialization the searched recipe atomic ratio \mathbf{w}
 - 2: Represent Vectors $\mathbf{V} \leftarrow \text{Embedding}(\text{Description} + \mathbf{E})$
 - 3: Fit PCA(\mathbf{V} , dimension= d)
 - 4: Latent subspace $\mathbf{S} \leftarrow \text{PCA}(\mathbf{wV})$
 - 5: **for** iteration t **do**
 - 6: The next batch suggestions $\mathbf{S}' \leftarrow \text{knowledge gradient}(\mathbf{S})$
 - 7: $\mathbf{V}' \leftarrow \text{PCA}^{-1}(\mathbf{S}')$
 - 8: Optimized wights $\mathbf{w}' \leftarrow \text{Minimization}(|\mathbf{V}' - \mathbf{w}'\mathbf{V}|)$
 - 9: Do the experiment with atomic ratio \mathbf{w}'
 - 10: Update the retrieved data $\mathbf{w} \leftarrow \text{Stack}(\mathbf{w}, \mathbf{w}')$
 - 11: **end for**
-

We further evaluated several open-source text embedding models both with [Qwen3-Embedding-0.6B, 4B and 8B⁴, nomic-embed-text-v1.5⁵] and without[all-MiniLM-L12-v2⁶, bge-large-en-v1.5⁷] MRL. For each model, we tested the first 32, 64, 128 and 256 dimensions and found that in nearly every case just five dimensions (and in a few cases four) were enough to exceed a cumulative explained variance ratio of 0.9. When these reduced representations were used in active learning, they achieved nearly identical performance (Fig. S5). These results demonstrate that our method leverages this information to avoid getting trapped in local optima and achieves strong performance across a wide range of experimental conditions. These experiments were conducted by adjusting our existing simulations, repeating each configuration ten times to calculate average performance, and using UCB ($\kappa = 2.5$) as the acquisition function. We found that selecting only a few representation dimensions led to large performance fluctuations, whereas increasing the number of selected dimensions produced progressively more consistent results across all models. In addition, we also compared our model performance with the recent transformer architecture (Fig. S31).

Supplementary Note 3

Methods for the active learning with dynamic exploration and exploitation

In active learning, balancing exploration and exploitation in Bayesian optimization (BO) is a longstanding challenge. Exploitation involves selecting points with the highest expected value in the surrogate model, while exploration focuses on points with high uncertainty. Striking the right balance is essential: over-exploitation can overlook valuable data points, while over-exploration may yield only low-value points. Traditionally, this balance relies on tuning a weight parameter, κ , in the following formula:

$$\max_x \mu(x) + \kappa\sigma(x),$$

where $\mu(x)$ is the surrogate mean value of data point x , and $\sigma(x)$ represents its uncertainty.

However, manual tuning can be infeasible if evaluating data points is costly. We addressed this by adapting the Extrinsic-Intrinsic Policy Optimization (EIPO) approach from reinforcement learning (RL), which balances objectives through constrained optimization. The main idea is that, given an exploratory and an exploitative strategy, the exploratory one should eventually outperform the exploitative one in terms of the true black box value f —demonstrating that exploration is beneficial. We formalize this using the following constraint when selecting a data point:

$$\max_x \mu(x) + \kappa\sigma(x) \quad \text{subject to} \quad f(x) \geq f(x'),$$

where x' is chosen by maximizing expected improvement (EI).

Following the original EIPO method, we solve this constrained problem using Lagrangian duality, resulting in a min-max optimization:

$$\min_{\lambda \geq 0} \max_x \underbrace{\mu(x) + \kappa\sigma(x) + \lambda(f(x) - f(x'))}_{\mathcal{L}(x, \lambda)},$$

where λ is the Lagrange multiplier and \mathcal{L} denotes the Lagrangian.

How to choose x ? The acquisition function for choosing datapoint x is the follows:

$$\max_x \mu(x) + \kappa\sigma(x) + \lambda f(x).$$

As we see that f appears in this objective, which makes optimization of x hard since querying the black-box function f is costly. Thus we approximate f by μ , turning the acquisition function to the following:

$$\max_x (1 + \lambda)\mu(x) + \kappa\sigma(x).$$

Note that the Lagrangian multiplier λ turns to be the weight of the value of the data point in optimization. The higher the λ is, the more important the value of the next data point should be. Yet, we haven't explained how the value of λ is tuned, which we will soon illustrate below.

How to update λ ? We update λ based on gradient descent, where the gradient of λ is given as:

$$\frac{\partial \mathcal{L}(x, \lambda)}{\partial \lambda} = f(x) - f(x').$$

Running gradient descent with the above gradient updates λ as follows:

$$\lambda \leftarrow \lambda - \frac{\partial \mathcal{L}(x, \lambda)}{\partial \lambda}.$$

This update yields an adaptive scheduling of the weight λ of data point's value in optimization. Let's consider the following cases:

- When the exploratory strategy x is *better* than the exploitative strategy, $f(x) - f(x') > 0$ and thus λ decreases. It means the exploratory strategy is better than the exploitative strategy.
- When the exploratory strategy x is *worse* than the exploitative strategy, $f(x) - f(x') < 0$ and thus λ increases. It means the exploitative strategy is better, and so the exploratory strategy should pay more attention on optimizing the expected value $\mu(x)$.

The implementation details are provided in the following pseudocode.

Algorithm 1 Active learning via EIPO

```

1: procedure EIPO(Batch size  $B$ )
2:   Initialize the surrogate model  $\mathcal{N}(\cdot|\mu, \sigma)$ 
3:   Initialize the Lagrangian multiplier  $\lambda$ 
4:   for  $t = 1 \dots T$  do
5:      $\{x_1 \dots x_{B/2}\} \leftarrow$  Sample  $B/2$  points  $x$  by the acquisition function:  $(1 + \lambda)\mu(x) + \kappa\sigma(x)$ 
6:      $\{x_{B/2} \dots x_B\} \leftarrow$  Sample  $B/2$  points  $x$  by the acquisition function:  $EI(x)$ 
7:      $\{y_1 \dots y_B\} \leftarrow$  Evaluate the datapoints  $\{x_1 \dots x_B\}$ 
8:     Update the surrogate model  $\mathcal{N}$  using  $\{(x_i, y_i)\}_{i=1}^B$ 
9:     Update the Lagrangian multiplier  $\lambda$ :  $\lambda \leftarrow \lambda - \frac{2}{B} \sum_{i=1}^{B/2} (y_i - y_{i+B/2})$ 
10:   end for
11: end procedure

```

Supplementary Note 4

Method for analysis of nanoparticle features by computer vision

To achieve this, we employed a depth prediction model, Depth Anything, to segment each image into foreground and background regions. Particle analysis was then restricted to the foreground, while the background was treated as void space.

The particle counting procedure involved the following steps:

1. Image sharpening to enhance structural contrast;
2. Depth estimation using DepthAnythingV2⁸ to distinguish foreground elements;
3. Adaptive thresholding to isolate potential particle regions;
4. Morphological operations to refine particle shapes;
5. Selection of particles based on absolute brightness above a defined threshold, with inclusion limited to those located exclusively within the foreground.

This approach significantly reduced sampling errors caused by substrate inhomogeneities.

In parallel, we addressed potential sampling bias arising from limited SEM image selection—a known issue when only a few fields of view are analyzed. To counter this, we incorporated SEM images spanning a wide range of field widths, thereby minimizing the risk of biased observations. As shown in Fig. S1, Fig. 2a-c, Extended Data Fig. 3, we performed particle size distribution fitting by integrating images covering an 8-fold variation in resolution. In addition to size distribution, we also extracted particle coverage and particle count per unit area as supplementary global descriptors, allowing for a more comprehensive characterization of the overall microstructure.

Supplementary Note 5

Calculation of the size of the high-dimensional materials design space

In our active learning workflow, the constraint with input variables is to satisfy an inequality. For example, if there are only 3 elements, each with a composition of a_1 , a_2 and a_3 , then the inequality is $a_1 + a_2 \leq 1$. The third composition, a_3 , could be easily derived from the rest of the compositions by subtracting from unity. Thus, the size of the space is essentially computing the area of a triangle of a square on a 2d plane, for the 3d materials. Given that the interval/precision (in terms of atomic percent) is 0.001 for our active learning workflow. The number of possible 3d materials candidates is $\frac{(10^3)^2}{2} = 5 \times 10^5$. Similarly, the number of possible n d materials candidates is $\frac{(10^3)^{n-1}}{(n-1)!}$.

Supplementary Note 6

Method of data processing for the *in situ* X-ray absorption experiments

Data reduction, data analysis, and EXAFS fitting were performed and analyzed with the Athena and Artemis programs of the Demeter data analysis packages⁹ that utilizes the FEFF6 program¹⁰ to fit the EXAFS data. The energy calibration of the sample was conducted through standard Pd foil and Pt foil, which as a reference were simultaneously measured. A linear function was subtracted from the pre-edge region, then the edge jump was normalized using Athena software. The $\chi(k)$ data were isolated by subtracting a smooth, third-order polynomial approximating the absorption background of an isolated atom. The k^3 -weighted $\chi(k)$ data were Fourier transformed after applying a HanFeng window function ($\Delta k = 1.0$). For EXAFS modeling, the global amplitude EXAFS (CN, R, σ^2 and ΔE_0) were obtained by nonlinear fitting, with least-squares refinement, of the EXAFS equation to the Fourier-transformed data in R-space, using Artemis software, EXAFS of the Pd foil and Pt foil was fitted respectively and the obtained amplitude reduction factor S_0^2 value (0.825 and 0.803) was set in the EXAFS analysis to determine the coordination numbers (CNs) in sample.

Supplementary Note 7

Method for construction of HEA models

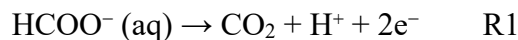
The models for Pd-based high entropy alloys were constructed based on three steps:

- (i) We initiated model construction from a face-centered cubic Pd structure and built a three-layer slab of the Pd(111) surface. This slab was expanded into a (4×5) supercell, comprising 60 atoms.
- (ii) Based on the target composition ($\text{Pd}_{0.381}\text{Pt}_{0.080}\text{Cu}_{0.009}\text{Au}_{0.004}\text{Ir}_{0.02}\text{Ce}_{0.086}\text{Nb}_{0.338}\text{Cr}_{0.082}$), we generated 100 random atomic configurations using a Python-based random shuffling algorithm to mimic the configurational entropy of HEAs, as illustrated in Fig. S16. Each configuration was fully relaxed, and the one with the lowest total energy was selected as the representative structure, hereafter referred to as HEA-8D. This energy-guided selection method ensures that the model captures the statistically reasonable local atomic arrangement.
- (iii) The energetically favorable HEA-8D configuration was subsequently used as the basis for all electronic structure and catalytic property calculations.

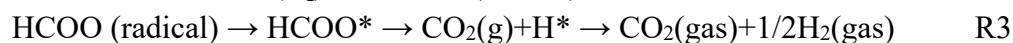
Supplementary Note 8

Energy reference selection and free energy treatment for formate oxidation

To accurately compute the free energy changes associated with formate oxidation, we adopted a reference scheme based on the neutral radical HCOO^\cdot , calibrated against the energy difference between HCOO^- and HCOO^\cdot . This approach avoids potential complications arising from the treatment of solvation effects, single electrons, and periodic charged systems in DFT. The overall target reaction is:



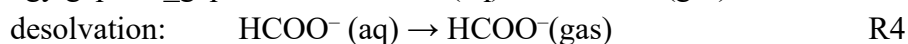
Instead of directly using the solvated formate anion as the reference, we decompose the process into the following steps:



The radical $\text{HCOO}^\cdot(\text{gas})$ can be readily treated using spin-polarized DFT, and the subsequent transformation (R3) proceeds via well-defined intermediates that are also accessible to direct DFT calculations. Under this reference, the free energy change from the reference state ($\text{HCOO}^\cdot + \text{catalyst surface}$) to the final state ($\text{CO}_2 + 1/2\text{H}_2 + \text{catalyst}$) is calculated to be approximately $\Delta G \approx -2.23 \text{ eV}$, consistent with values reported in the literature (ACS Energy Letters, 2023, 8, 3945–3954).

However, referencing to HCOO^\cdot introduces an offset relative to the actual reactant, $\text{HCOO}^-(\text{aq})$.

This energy gap ΔG_{gap} between $\text{HCOO}^-(\text{aq})$ and $\text{HCOO}^\cdot(\text{gas})$ arises from:



$\Delta G(\text{R4})$ represents the solvation free energy, computed via DFT (PBE functional, VASP). The primary challenge in this step is handling the charged species in a periodic cell. $\Delta G(\text{R5})$, the ionization step, can be approximated using the electron affinity (EA) of HCOO^- , derived from the eigenvalue of the LUMO. This gives $\Delta G(\text{R5}) \approx -1.80 \text{ eV}$. The total gap is then:

$$\Delta G_{\text{gap}} = \Delta G(\text{R4}) + \Delta G(\text{R5}) \approx 1.80 \text{ eV}$$

This value is catalyst-independent and can be used as a constant offset when comparing energies across different surfaces.

Supplementary Note 9

Understanding catalytic activity of HEA surfaces through site-specific energetics and DOS analysis

We have also considered all possible adsorption sites on the HEA surface model. Among them, eight sites were identified as catalytically active and capable of stably completing the reaction, as illustrated in Fig. S17. These include three Pd sites (Pd-1, Pd-2, and Pd-3) located on the surface but differing in coordination number and local atomic environment; two Pt sites (Pt-1 and Pt-2) with distinct geometries; as well as individual Ce (HEA-Ce-1), Cu (HEA-Cu-1), and Nb (HEA-Nb-1) sites. All adsorption sites are depicted in Fig. S17, and their corresponding reaction pathways and free energy profiles are detailed in Fig. S18-25. By summarizing the potential-determining steps (PDS) along both the direct and indirect reaction pathways, we re-evaluated the energy barriers associated with each active site on the HEA surface. The PDS energy barriers for the direct pathway are presented in Fig. S26, while those for the indirect pathway are shown in Fig. S27. Our calculations reveal that the Pd site in the HEA exhibits significantly lower energy barriers compared to pure Pd. Notably, for the indirect pathway, the Pd site of HEA shows a substantially reduced PDS of -0.005 eV, in stark contrast to 0.706 eV for pure Pd. This highlights the critical role of local structural modulation in enhancing the catalytic performance of HEAs.

To provide further mechanistic insight, we performed projected density of states (DOS) analysis for the three Pd sites (Fig. S28). The calculated d-band centers for Pd-1, Pd-2, and Pd-3 are all lower than that of pure Pd, indicating that the HEA environment weakens the strong adsorption of H and CO on Pd sites, thereby mitigating potential H or CO poisoning.

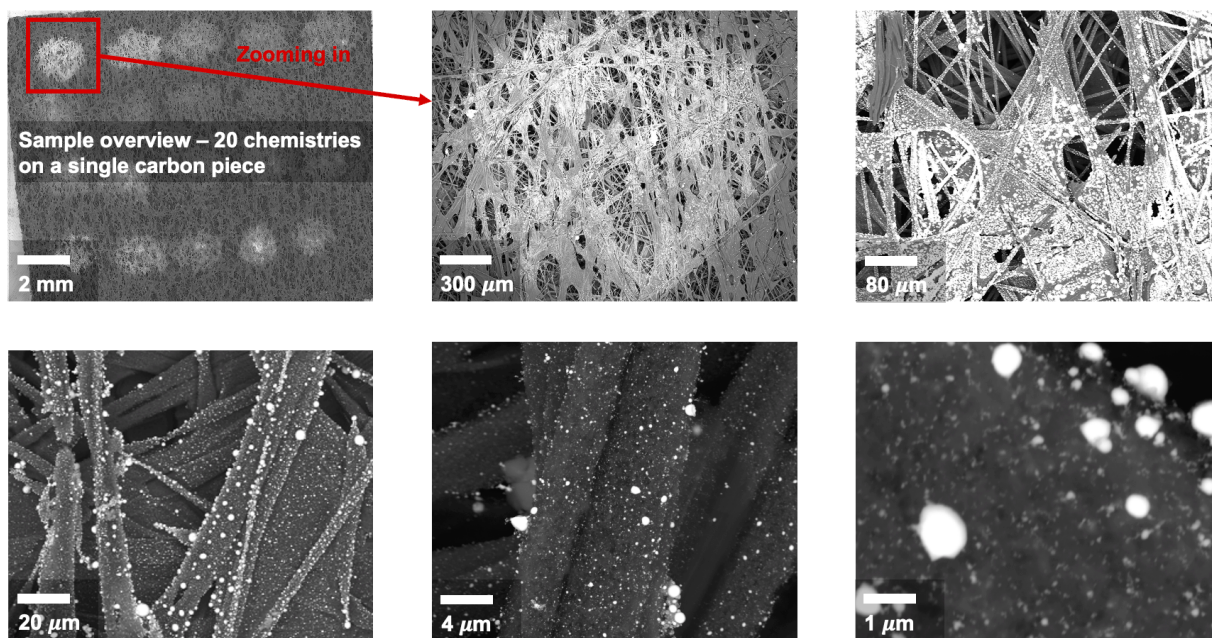


Fig. S1. An example of stitched Scanning Electron Microscopy (SEM) images for recipes with different resolutions. In this case, 20 different chemistries were integrated onto a single carbon piece (of size ~ 1.5 cm by 2.5 cm) to increase the imaging throughput (note that this current size of 20 recipes per batch of imaging is limited by the size of the sample stage of our desktop SEM, but may be fully expanded in other SEMs with larger sample stages). A script-driven SEM was utilized to automatically zoom in/out, focus, stigmatize, and take an image. For each recipe (which appears as a single ‘droplet’ in the top left figure), we elected to take images at multiple resolutions to enable the characterization of particles of various dimensions. Big, overall images could be stitched together from small, local images for computer vision analysis.

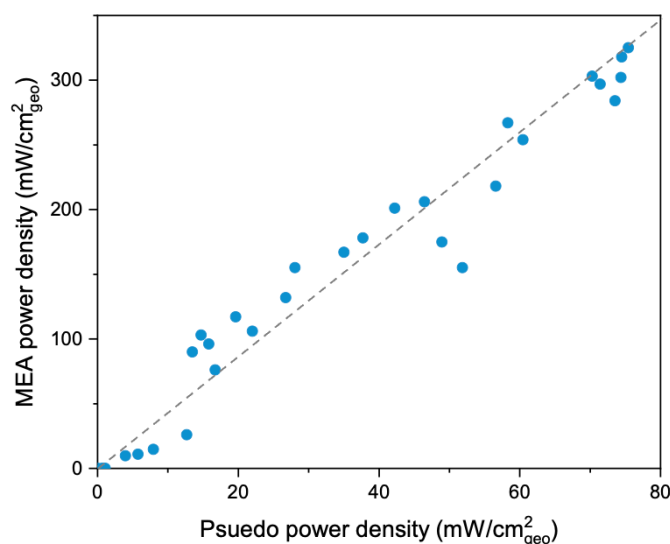
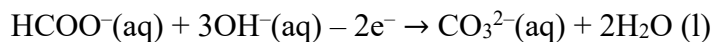


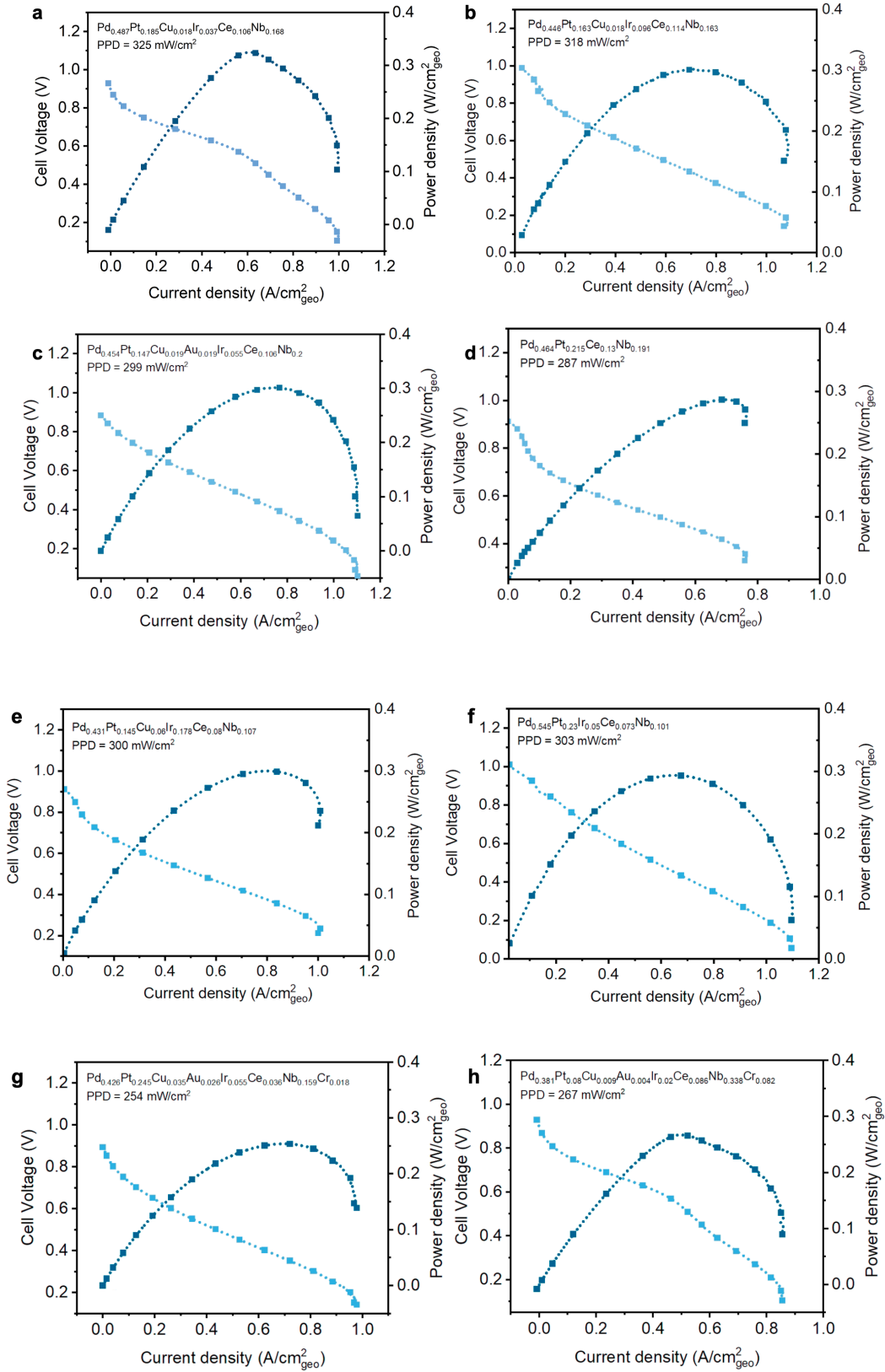
Fig. S2. Correlation between the pseudo power density derived from the three-electrode testing and the power density tested in the direct formate fuel cell. We selected 30 different chemistries for the membrane electrode assembly (MEA) tests. The Pearson coefficient r^2 for these two tests is 0.92, which shows a strong correlation between the results from the three-electrode tests and the MEA testing. The specific recipe and polarization curve for each MEA tested were shown in Fig. S3.

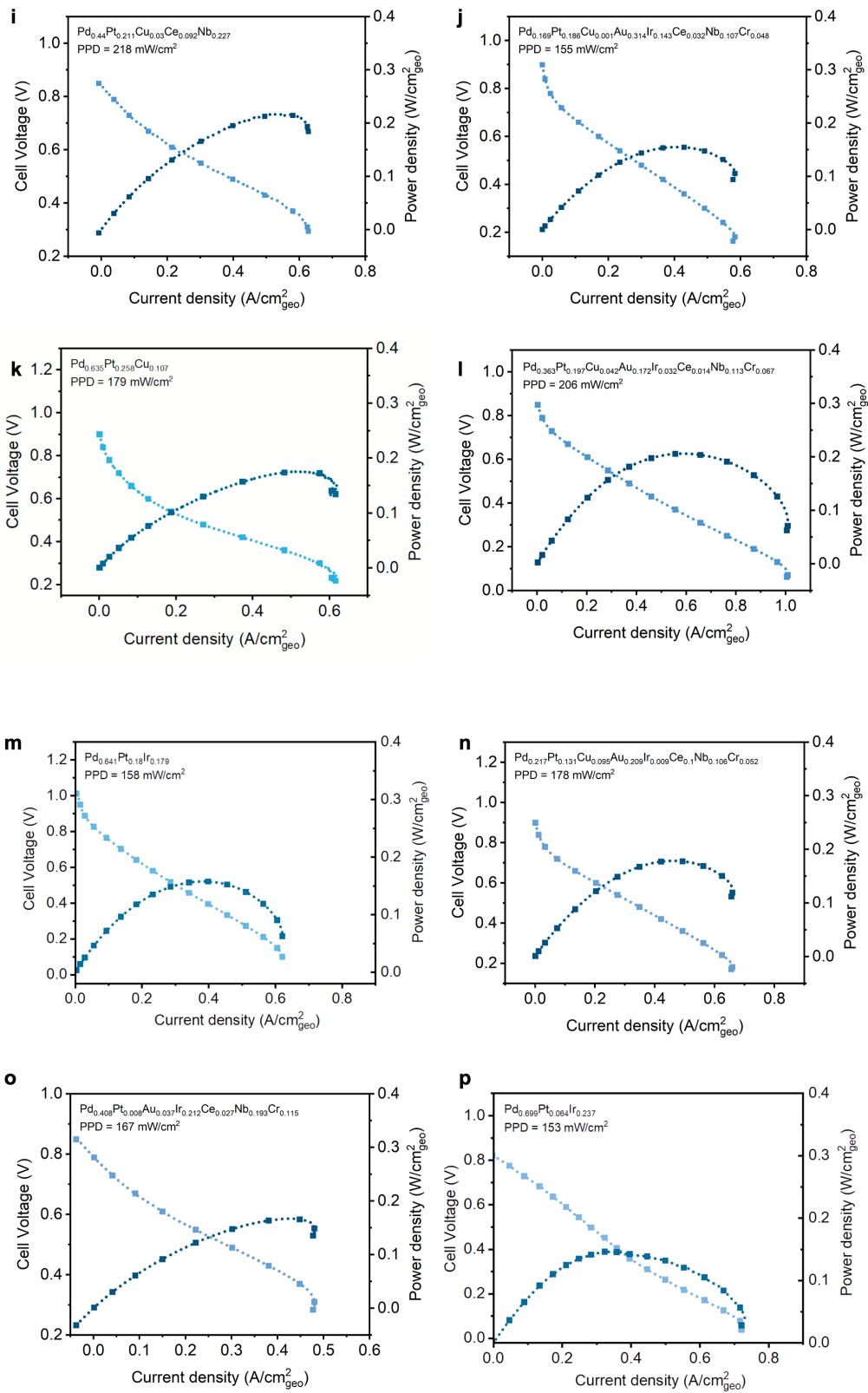
For these testing, we only changed the catalyst chemistry of the formate oxidation electrode, but kept all other variables and parameters fixed, including the catalyst loading density, substrate, counter electrode, membrane, electrolyte, gas flow rate, temperature, etc. In addition, another point to notice is that we were investigating an electrochemical reaction that does not involve a vigorous gas phase. In specific, the formate oxidation half reaction that occurs in 1.0 M KOH solution could be written as:

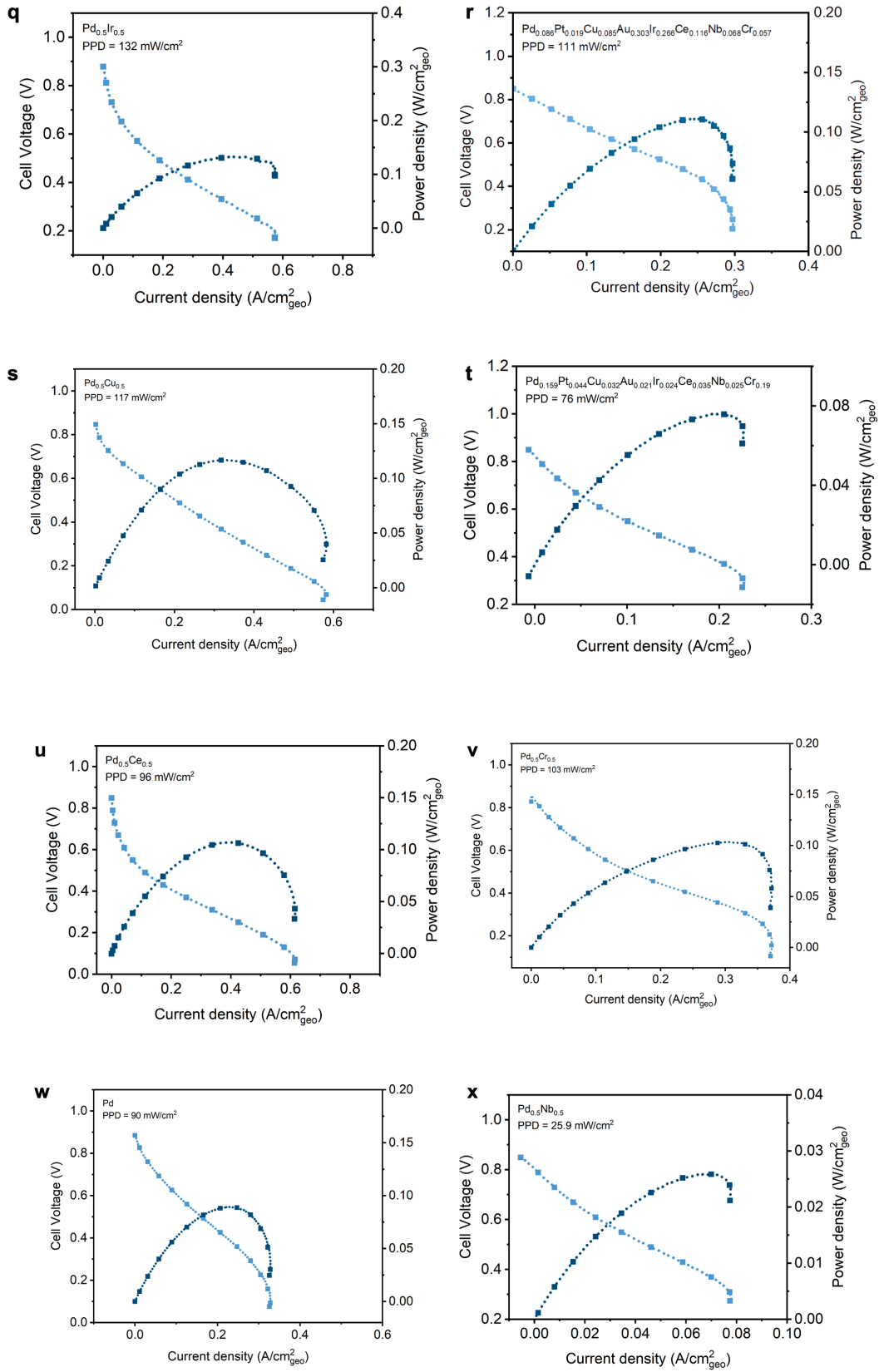


Thus, we believe that this reaction could be considered as a liquid-to-liquid conversion process. Although concerns may be that small amount of CO_2 gas still evolves, the concentrated 1.0 M KOH aqueous solution that we use is considered as an efficient CO_2 capture solvent. In fact, we did not observe obvious bubble effect at the formate anode when reacting in 1.0 M KOH. We believe that the ‘liquid-to-liquid conversion’ nature of the reaction may also help partly explain the good consistency between the three-electrode testing and MEA testing in this case. To some extent, the discrepancy of different testing methods largely comes from the participation of a gas phase in the reaction, specifically the very different mass transport property of gas in the liquid and gas phase. For example, in the second paragraph of the Introduction section of this paper¹¹, the authors wrote: ‘... However, the RDE suffers from an inherent drawback when it comes to characterize reactions involving gaseous reactants (like those of PEMFC and the ORR in

particular): the gas (here O_2) is transported to the catalyst interface whilst dissolved in the electrolyte, the solubility being small (*ca.* 1 mmol L^{-1}) and the diffusion coefficient also (*ca.* $10^{-6} \text{ cm}^2 \text{ s}^{-1}$). This means that O_2 transport is roughly four decades slower than in the gas phase; this is why one observes diffusion-convection (mass-transport)-limited plateaus in RDE ...'. Also, in Table 1 of another paper¹², we find that all the reactions in comparison involve a gas phase, including OER, HER, HOR, ORR.







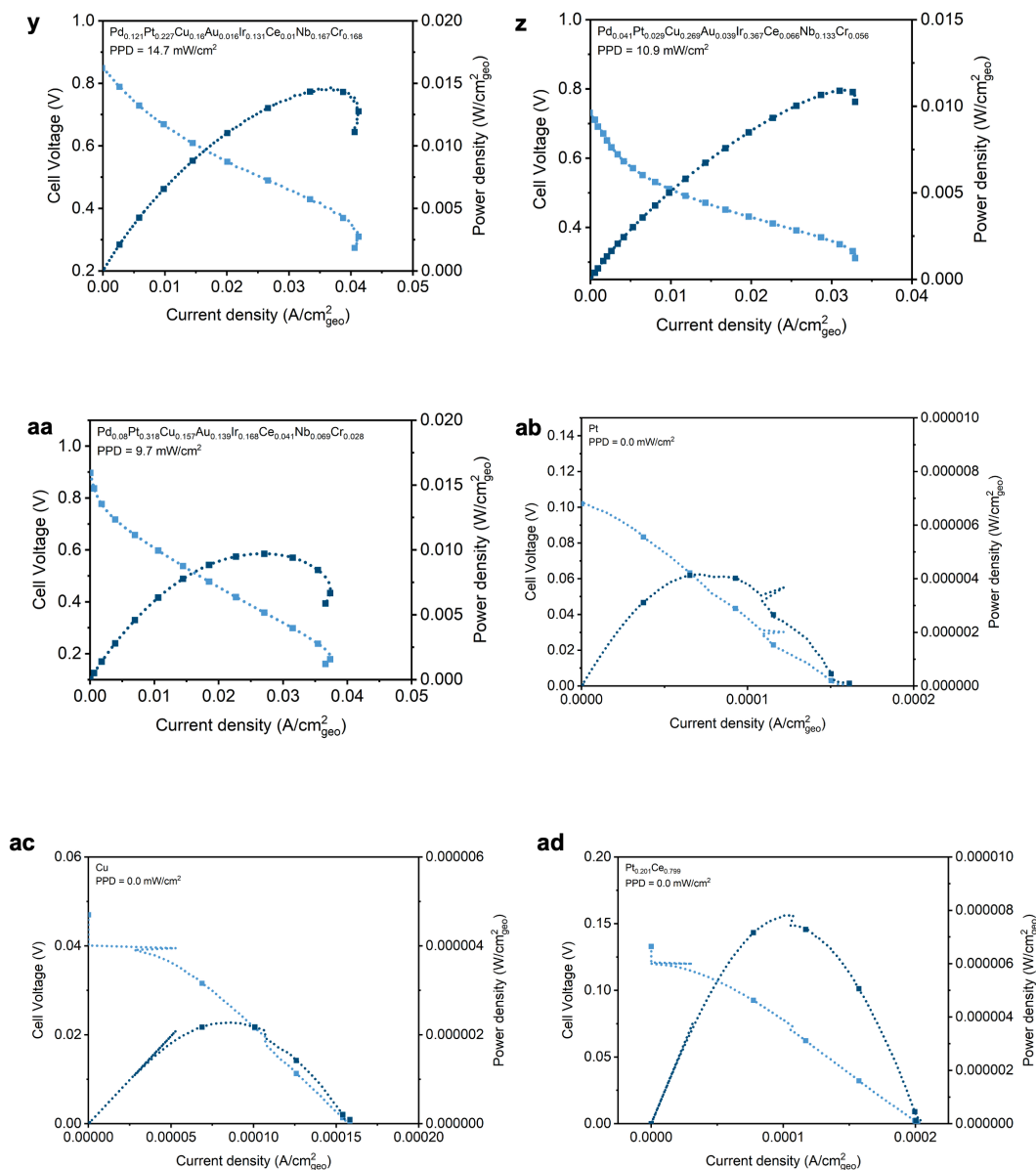


Fig. S3. Polarization curve for 30 recipes tested in the MEA. The polarization curve of 30 selected chemistries tested in the direct formate fuel cells at 60 °C in an electrolyte of 1.0 M KOH and 2.0 M KOOCH with chemistry labelled on each plot. To enable comparison with literature, all samples had a total catalyst loading of 2.0 mg/cm². Testing details are in the Method.

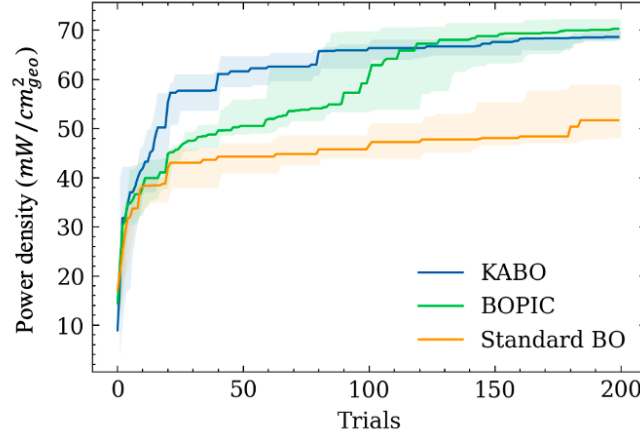


Fig. S4. Comparison plot for the convergence efficiency for the three algorithms by simulation. The plot compares convergence rates across different optimization strategies, plotting power density and trial numbers on the y and x axes. The solid line shows the mean value, and the shade indicates the uncertainty. When implementing 20 random seeds, the KABO (with the text-embedding-ada-002 by OpenAI) and BOPIC active learning algorithms achieved faster convergence than standard BO. The solid lines represent median values, while the shaded regions indicate the range between the first and third quartiles. Each curve was obtained by averaging over 20 independent runs, and it could be observed that on average, the standard BO algorithm tends to be stuck in local optima, while both the KABO and BOPIC approach would continuously make progress and find the global optima.

We conducted a quantitative analysis based on the results of our evaluation framework using these three algorithms. First, we measured how much the power density value could increase under the same trials compared to standard BO, which is the Enhancement Factor. Then, we also measured how many fewer experiments could be expected to find the same power density compared to the general baseline, which is the Acceleration Factor. The measurement of AF only adopted on the values are expected to exceed 5 times the performance of a standard pure Pd sample.

$$\text{Enhancement Factor} = \sum_{i=1}^n \frac{Top_i / Top_i^{BO}}{n}$$

$$\text{Acceleration Factor} = \sum_{y=60}^{\max(y_{200}^{BO})} \frac{\min \{i | Top_i > y\} / \min \{i | Top_i^{BO} > y\}}{\max(y_{200}^{BO}) - 5 \times y_{Pd}}$$

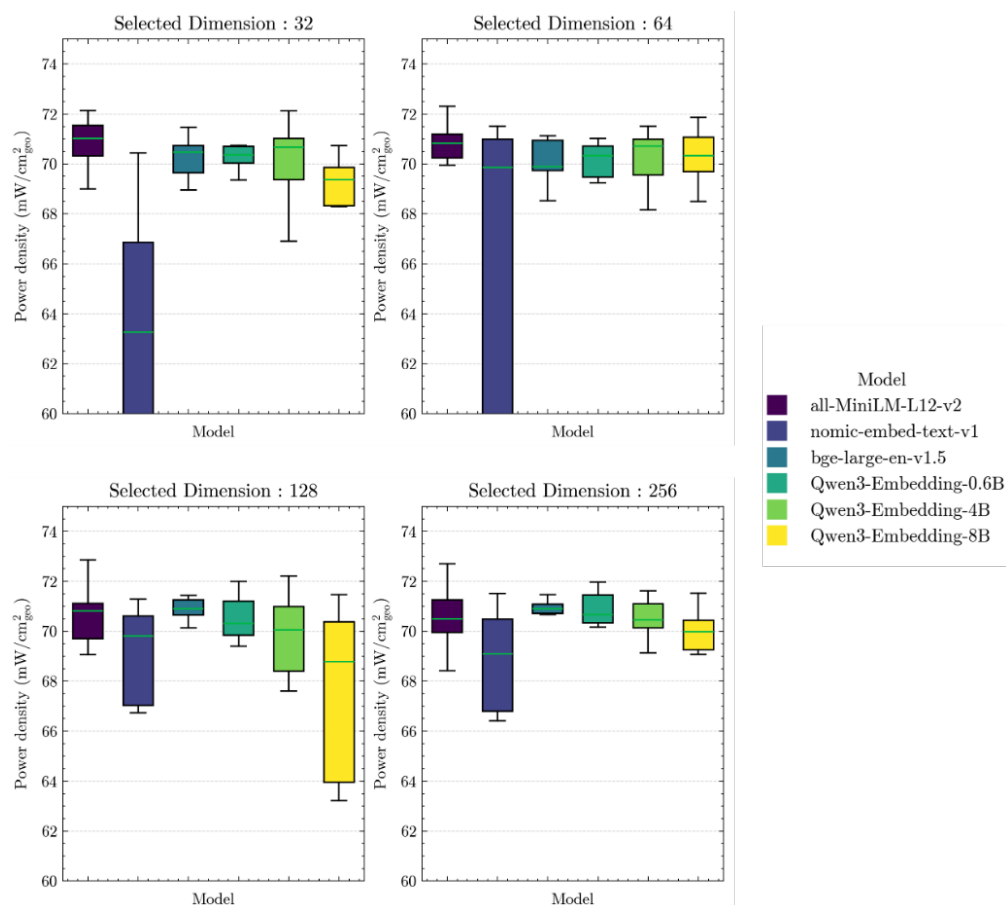


Fig. S5. Performance comparison of leveraging different open-source text embedding models to perform active learning with different dimensions of features. In all cases, a high performance value of around $70 \text{ mW}/\text{cm}^2$ has been searched, which shows high performance of all models with our PCA method. As a reminder, the benchmark Pd catalyst shows a low performance of around $13.5 \text{ mW}/\text{cm}^2$.

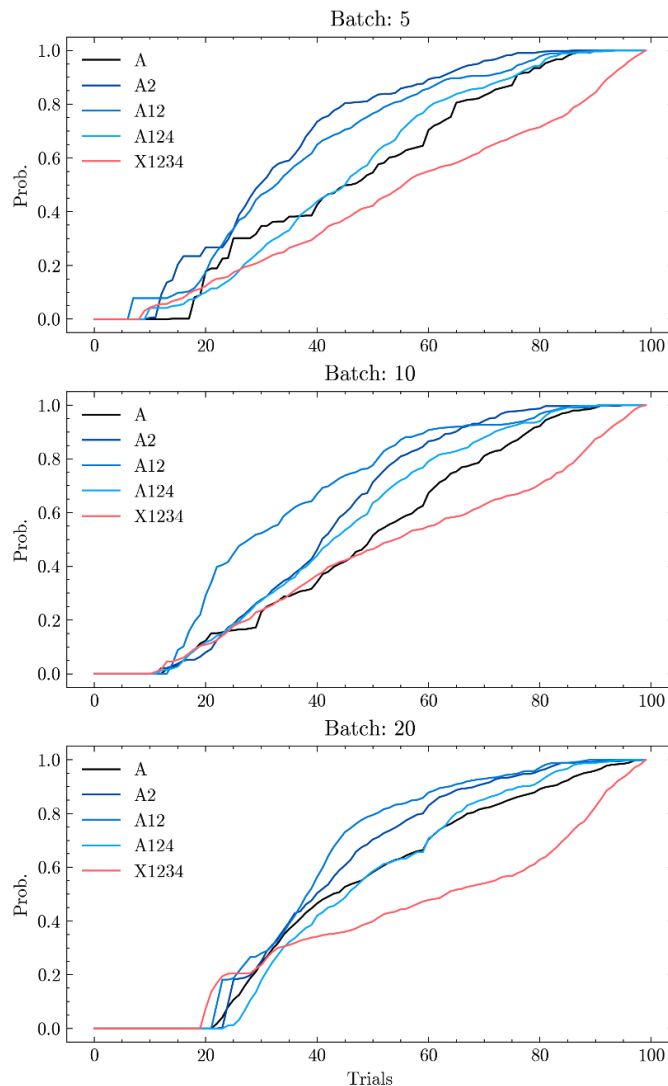


Fig. S6. Performance of active learning when using different combinations of elemental and morphological input (A, A2, A12, A124, X1234) with different batch size (5, 10, and 20) by simulation. Herein, we prepared an experimental dataset of 100 recipes, each with chemical ratios, SEM images, and power density performance. We simulate the probability of successfully locating the best-performing recipes (y-axis) with trial numbers (x-axis).

The notations are defined as below.

- (1) *A*: Elemental vector (8D). This configuration only considers the elemental input and thus takes an 8D input vector.
- (2) *A2*: Elemental vector (8D) combined with the intercept of the particle size distribution function (1D). This configuration takes a 9D input vector.

(3) *A12*: Elemental vector (8D) combined with both the slope and intercept of the particle size distribution function (2D). This configuration takes a 10D input vector.

(4) *A124*: Elemental vector (8D) combined with the slope and intercept of the particle size distribution function, along with the total particle count per unit area (3D). This configuration takes a 11D input vector.

(5) *X1234*: Only microstructural features, which include the slope and intercept of the particle size distribution function, together with the particle coverage ratio and total particle count per unit area (i.e., excluding elemental features) (4D). This experiment serves as the ablation experiment where the search was performed using only these auxiliary microstructural features. This configuration takes a 12D input vector.

For each condition, we conducted 50 independent trials using different random seeds and averaged the results. We examined three acquisition batch sizes—5, 10, and 20—to simulate the iterative selection process in high-throughput experimentation. As could be seen, the curves (*A2*, *A12*, *A124*) consistently converge faster when both elemental and microstructural features are considered, while *only* considering microstructural features (*X1234*) leads to worse results than the baseline result (*A*, considering elemental input only). This result underscores the importance of jointly considering elemental and microstructural factors. This result also shows the effectiveness of using the four suggested morphological features as input (along with chemical input) for active learning in our application.

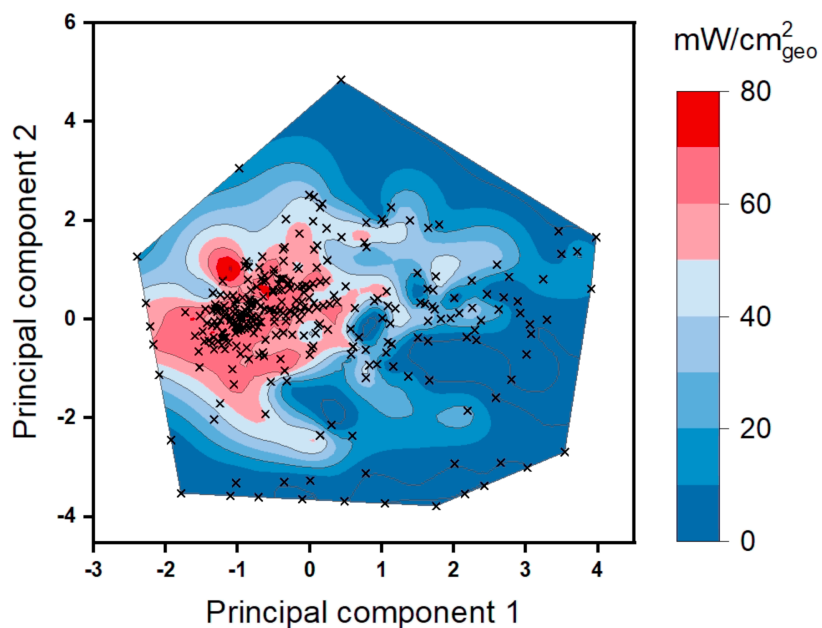


Fig. S7. Principal component analysis (PCA) plot with contour lines for octonary alloys (Pd-Pt-Cu-Au-Ir-Ce-Nb-Cr). This is the same figure as Fig. 1c with axis scale shown. The physical significance of PCA could be best explained by the *Principal Component Analysis (PCA) loading matrix*, which shows the correlation between original variables (herein the ratio of the eight elements) and the newly formed principal components (PC_1 , PC_2 ,). These loadings indicate how much each original variable contributes to each component, with values ranging from -1 to 1 . Briefly, to obtain the *PCA loading matrix*, we first standardize each element ratio to a z -score:

$$z_j = \frac{x_j - \mu_j}{\sigma_j}$$

Where x_j is raw ratio of element j , μ_j is the mean of that element, and σ_j is the standard deviation of that element across all recipes.

We then perform singular-value decomposition (SVD), and for a centered-and-scaled data set, the raw eigenvector coefficients already serve as loadings—they quantify how strongly each original variable contributes to each principal component. The computed PCA loading matrix is as below:

Element	Pd	Pt	Cu	Au	Ir	Ce	Nb	Cr
PC ₁	-0.602	-0.352	0.305	0.283	0.372	0.115	-0.036	0.434
PC ₂	0.283	-0.433	-0.278	-0.010	0.220	-0.489	0.560	0.231

To project the recipe ratio (8D space) with corresponding loading matrix coefficients back to the principal component space (2D space, as shown in Fig. 2c of the main text), we could simply use linear combination:

$$PC_k = \mathbf{z}^T \mathbf{w}_k = \sum_{j=1}^8 z_j w_{jk}$$

Where w_k is the loading vector for PC_k . w_k is an orthonormal eigenvector.

How to read PC₁?

- Large negative scores (left side of this figure) arise from alloys enriched in Pd and Pt—the two platinum group metals (PGMs) with the largest negative loadings in the PCA loading matrix. Both elements are also well-established catalysts for the formate/formic-acid oxidation reaction.
- Large positive scores (right side of this figure) correspond to compositions where Pd/Pt content is lower and Au, Cr, Cu, plus Ir become comparatively more prominent.
- Correlation with performance: PC₁ shows a strong negative correlation with power density. Thus, Pd/Pt-rich samples generally deliver higher power densities, although a few Ir/Au-rich outliers perform well too.

Interpretation: Thus, PC₁ could be thought as a ‘Pd/Pt or PGM content axis’ that arranges the library from Pd/Pt-dominated formulations (left) to compositions richer in Ir, Au, Cr, and Cu (right). It could be observed that a negative PC₁ of ~ -1 leads to the best performance.

How to read PC₂?

- Positive PC₂ scores pick out alloys rich in Nb (largest positive loading) and, to a lesser extent, Ir and Cr.
- Negative PC₂ scores mark compositions that are Ce-rich and Nb-poor, with some additional weighting from higher Pt. Ce carries the most negative loading; Pt is negative as well.
- The correlation with power density is modest. It hints that adding a moderate amount of Nb (and perhaps Ir) can boost performance once the overall Pd/Pt balance captured by PC₁ is already favorable.

Interpretation: PC₂ behaves like a ‘dopant trade-off axis’. It tracks how Nb (and secondarily Ir/Cr) are balanced against Ce (and a bit of Pt) within a Pd/Pt-dominated matrix. The highest power densities occur around PC₂ ≈ 0 , suggesting that an optimal Nb \leftrightarrow Ce (and Ir/Pt) balance—rather than an extreme in either direction—gives the best results when Pd/Pt content is already optimized.

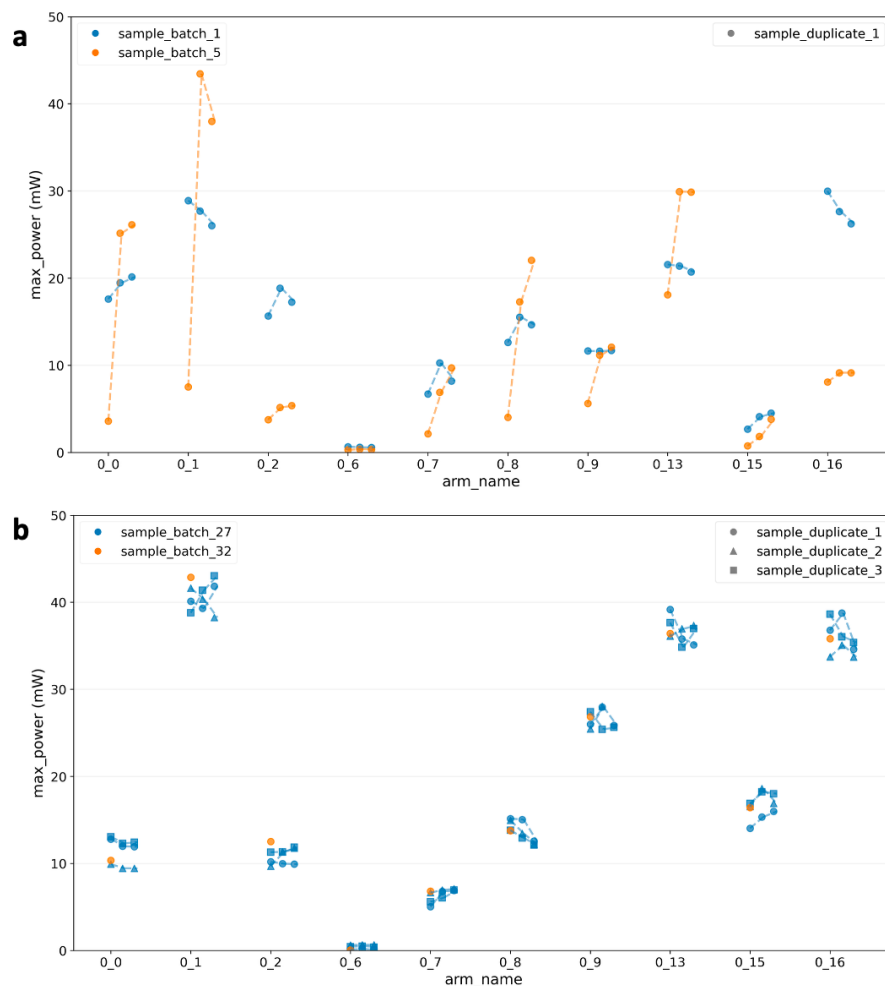


Fig. S8. Plots of reproducibility check. (a) Comparison plots of results from batches 1 and 5. (b) Comparison plots of results for batches 27 and 32 (Supplementary Data 1). The x -axis is the arm name, which stands for a specific recipe in this active learning regime, the y -axis is the maximum pseudo power density derived from the linear sweep voltammetry (LSV) scan, which is the objective to be optimized. The same batch of materials are synthesized at the same time. However, within the same batch, we may synthesize multiple samples of the same recipe. Also, each sample may be tested multiple times, which are connected by the dotted lines. Different batches of samples are synthesized at different times. For example, in batch 1, we synthesized 1 sample with recipe 0_0, and tested it in triplicate. Batch 32 was synthesized two months after batch 27, for the purpose of checking long-term baseline shift. Ideally, the performances with the same arm name should all align. Each sample is 1 cm^2 . The legend in grey at the top right corner only represents the shape symbol. All the experiments were conducted with the high-throughput robotic workflow, which shows that robots do not guarantee reproducibility.



Fig. S9. Example of AI for experimental diagnostics with the wood sample stages. Comparison of a failed sample stage (left) versus a successfully cut sample stage (right). For both sample stages, they were cut at the same location by the same laser device, with the same laser parameters. CRES^t ‘experimental diagnostician’ observed charring and unsuccessful cutting of the sample stage on the left. It pointed out that the issue came from the non-uniform surface texture and property of the wood. Thus, after selecting a higher-quality plywood, we successfully obtained a customized sample stage. Each slot is of size 1.0 cm \times 1.3 cm.

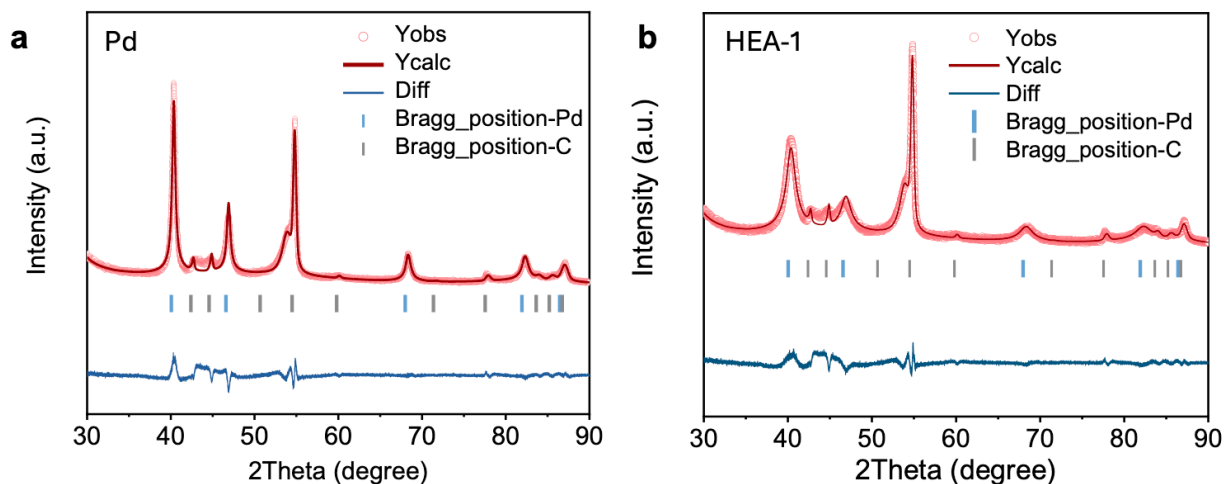


Fig. S10. X-ray diffraction (XRD) with Rietveld refinement for materials analysis. The catalyst Pd (a) and HEA-1 (b) represent the chemistry Pd and $\text{Pd}_{0.487}\text{Pt}_{0.185}\text{Cu}_{0.018}\text{Ir}_{0.037}\text{Ce}_{0.106}\text{Nb}_{0.168}$, respectively. Similar singular face-centered cubic (FCC) phase could be observed for the two chemistries. Details of methods, lattice constants and fitting parameters are in the Table S2.

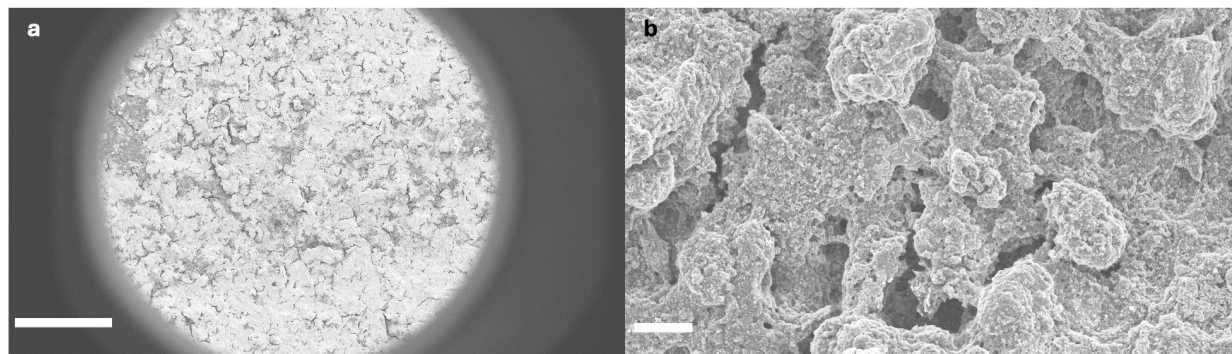


Fig. S11. SEM images of carbon paper loaded with 2.0 mg/cm² palladium black. The entire carbon paper was uniformly loaded with Pd black. This makes sure that our benchmark sample was properly prepared. The scale bar in (a) and (b) is 500 μm and 15 μm , respectively.

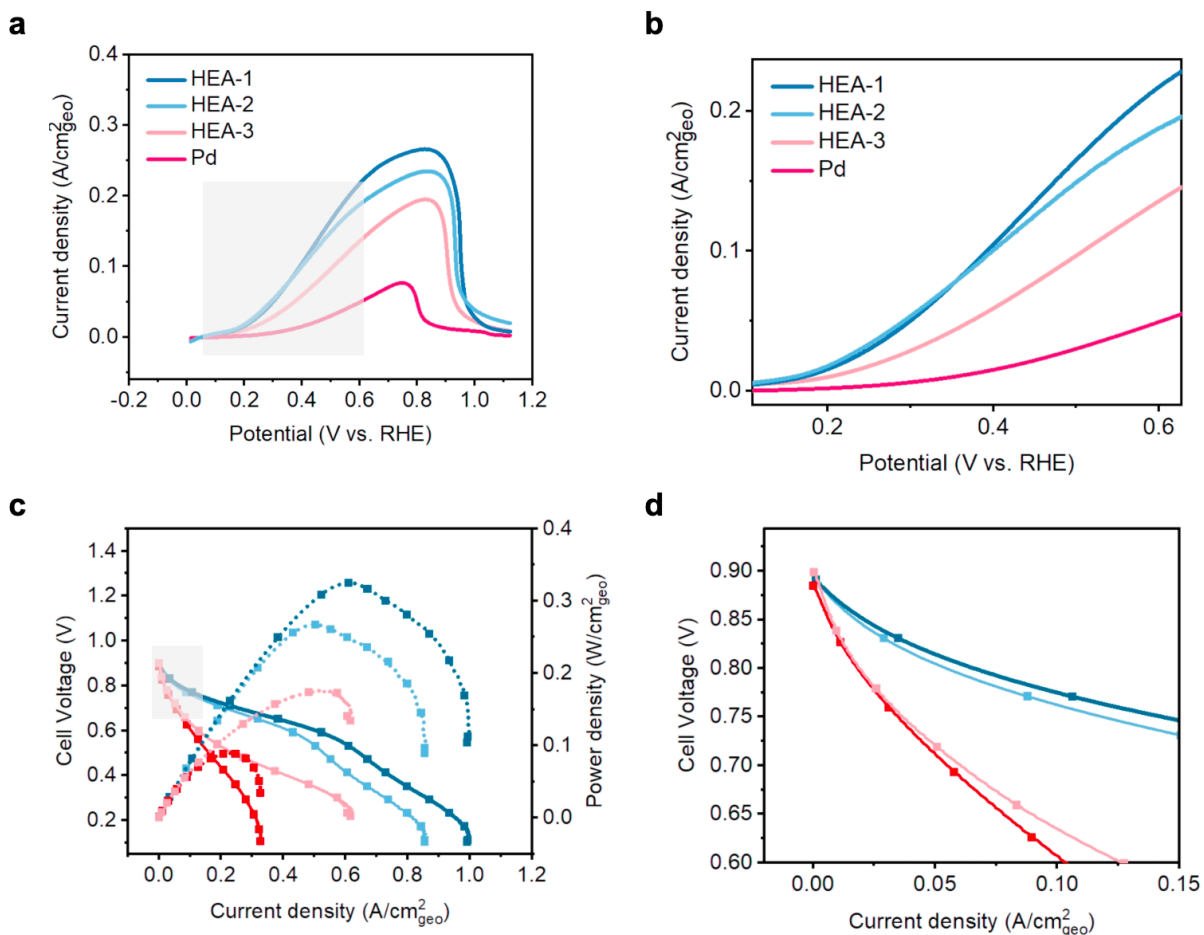


Fig. S12. Comparison of the activation loss region for tests in the three-electrode cell and the membrane electrode assembly. (a, b) Linear sweep voltammetry (LSV) test in the three-electrode cell. Scan rate = 10.0 mV/s. (b) is the zoomed-in region of the grey block region in (a). (c, d) Polarization test in the membrane electrode assembly. (d) is the zoomed-in region of the grey block region in (c). The catalyst HEA-1, HEA-2, HEA-3 and Benchmark represent the chemistry composition of $Pd_{0.487}Pt_{0.185}Cu_{0.018}Ir_{0.037}Ce_{0.106}Nb_{0.168}$, $Pd_{0.381}Pt_{0.080}Cu_{0.009}Au_{0.004}Ir_{0.02}Ce_{0.086}Nb_{0.338}Cr_{0.082}$, $Pd_{0.635}Pt_{0.258}Cu_{0.107}$ and Pd, respectively. In terms of intrinsic activity, a consistent trend of HEA-1 > HEA-2 > HEA-3 > Pd was found in both tests.

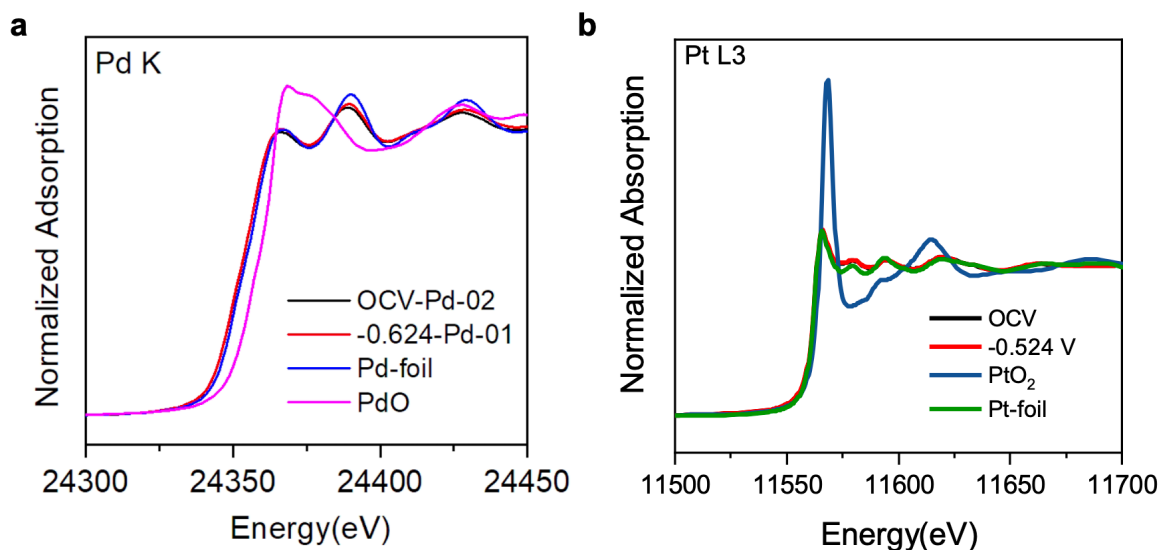


Fig. S13. XANES analysis for Pd and Pt in the HEA-8D catalyst. (a) Pd K-edge XANES spectra of the HEA-8D catalyst at OCV and -0.524 V (vs. Hg/HgO in 1.0 M KOH and 1.0 M HCOOK aqueous solution), compared with reference spectra of Pd foil and PdO, revealing the metallic state of Pd under reaction conditions. (b) Pt L₃-edge XANES spectra of the HEA-8D catalyst at OCV and -0.524 V (vs. Hg/HgO in 1.0 M KOH and 1.0 M HCOOK aqueous solution), compared with Pt foil and PtO₂, highlighting the electronic structure and stability of Pt.

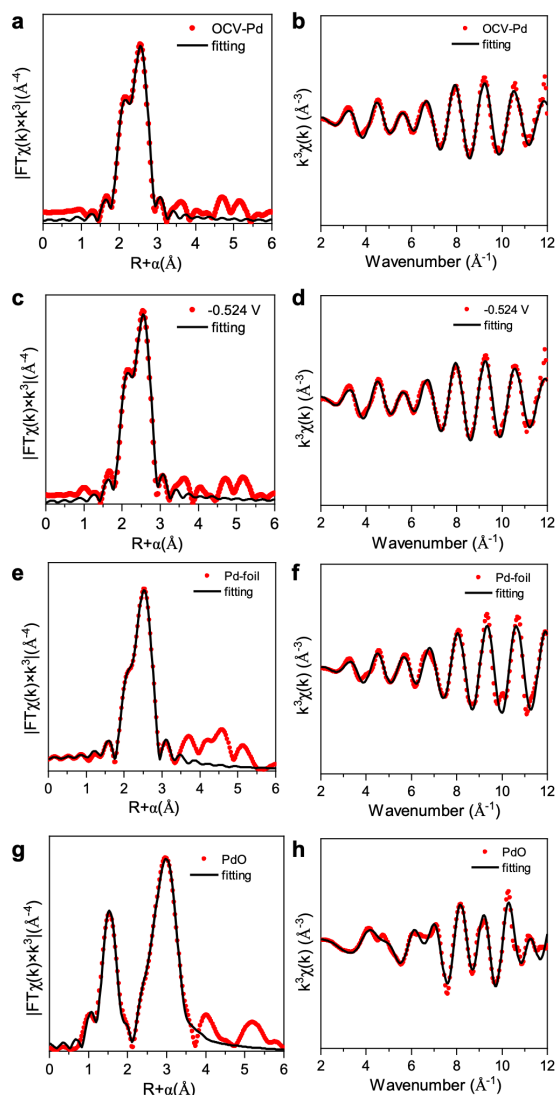


Fig. S14. EXAFS fitting curves at the Pd K-edge in R-Space and K-Space for HEA-8D and reference catalysts. (a) EXAFS fitting results at the Pd K-edge in R-space for HEA-8D at OCV. (b) EXAFS fitting results at the Pd K-edge in K-space for HEA-8D at OCV. (c) EXAFS fitting results at the Pd K-edge in R-space for HEA-8D at -0.524 V (vs. Hg/HgO in 1.0 M KOH aqueous solution). (d) EXAFS fitting results at the Pd K-edge in K-space for HEA-8D at -0.524 V (vs. Hg/HgO in 1.0 M KOH aqueous solution). (e) EXAFS fitting results at the Pd K-edge in R-space for Pd foil. (f) EXAFS fitting results at the Pd K-edge in K-space for Pd foil. (g) EXAFS fitting results at the Pd K-edge in R-space for PdO. (h) EXAFS fitting results at the Pd K-edge in K-space for PdO. The fitting curves at the Pd K-edge show excellent agreement with experimental data, validating the accuracy and reliability of the structural parameters derived for Pd. These results provide detailed insights into the local atomic environment of Pd in the HEA-8D catalyst under various conditions and in reference systems, highlighting the structural factors critical to the catalytic functionality of Pd.

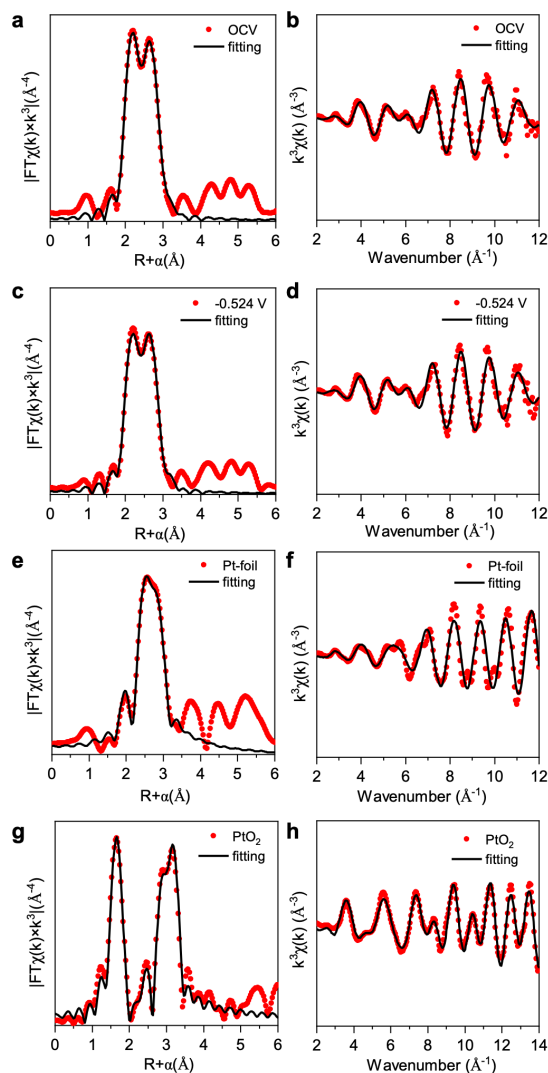


Fig. S15. EXAFS fitting curves at the Pt L₃-edge in R-Space and K-Space for HEA-8D and reference catalysts. (a) EXAFS fitting results at the Pt L₃-edge in R-space for HEA-8D at OCV. (b) EXAFS fitting results at the Pt L₃-edge in K-space for HEA-8D at OCV. (c) EXAFS fitting results at the Pt L₃-edge in R-space for HEA-8D at -0.524 V. (d) EXAFS fitting results at the Pt L₃-edge in K-space for HEA-8D at -0.524 V. (e) EXAFS fitting results at the Pt L₃-edge in R-space for Pt foil. (f) EXAFS fitting results at the Pt L₃-edge in K-space for Pt foil. (g) EXAFS fitting results at the Pt L₃-edge in R-space for PtO₂. (h) EXAFS fitting results at the Pt L₃-edge in K-space for PtO₂. The fitting curves at the Pt L₃-edge exhibit excellent agreement with experimental data, confirming the reliability and accuracy of the derived structural parameters for Pt. These results offer valuable insights into the local atomic environment of Pt in the HEA-8D catalyst under different conditions, providing comparative benchmarks with Pt foil and PtO₂, and highlighting structural factors influencing the catalytic behavior of Pt.

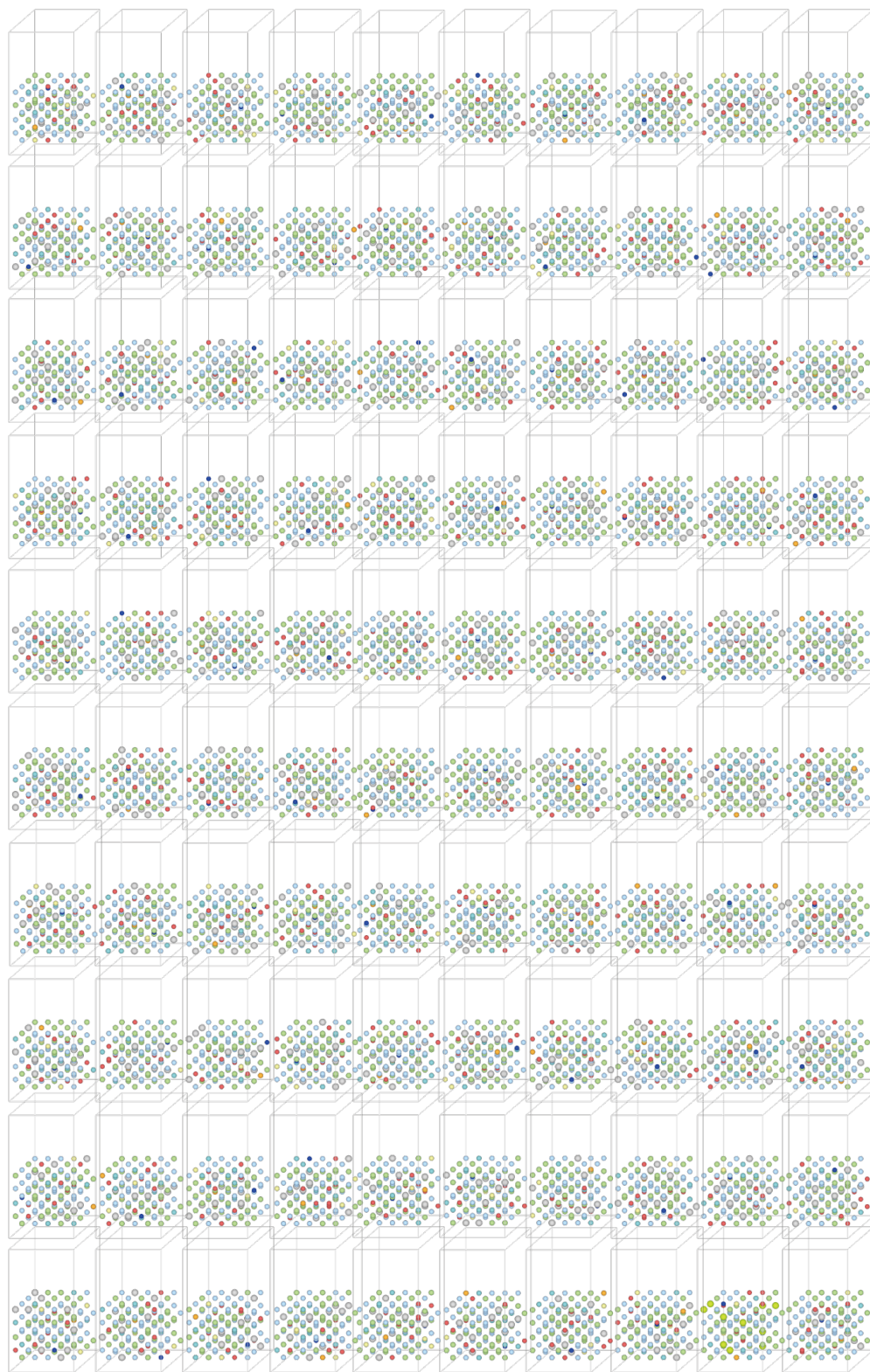


Fig. S16. Atomic configurations of 100 randomly generated octonary alloy surface models for structural screening.

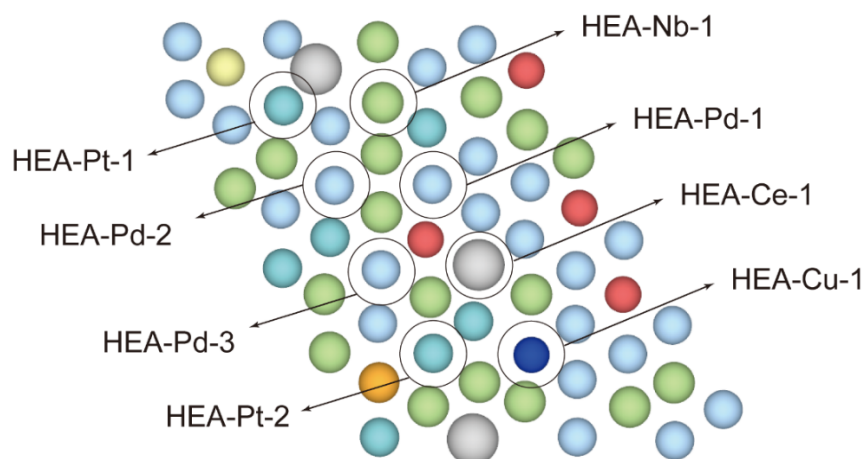


Fig. S17. Identified catalytically active sites on the octonary alloy surface model, including Pd, Pt, Ce, Cu, and Nb atoms with distinct local environments.

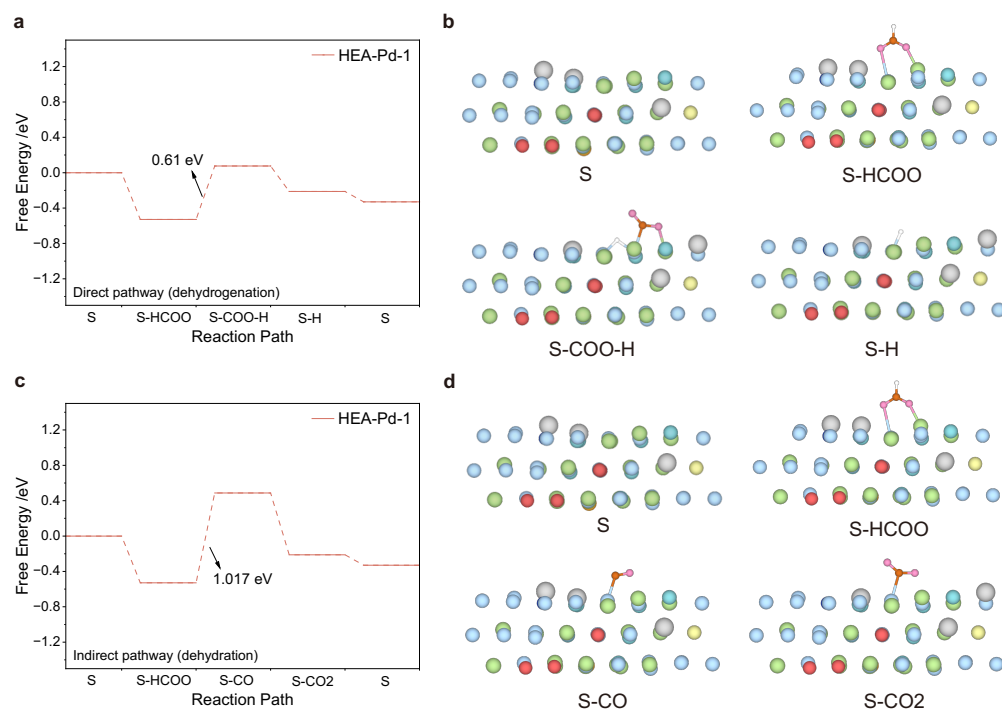


Fig. S18. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Pd-1. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Pd-1.

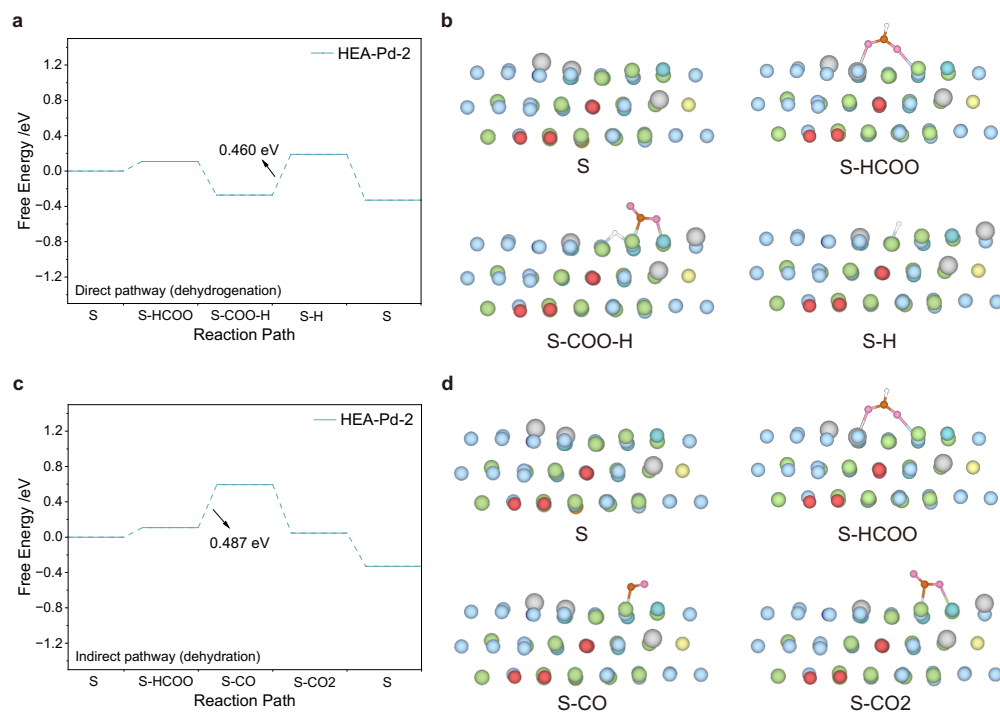


Fig. S19. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Pd-2. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Pd-2.

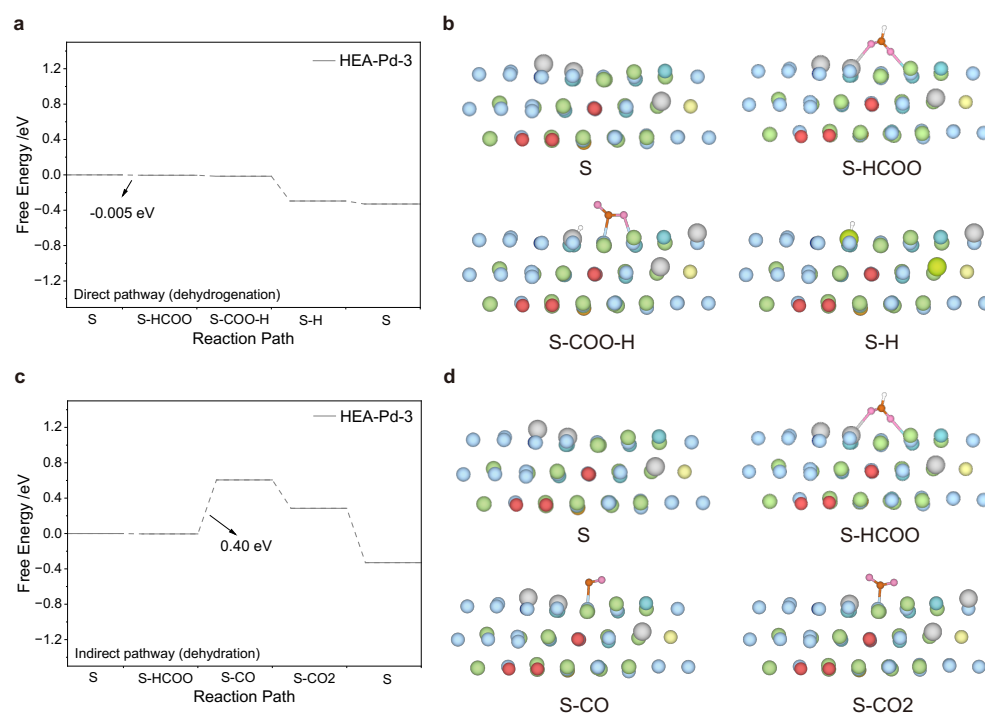


Fig. S20. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Pd-3. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Pd-3.

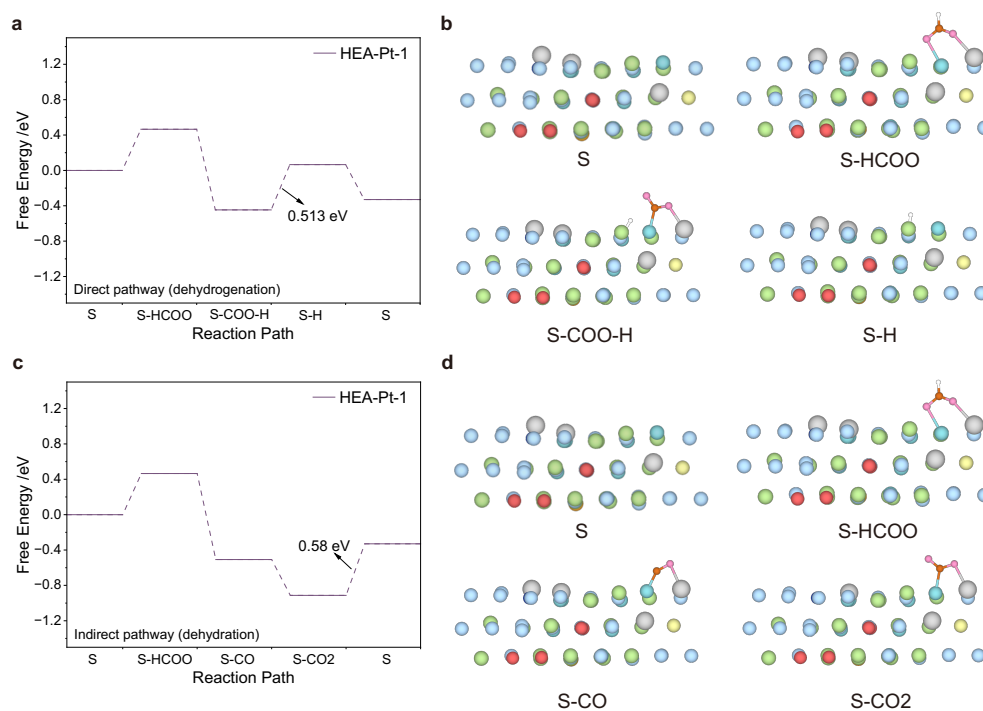


Fig. S21. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Pt-1. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Pt-1.

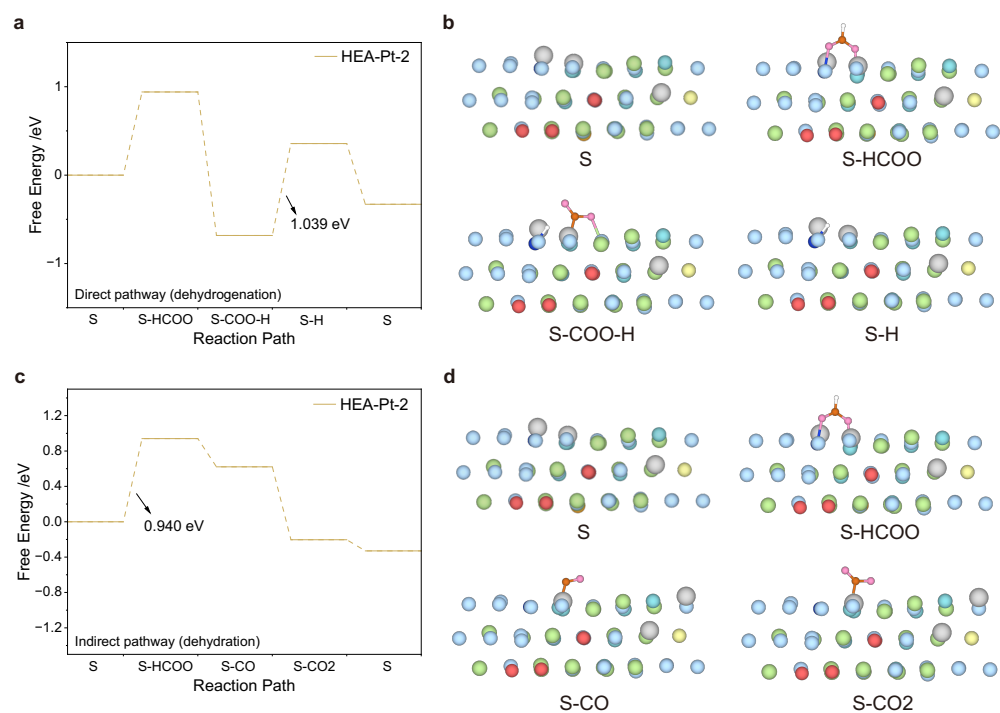


Fig. S22. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Pt-2. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Pt-2.

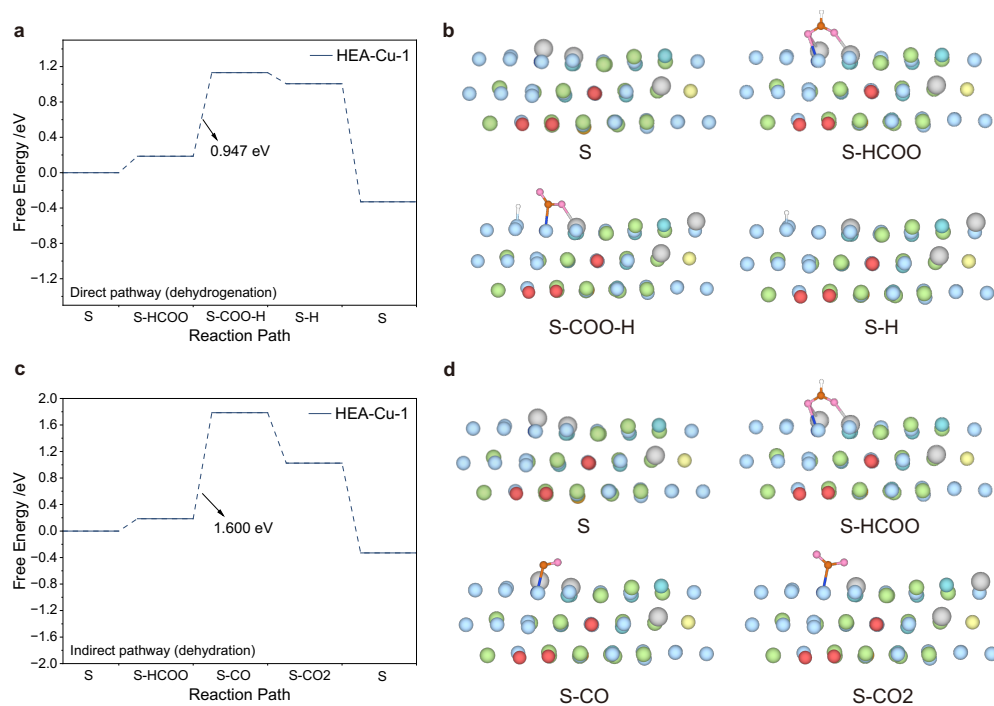


Fig. S23. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Cu-1. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Cu-1.

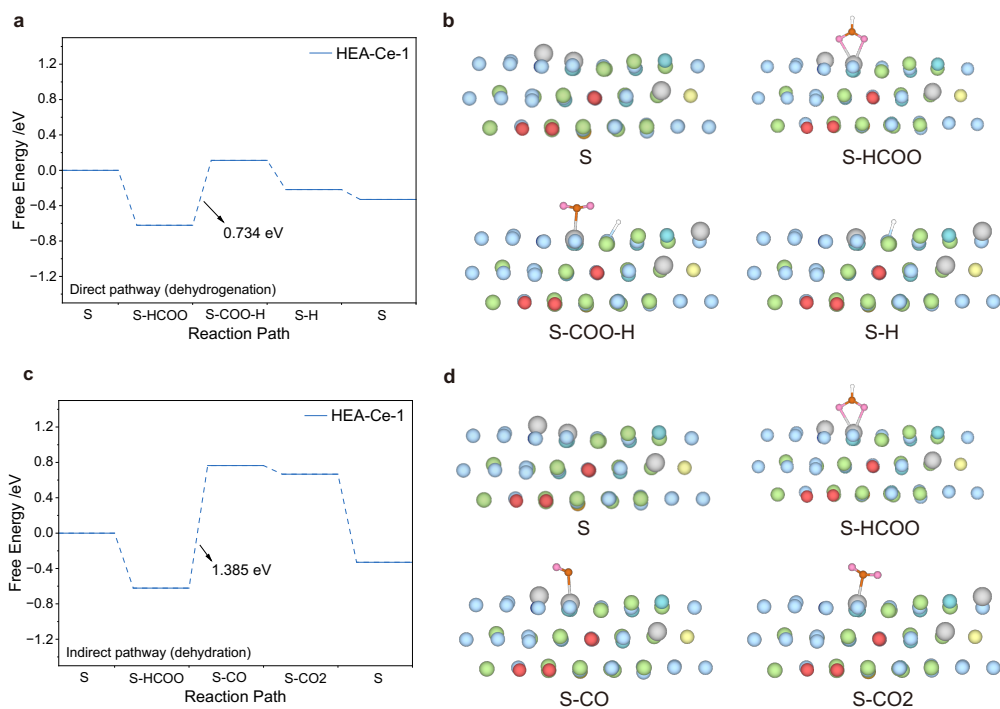


Fig. S24. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Ce-1. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Ce-1.

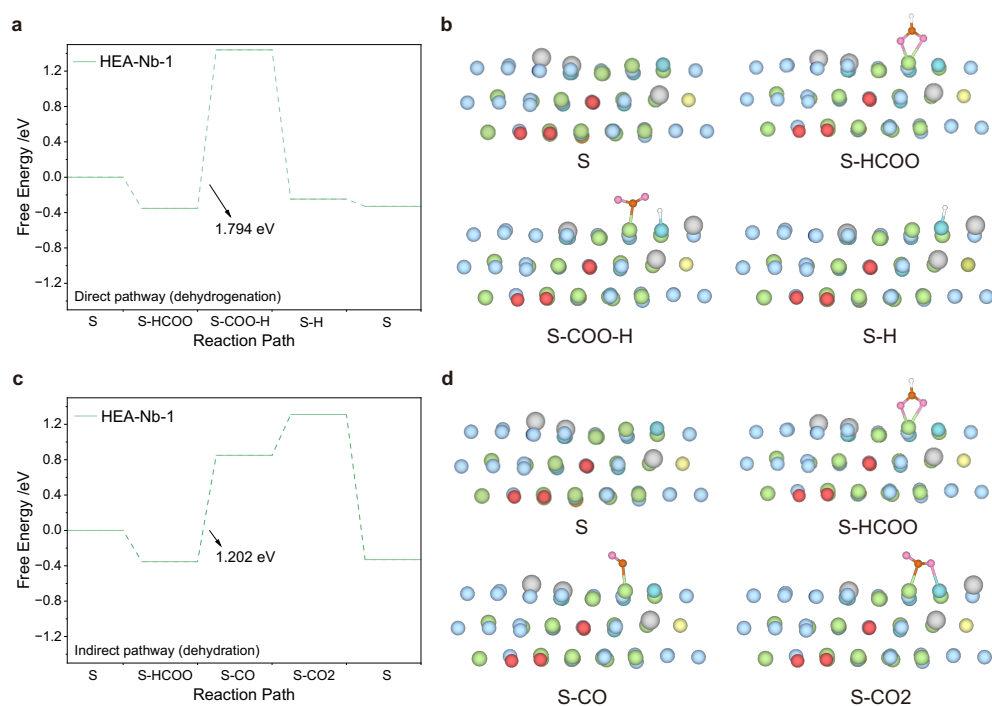


Fig. S25. (a–b) Free energy profile (a) and corresponding optimal configurations of direct pathway (b) on HEA-Nb-1. c–d, Free energy profile (c) and corresponding optimal configurations of indirect pathway (d) on HEA-Nb-1.

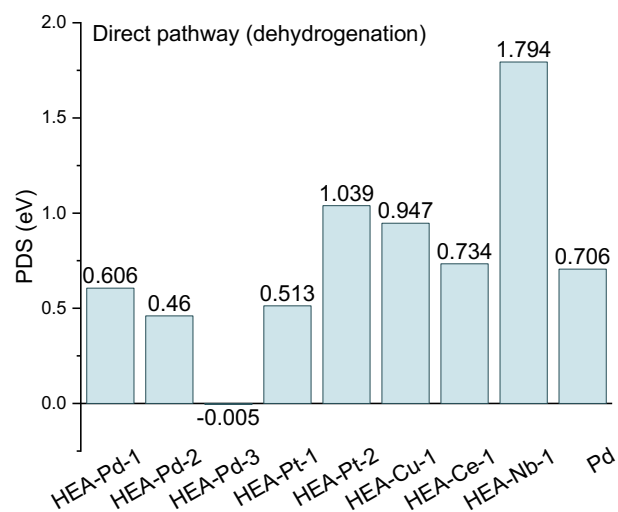


Fig. S26. Comparison of potential-determining step (PDS) free energy barriers at different HEA active sites along the direct reaction pathway. Structural tuning at the Pd site lowers the PDS to -0.03 eV, in sharp contrast to 0.71 eV for pure Pd.

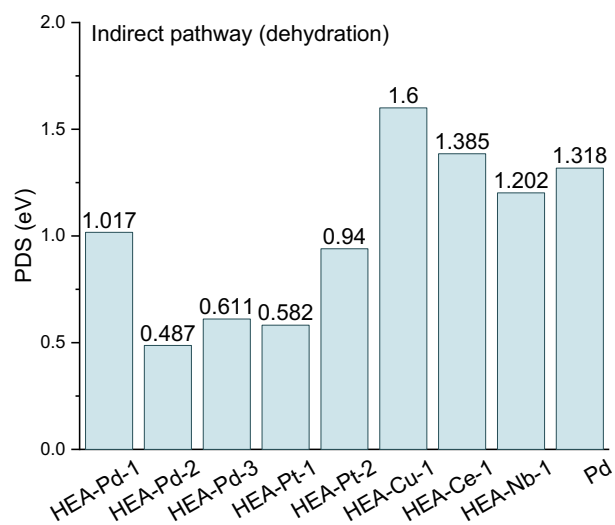


Fig. S27. Comparison of potential-determining step (PDS) free energy barriers at different HEA active sites along the indirect reaction pathway. The Pd site in the HEA exhibits a significantly reduced PDS (0.487 eV) compared to pure Pd (1.318 eV), highlighting the effect of local structural modulation.

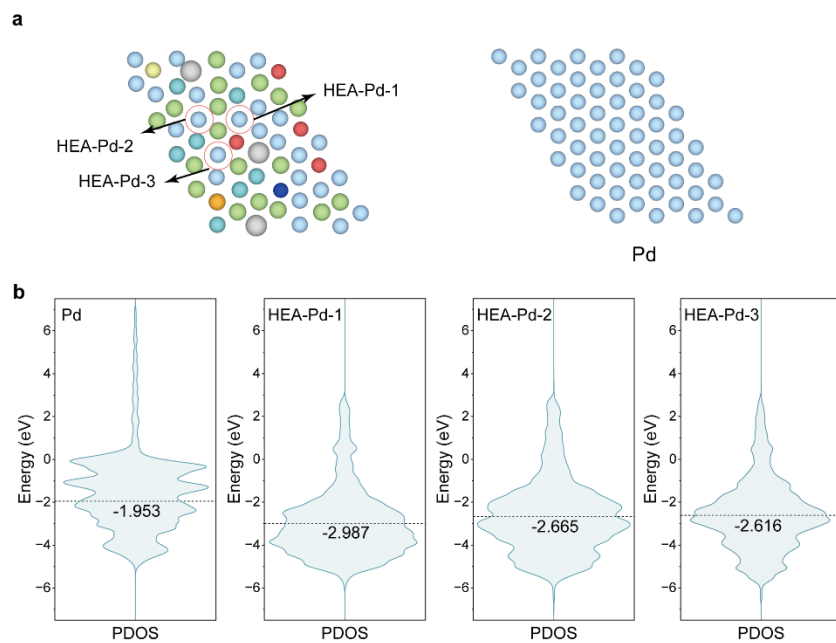


Fig. S28. Partial density of states (PDOS) analysis for the Pd site on the pure Pd surface and three representative Pd sites on the HEA surface. The calculated d-band centers are indicated in the figure.

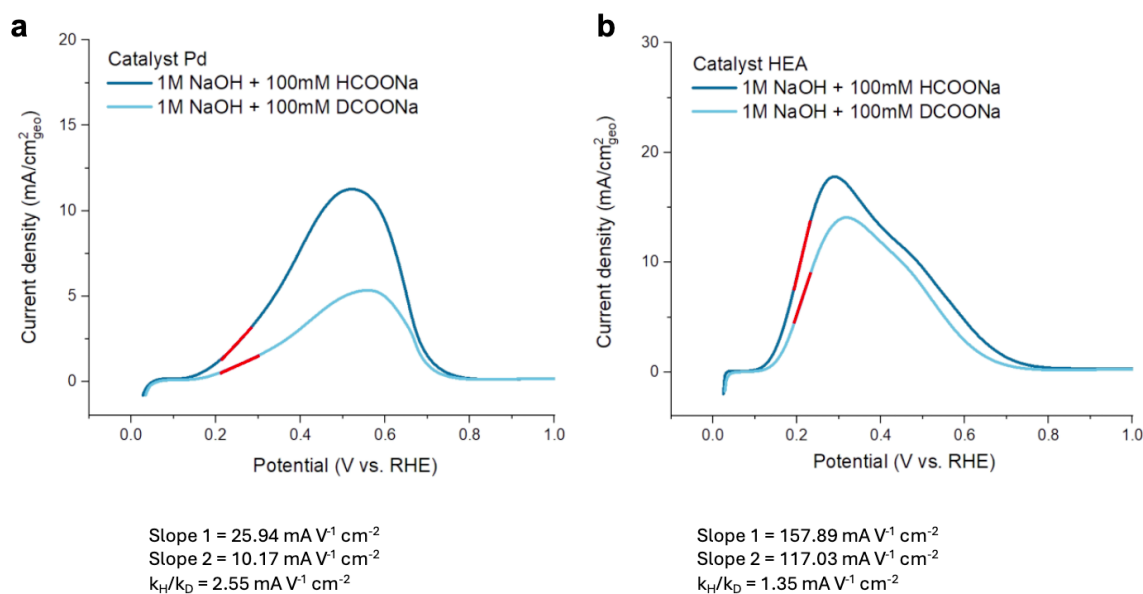


Fig. S29. Isotope experiments to study the kinetics of formate oxidation. (a) Linear sweep voltammetry (LSV) of a pure Pd catalyst. (b) LSV test of our optimized HEA catalyst. Both samples have been sufficiently activated with 30 scans of cyclic voltammetry (CV) test before conducting the LSV measurement. The reaction occurred in a mixture of 1.0 M NaOH and 100 mM HCOONa (or DCOONa) in a standard three-electrode cell setup. The reference electrode was Hg/HgO, and the counter electrode was the Pt foil. The LSV scan rate was 10.0 mV/s. The catalyst HEA refers to Pd_{0.381}Pt_{0.080}Cu_{0.009}Au_{0.004}Ir_{0.02}Ce_{0.086}Nb_{0.338}Cr_{0.082}. For this experiment, the same molar loading amount of catalysts was applied for the pure Pd sample and the optimized HEA sample. Following methods in the reference, we compared the LSV slope ratio at around 100 mV more positive than the onset potential of formate oxidation (regions corresponding to the oxidation of H_{ads}). For the pure Pd sample, an LSV slope ratio (indicative of the H/D KIE value) of 2.55 mA V⁻¹ cm⁻² was observed, which is very close to the value of 2.40 mA V⁻¹ cm⁻² reported in the literature¹³. While for our optimized multi-element catalyst, a much reduced slope ratio of only 1.35 mA V⁻¹ cm⁻² was observed. This shows that for the pure Pd sample, similar to what was found in literature, the H_{ads} formed by the oxidation of HCOO⁻ serves as the site-occupying species that suppress the formate oxidation reaction, and the oxidation of surface adsorbed H_{ads} (H_{ads} + OH⁻ → H₂O + e⁻) is the rate determining step (RDS). While for our optimized catalyst, due to the tuned surface chemistries, an H/D KIE value that is smaller than 2.0 was observed and thus this oxidation of H_{ads} is not considered as the RDS (which in turn suggests that the H_{ads} oxidation is much more facilitated on our optimized catalyst)¹³. This indicates a mechanism change for the two catalysts. Other elementary steps such as (HCOO⁻ → H_{ads} + COO_{ads}⁻) or (COO_{ads}⁻ → CO₂ + e⁻) may serve as the RDS in this reaction. Also, a much larger LSV slope for the HEA catalyst was observed, and this indicated an overall facilitated kinetics for the HEA catalyst.

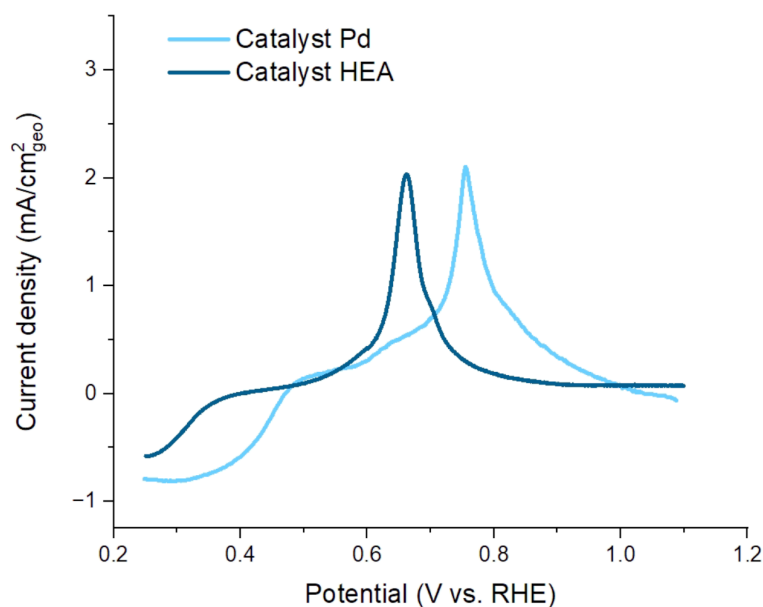


Fig. S30. CO stripping experiments. The CO stripping curve for a pure Pd catalyst and our optimized HEA catalyst. For both curves, the background was subtracted. Both samples have been sufficiently activated with 30 scans of cyclic voltammetry (CV) test before conducting the CO stripping measurement. The reaction occurred in 1.0 M KOH in a standard three-electrode cell setup. The reference electrode was Hg/HgO, and the counter electrode was a carbon rod. For the CO stripping experiment, the CO gas was first bubbled into the electrolyte for 20 min while holding the potential at 0.1 V vs. RHE. Next, the CO gas was switched to the N₂ gas for 30 min while keeping the same potential. Then, the LSV was tested, and the scan rate was 20.0 mV/s. The catalyst HEA refers to Pd_{0.381}Pt_{0.080}Cu_{0.009}Au_{0.004}Ir_{0.02}Ce_{0.086}Nb_{0.338}Cr_{0.082}. For the pure Pd catalyst, the CO stripping peak at 0.754 V (vs. RHE) was observed, which matches well with the literature value of 0.75 V¹³. An obvious negative shift (of around -0.1 V) of the primary CO stripping peak was observed for our optimized catalyst, as compared with the pure Pd catalyst. This indicates that our optimized catalyst is less susceptible to CO poisoning as the CO_{ads} could be stripped off the surface at a more negative potential.

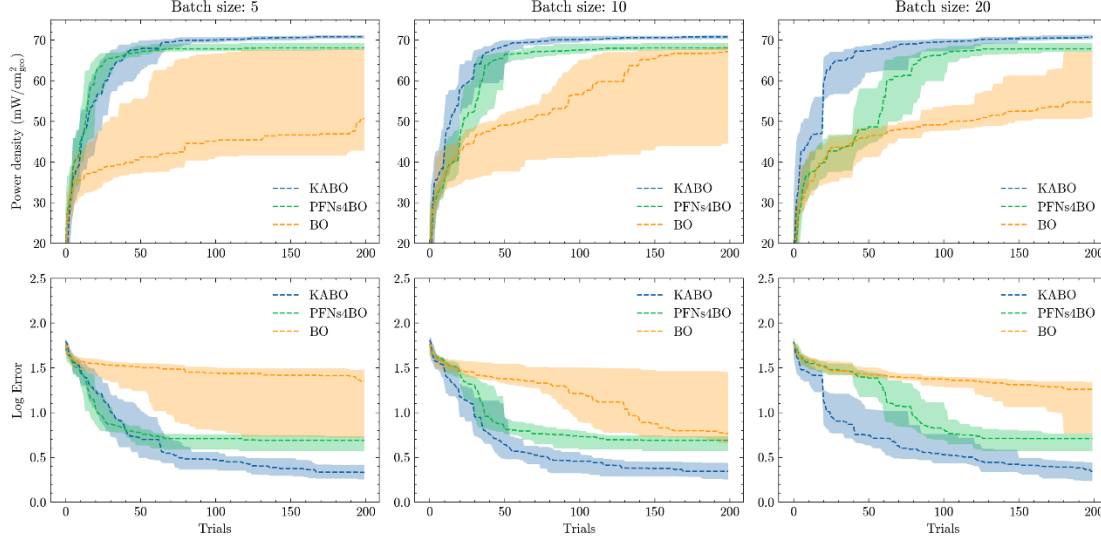


Fig. S31. Comparison of results (performance and convergence) for active learning with our knowledge-assisted Bayesian optimization (KABO) method, the PFNs4BO transformer method, and the standard BO with different batch size. As could be seen, both our KABO method and the transformer method exhibits superior performance compared with standard BO, showing the feasibility for both methods for active learning. The dotted line shows the mean value, and the shaded region shows the uncertainty, defined by the interquartile range between the 25th percentile (Q1) and the 75th percentile (Q3).

To broadly examine how BO evolution can improve performance, we have now included the PFNs4BO framework for comparison¹⁴. This network uses a *Transformer architecture* to fit prior data via in-context learning and has demonstrated strong results in multiple studies. Also, we briefly explored an alternative approach that involves direct parameter optimization through iterative natural language conversation with LLMs, rather than traditional numerical methods. However, this strategy exhibited significant limitations. Conversational interface constraints restrict query scope, and pervasive LLM hallucination consistently hampered further progress once a preliminary optimization level was reached. For example, our preliminary trials with this method converged to a suboptimal power density value of 56 mW/cm² within approximately 50 iterations, after which improvement stalled. Consequently, this approach was not the primary focus of the current study.

Constrained by the data processing methodology of PFNs4BO, we re-implemented the entire experimental pipeline using BoTorch. We assessed performance differences across batch sizes of 5, 10 and 20, while maintaining a fixed experimental budget of 200 trials. Each trial was repeated 32 times with different random seeds to ensure robustness. For the PFNs4BO experiments, we employed the most proficient ‘hebo_plus_model’ as the pre-trained model, utilizing a UCB acquisition strategy.

In terms of optimizing the objective function, beyond merely displaying the best value identified by the current active learning iteration, we also plotted its deviation from the literature's established optimal power density of around 74 mW/cm². Under experimental conditions involving small batch sizes and numerous iterations, the PFNs4BO method indeed demonstrated significant advancements compared to traditional baseline Bayesian optimization, and its performance closely paralleled our KABO algorithm. Conversely, with larger batch sizes and fewer iterations, our KABO algorithm exhibited a notable advantage. This superiority stems from KABO's ability to improve the data distribution space under zero-shot conditions, rather than solely refining subsequent active learning algorithms.

Another noteworthy observation pertains to the inference time for both methods, assessed on a CPU. In trials with a batch size of 5, PFNs4BO required approximately 912 seconds to complete a single experiment. In contrast, KABO, aligning with the optimization efficiency of standard BO approaches, only needed 126 seconds for the same task. This substantial difference in computational overhead is a crucial consideration for applications demanding high-throughput, real-time optimization.

Pd	Pt	Cu	Au	Ir	Ce	Nb	Cr	max_power (mW cm ⁻²)	Normalized Cost (\$)	Ratio (mW cm ⁻² \$ ⁻¹)
0.487	0.185	0.018	0.000	0.037	0.106	0.168	0.000	75.450	66.324	1.138
0.454	0.147	0.019	0.019	0.055	0.106	0.200	0.000	74.400	63.842	1.165
0.464	0.215	0.000	0.000	0.000	0.130	0.191	0.000	73.540	62.983	1.168
0.459	0.197	0.066	0.000	0.000	0.112	0.165	0.000	69.970	61.094	1.145
0.412	0.157	0.015	0.000	0.031	0.090	0.142	0.153	66.990	56.123	1.194
0.405	0.154	0.015	0.000	0.031	0.088	0.140	0.167	65.240	55.192	1.182
0.403	0.143	0.040	0.000	0.018	0.121	0.263	0.012	63.170	52.980	1.192
0.381	0.080	0.009	0.004	0.020	0.086	0.338	0.082	58.300	46.563	1.252
0.371	0.068	0.061	0.055	0.019	0.070	0.296	0.060	58.270	50.204	1.161
1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	13.500	100.000	0.135

Table S1. Rankings of the 10 recipes with the highest power density per catalyst cost, in the order of power density, in comparison with pure Pd (at the bottom row). The catalyst cost is normalized by the pure Pd catalyst and referenced in this table¹⁸. The ratio in the table is the maximum power divided by the catalyst cost.

Sample	Lattice parameter (Å)	Cell volume(Å ³)	Rp	Rwp
Pd	3.89643	59.156	3.44%	5.17 %
HEA-1	3.89787	59.222	3.58 %	4.97 %
HEA-2	3.89942	59.293	3.73 %	5.15 %

Table S2. Results for the lattice constants from the XRD analysis. The XRD patterns were analyzed using Rietveld¹⁵ or Le Bail¹⁶ methods in TOPAS software¹⁷. The Le Bail method was employed for the carbon paper electrode due to its broader diffraction peaks and partially disordered structure, while the Rietveld method was used for the catalyst to ensure accurate structural determination. The catalyst crystallizes in the Fm-3m space group, whereas the carbon paper electrode adopts the P6₃/mmc space group. The catalyst HEA-1 and HEA-2 represent the chemistry $\text{Pd}_{0.487}\text{Pt}_{0.185}\text{Cu}_{0.018}\text{Ir}_{0.037}\text{Ce}_{0.106}\text{Nb}_{0.168}$ and $\text{Pd}_{0.381}\text{Pt}_{0.080}\text{Cu}_{0.009}\text{Au}_{0.004}\text{Ir}_{0.02}\text{Ce}_{0.086}\text{Nb}_{0.338}\text{Cr}_{0.082}$, respectively.

Sample	Shell	CN ^a	R (Å) ^b	σ ² (Å ²) ^c	ΔE ₀ (eV) ^d	K-range (Å ⁻¹)	R-range (Å)	R factor
Pd <i>k</i> -edge								
Pd foil	Pd-Pd	12*	2.740±0.002	0.0054±0.0003	3.0±0.4	2.9-11.6	1.0-3.0	0.0035
PdO	Pd-O	4.0±0.3	2.021±0.007	0.0013±0.0009	-2.0±0.8	2.0-12.0	1.0-3.7	0.0074
	Pd-Pd	3.4±0.8	3.022±0.012	0.0033±0.0016	-8.2±1.5			
	Pd-Pd	7.5±1.8	3.414±0.010					
−0.524V	Pd-Pt	3.5±0.6	2.737±0.017	0.0054±0.0010	1.9±0.7	2.5-12.0	1.0-3.0	0.0103
	Pd-Pd	6.1±0.7	2.740±0.008					
OCV	Pd-Pt	3.5±0.4	2.742±0.012	0.0055±0.0007	1.3±0.5	2.5-12.0	1.0-3.0	0.0056
	Pd-Pd	5.8±0.5	2.746±0.006					
Pt <i>L3</i> -edge								
Pt foil	Pt-Pt	12*	2.763±0.003	0.0046±0.0003	7.4±0.4	2.5-12.0	1.3-3.2	0.0046
PtO ₂	Pt-O	6.0±0.5	2.022±0.008	0.0018±0.0009	11.2±1.1	3.0-14.0	1.2-3.6	0.0118
	Pt-Pt	8.6±1.8	3.103±0.005	0.0036±0.0008	8.8±1.1			
	Pt-O	9.4±2.1	3.648±0.025					
−0.524 V	Pt-Pt	4.3±0.4	2.720±0.009	0.0044±0.0010	4.9±0.7	3.0-12.0	1.0-3.2	0.0083
	Pt-Pd	4.0±0.5	2.725±0.007					
OCV	Pt-Pt	4.4±0.4	2.726±0.011	0.0043±0.0012	5.2±0.8	3.0-12.0	1.4-3.2	0.0079
	Pt-Pd	3.9±0.5	2.727±0.009					

Table S3. EXAFS fitting parameters at the Pd *K*-edge & Pt *L3*-edge for various samples. ^aCN, coordination number; ^bR, the distance to the neighboring atom; ^c σ^2 , Debye-Waller factor, the Mean Square Relative Displacement (MSRD); ^d ΔE_0 , inner potential correction; R factor indicates the goodness of the fit. S_0^2 was fixed to 0.825 and 0.803 respectively, according to the experimental EXAFS fit of Pd foil and Pt foil by fixing CN as the known crystallographic value. *This value was fixed during EXAFS fitting, based on the known structure of Pd and Pt. Error bounds that characterize the structural parameters obtained by EXAFS spectroscopy were estimated as CN ± 20%; R ± 1%; σ^2 ± 20%; ΔE_0 ± 20%. A reasonable range of EXAFS fitting parameters: $0.700 < S_0^2 < 1.000$; CN > 0; $\sigma^2 > 0 \text{ Å}^2$; $|\Delta E_0| < 15 \text{ eV}$; R factor < 0.02.

Appendix

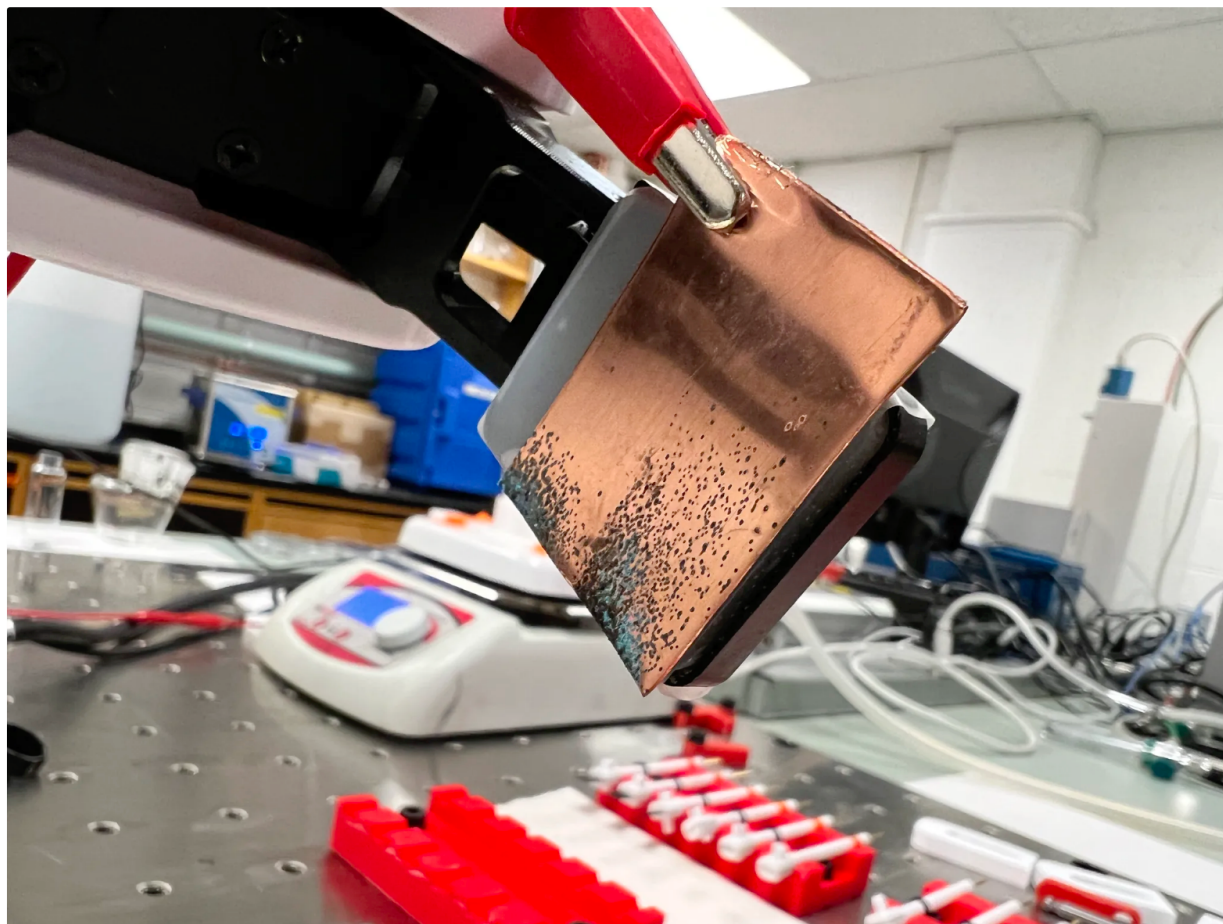
Example question-answer set for evaluation of the vision language model (VLM) to debug real-world materials science experiments. More examples could be found in li.mit.edu/cr/.

General description:

You are an expert in materials science and electrochemistry experiments. For each question, you will be shown one or a few pictures of real experimental scenarios and/or experimental data collected under that experimental scenario. You are supposed to point out (1) if there is any issue with the experimental setup that you have observed, and if there is, (2) what is the issue, and (3) how to properly address it.

Question 1:

This is a piece of thick copper foil that is used as a current collector, which will be in contact with the electrode sample holder for electrochemical testing. The electrolyte contains 1.0 M KOH. When performing electrochemical test, the copper foil stays around a few centimeters above the electrolyte.



Answer by OpenAI o3 model:

Step	Observation	Why it is a problem	How to fix / best practice
(1) Foil material: bare Cu in strongly alkaline environment	Dark, patchy film already visible \Rightarrow $\text{Cu}_2\text{O}/\text{CuO}$ growth. Even a few centimeters above the meniscus, KOH vapor, splash and	<i>Chemical instability</i> Cu oxidizes to $\text{Cu}_2\text{O}/\text{CuO}$ and eventually dissolves as $\text{Cu}(\text{OH})_4^{2-}$. The mixed oxide layer raises contact resistance \rightarrow drifting iR-drop and noisy data flakes/dissolves $\rightarrow \text{Cu}^{2+}$	Replace or isolate the copper. Use Ni sheet, Ni-plated Cu, Ti, 316 SS, or Au-plated Cu for anything within “spray distance” of the electrolyte. If Cu must

Step	Observation	Why it is a problem	How to fix / best practice
	capillary creep along the sample holder wet the foil.	can redeposit on/poison your working electrode changes geometry with time → irreproducible current distribution	stay (thermal bus bar): electroless-plate $\geq 5\ \mu\text{m}$ Ni or $\geq 1\ \mu\text{m}$ Au on all faces likely to be exposed to alkaline mist. Cover non-contact areas with chemically-resistant heat-shrink or epoxy to stop wicking.
(2) Mechanical contact: small crocodile clip on oxidized foil	Limited, uneven pressure on a surface that is already partly oxidized.	<i>Electrical unreliability</i> High and variable contact resistance; localized heating at high current; possible arcing if contact is intermittent.	Provide a clean, broad, stable interface. Wet-polish the Cu (or the new Ni/Au surface) immediately before assembly. Use a spring-loaded flat clamp, screw post, or gold-plated leaf spring to achieve uniform pressure over $\geq 1\ \text{cm}^2$. Verify $< 2\ \text{m}\Omega$ with a four-wire meter.
(3) Stray Cu contamination path to the cell	Foil and clip are only “a few cm” above the bath; KOH splashes/aerosol are common during degassing, stirring, or gas evolution.	<i>Electrolyte cross-contamination</i> Even trace Cu^{2+} can electrodeposit on catalytic surfaces—altering kinetics and masking real degradation/stability trends.	Physically segregate current bus from electrolyte space. Mount bus bar higher or behind a PTFE splash shield. Vent cell with a lid or baffle that stops alkaline droplets from reaching the collector. Rinse and dry current collector after every run.

Step	Observation	Why it is a problem	How to fix / best practice
(4) Long-term reproducibility	Progressive oxide growth is visible already. Over weeks/months each test starts with an increasingly resistive, rougher surface.	Drift in “baseline” ohmic drop; harder to compare datasets taken days apart.	Implement a maintenance schedule. If you insist on Cu: repolish or replace after a fixed number of runs. Better: switch to Ni or Ti—forms a self-passivating layer that stabilizes quickly and then stays constant.

Quick checklist for a robust alkaline-cell current collector

1. **Material stable in 1 M KOH:** Ni \geq 99 %, 316 SS, Ti, graphite, or Au-plated Cu.
2. **Large, flat, polished contact area** secured by spring clamp or bolted lug.
3. **Isolation from electrolyte spray:** PTFE/PVDF shield, heat-shrink on leads.
4. **Measure contact resistance** before each campaign; target \lesssim 2 m Ω .
5. **Document maintenance:** date of polishing/plating, measured resistance, photo log of surface condition.

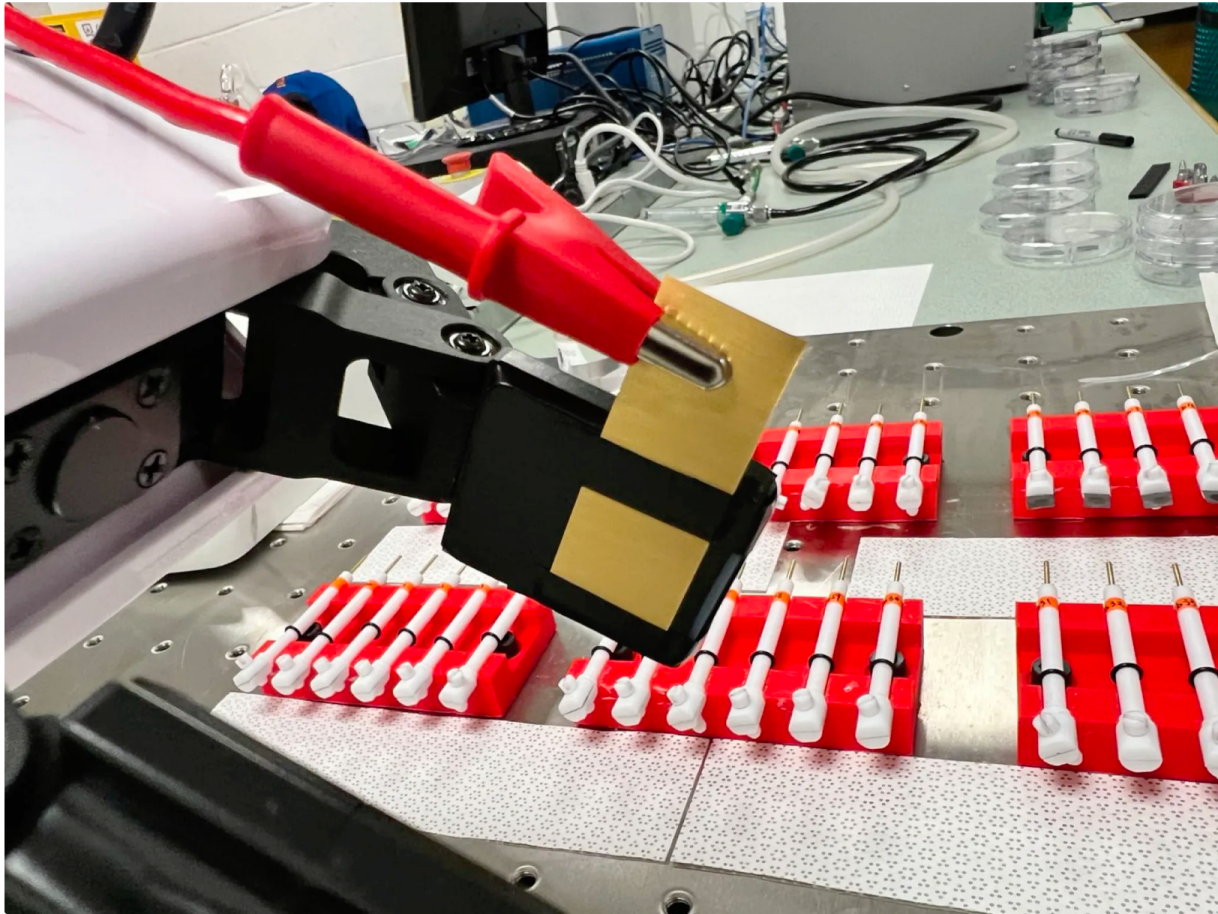
Implementing these measures eliminates Cu dissolution and contact drift, prevents copper contamination of your catalyst, and yields stable, reproducible electrochemical data in alkaline media.

Human Rating: Correct

Human Comments on the AI answer:

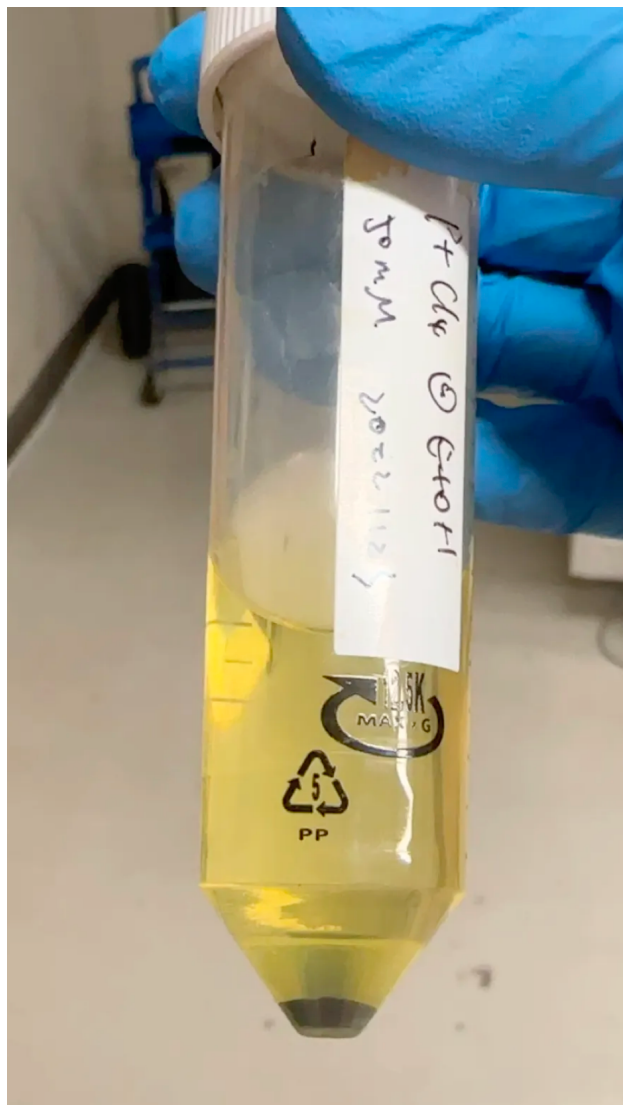
The AI clearly observes the ‘Dark, patchy film already visible’ on the copper foil by just observing the image. The AI reasons about potential issues such as KOH vapor/splashing/etc. (even though the copper foil is not in direct contact with the electrolyte), which is exactly what a human expert has thought. Further, the AI proposes to use materials such as Au foil, which is again exactly what we used in lab for practice, as could be seen below. **Please note that, when asking the AI the question, we only showed it the picture of the Copper foil. We did not tell the AI the copper is corroded (this implies that AI already knows what metal corrosion is, and importantly, what metal corrosion looks like), nor did we show it the picture of the Gold foil, which was**

proposed by a human expert as a strategy to solve the problem (this implies that the AI has the capability to solve such real experimental issues).



Question 2:

This is a bottle of 50 mM of PtCl_4 solution in ethanol. Initially, all the salts were dissolved. This is what we have after around a month.



Answer by OpenAI o3 model:

#	Observation	Underlying cause	Consequences for experiments	Corrective actions / best practice
1	Dark, metallic-looking sediment at the conical tip (Pt “black”)	Ethanol is a mild reducing agent. Over time it reduces $\text{Pt(IV)} \rightarrow \text{Pt(0)}$, producing elemental Pt nanoparticles that	Actual Pt(IV) concentration is now unknown—dosing becomes inaccurate. Colloidal Pt can seed	→ Do not store PtCl_4 in alcohols. Prepare a fresh solution immediately before use.→ If long-term storage is essential,

#	Observation	Underlying cause	Consequences for experiments	Corrective actions / best practice
		agglomerate and settle. Trace water and light accelerate the process.	unwanted nucleation on electrodes or glassware. • Solution composition drifts between replicates.	dissolve $\text{PtCl}_4 \cdot x\text{H}_2\text{O}$ in 0.1–1 M HCl(aq) , keep in an amber glass bottle under N_2 at 4 °C. The acidic chloride matrix suppresses reduction.
2	Yellow color has lightened / changed vs. fresh stock	Partial hydrolysis / ligand exchange: $\text{PtCl}_4 + \text{EtOH} + \text{H}_2\text{O} \rightleftharpoons [\text{PtCl}_4(\text{EtOH})_2] + \text{HCl}$, then reduction to Pt(II) and Pt(0).	Optical absorption no longer correlates with Pt(IV) concentration; any spectroscopic calibration is invalid.	→ For organometallic routes that require an organic solvent, use a freshly prepared $\text{Pt}(\text{acac})_2$ or H_2PtCl_6 solution in ethanol/ethylene glycol right before the synthesis step.
3	Polypropylene centrifuge tube (PP) used as container	PP is slightly permeable to $\text{O}_2/\text{H}_2\text{O}$ vapor and not completely light-blocking.	Oxidative and photochemical pathways speed up Pt(IV) reduction; Cl^- can leach tiny amounts of impurities from PP over time.	→ Store precious metal precursors in amber, air-tight glass vials (borosilicate or Schott), sealed with PTFE-lined caps, kept in the dark.
4	One-month storage without inert atmosphere	Dissolved O_2 assists ethanol oxidation (forming acetaldehyde/acetate) which in turn reduces Pt(IV).	Faster degradation than expected; may complete in days at room T.	→ Degas solution with N_2/Ar if you must keep it a few days; but the safer route is “make-fresh, use-immediately”.

Human Rating: Correct

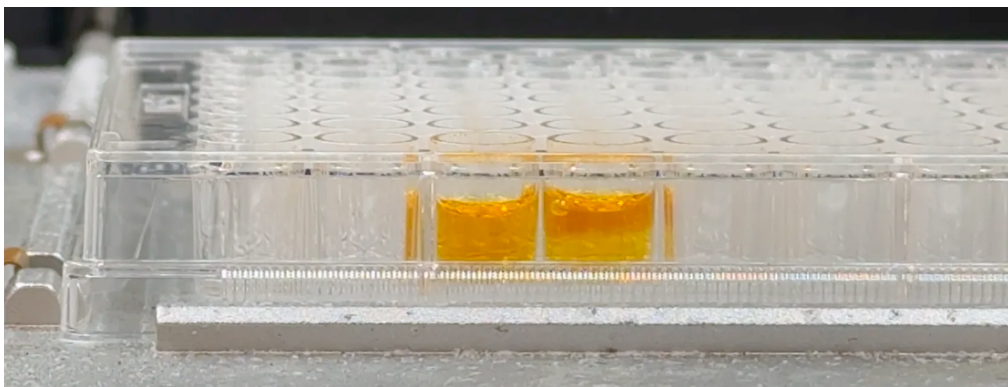
Human Comments on the AI answer:

The AI clearly observes the ‘Dark, metallic-looking sediment at the conical tip (Pt black)’ and attributes this to the slow reduction of Pt(IV) by ethanol. Importantly, it proposes a practical

strategy which is to add HCl to the solution or use H_2PtCl_6 , which is exactly what we do in the lab. In fact, with 50 mM H_2PtCl_6 ethanol solution, we did not observe any sediment even for a year.

Question 3:

This a well-plate for liquid mixing. In each of the well, we mix two metal precursors (both 50 mM metal chloride solution in ethanol, one of them is yellowish color, and the other is dark yellowish color). For the left and right well, we use two mixing protocols, and this is what is looks like after mixing.



Answer by OpenAI o3 model:

Feature	Left-hand well (Protocol 1)	Right-hand well (Protocol 2)	What this tells us
Vertical color profile	Very sharp, horizontal boundary. Bottom $\approx 60\%$ of the column is deep orange; the upper layer is noticeably paler.	Boundary is curved and diffuse; orange tone reaches $\geq 80\%$ of the height, only a thin light band remains on top.	P 2 induced more fluid motion (better but still partial mixing). P 1 left the two liquids essentially stacked.
Interface geometry	Flat meniscus \rightarrow liquid B was deposited gently from above; flow stopped almost instantly.	Interface tilts upward toward the right wall \rightarrow dispense jet entered deeper or at an angle, producing a single convection roll.	Tip depth / dispense trajectory differ.
Time needed to reach homogeneity (qualitative)	Minutes-to-hours (diffusion-limited).	Tens of seconds (weak natural convection + diffusion).	Neither protocol delivers “instant mix” that high-throughput assays assume.
Risk for screening accuracy	High – concentration at the analysis zone depends on time and sampling height;	Better, but early-time kinetics still drift; well-	A dedicated mixing step is required for both.

Feature	Left-hand well (Protocol 1)	Right-hand well (Protocol 2)	What this tells us
	early measurements will vary widely.	to-well RSD will remain significant.	

Why both are inadequate

- **Ethanollic metal-chloride solutions differ in density and viscosity** once their chloride/ligand coordination spheres diverge; sequential addition therefore creates stable stratification.
- At 20 μL , the Reynolds number of the dispense jet is low; unless the tip moves or re-aspirates, flow stays laminar and dies out within a few milliseconds.
- Result: concentration in the analysis zone (bottom 0.5–1 mm) changes with time and wall position, so any color-metric, spectroscopic, or nanoparticle-nucleation read-out will be history-dependent.

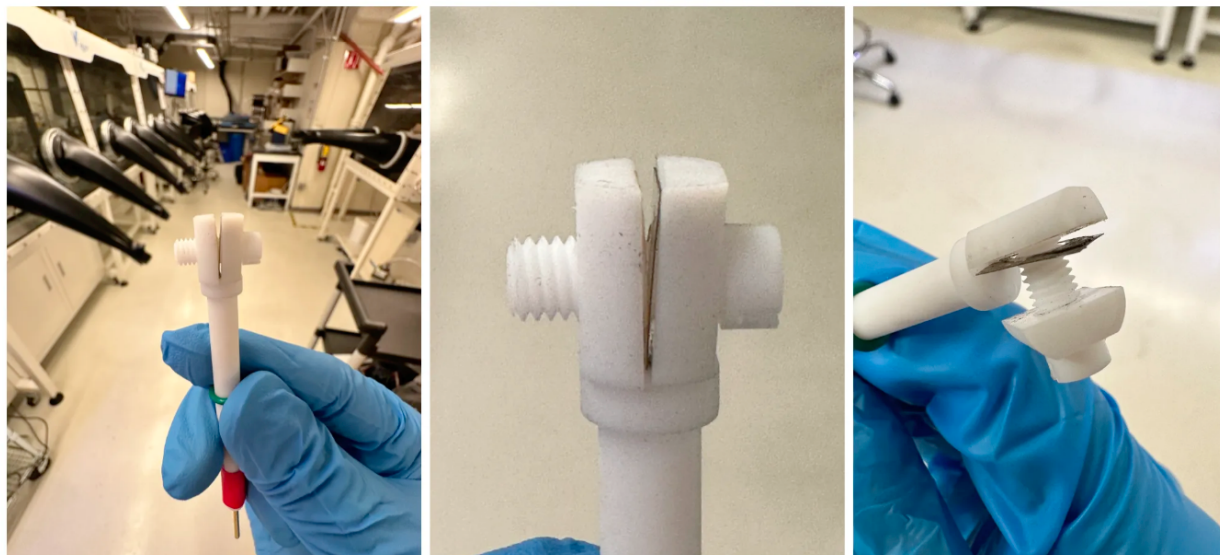
Human Rating: Wrong

Human Comments on the AI answer:

For this question, even though trying multiple times, the AI failed to recognize that the left-hand well had a more uniform mixing of solution and thus a better protocol, compared with the right-hand well in the picture. However, one encouraging thing is that the AI mentioned the concept of ‘molecular diffusion’, which is a rather slow process (minutes-to-hours), and this is exactly why a vigorous mixing protocol was required to achieve uniform mixing and avoid layering of liquids. This means that if we use the liquid handling pipette to gently dispense some liquid into/onto a bottom layer of liquid, it probably takes a long time for them to mix well under static conditions at room temperature.

Question 4:

The pictures below are electrode sample holders. When doing electrochemical experiments, typically, a carbon paper sample will be clamped between the metal foil in the middle and the white PTFE body. For a new sample holder, the metal foil looks like the picture in the middle. For a long-term used sample holder, it looks like the picture on the right. Sometimes, the performance is normal, but sometimes, there is no performance at all.



Answer by OpenAI o3 model:

Photo label	What has happened to the holder	Why this gives “sometimes works / sometimes dead” readings
Middle panel (new) <i>narrow slot, parallel jaws, clean foil</i>	PTFE jaws still square; screw threads are crisp → screw can press the metal foil tightly against the carbon paper.	Low, stable contact resistance ($\lesssim 20\text{--}30\text{ m}\Omega$) → cell behaves normally.
Right panel (after months of use) <i>slot flared open, foil edge bent, PTFE surface scarred</i>	1. Cold-flow / creep of PTFE: every tightening cycle plastically deforms the soft polymer; jaws spread, the slot widens.2. Thread wears in PTFE: set-screw bottoms before it generates real preload.3. Foil edge work-hardens & oxidizes loses flatness; oxide layer adds $\geq 100\text{ m}\Omega$.	<ul style="list-style-type: none">• Some assemblies still make a thin line of contact → “normal” run.• Slight misalignment or fresh oxide film breaks the circuit → essentially open cell (“no performance”).• Behavior flips run-to-run because clamping pressure is now below the threshold needed for a reliable metallic junction.

How to fix the existing holders

1. Restore clamping force

- Slip two 0.1-0.2 mm stainless or Ni shims into the slot faces so the screw compresses metal, not PTFE.
- Add a Belleville (spring) washer under the screw head to keep constant preload while the polymer creeps.

2. Renew the current-collector interface

- Replace or re-polish the foil each day; if Cu/Ni is used, plate 1-2 μm Au to prevent oxide build-up.
- Wipe the carbon paper contact zone with IPA and blow dry just before clamping.

3. Verify the joint before immersing

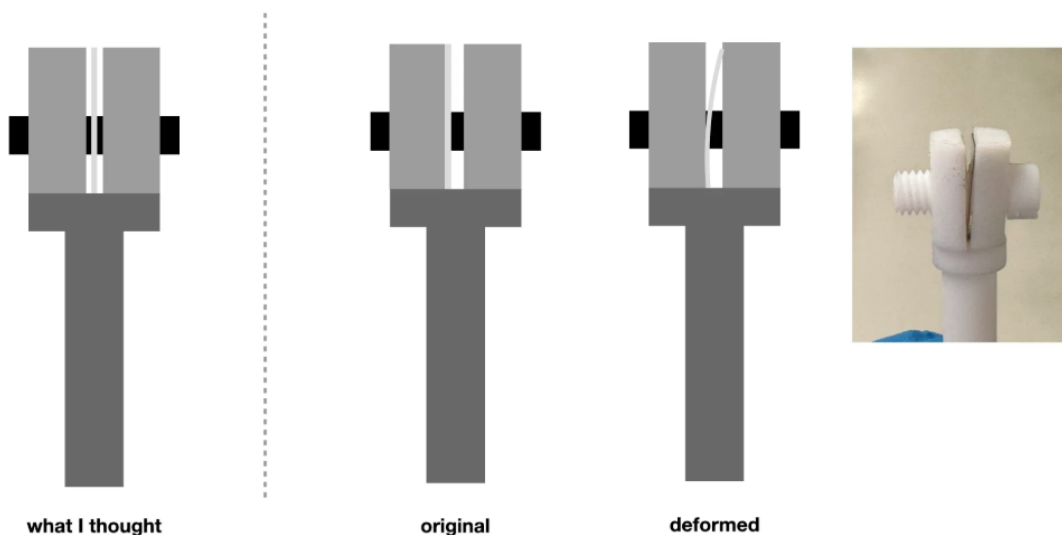
- Four-wire measure resistance from the external lead to the working-electrode surface; target $\leq 30\text{ m}\Omega$.
- Re-tighten or rebuild if higher.

Human Rating: Wrong

Human Comments on the AI answer:

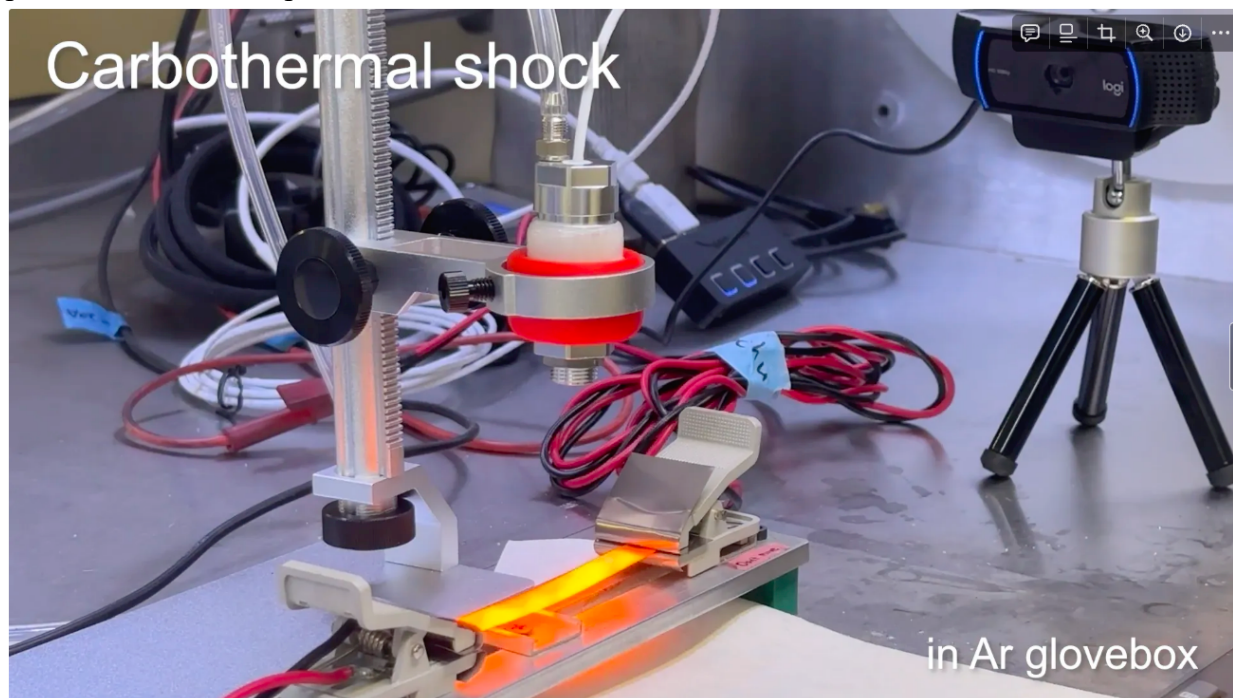
There seems to be some serious misunderstanding of these pictures from the AI. Maybe because the picture in the middle and the picture on the right was taken from different camera view, the AI thinks that the white PTFE polymer was deformed, and thus attributed this issue to a mechanical problem.

In fact, the truth is, for this batch of sample holders (with Titanium metal foil in the middle), the two sides of the Ti foil are different (see schematics below). Only one side is conductive, while the other side is coated with non-conductive adhesive tape. However, after long-term use (in acid or base), the adhesion gradually fails. Thus, if the sample is unfortunately clamped at the wrong side of the Ti foil, there might be no or little current response, which will lead to misleading experimental results, especially for an unknown material.



Question 5:

The picture below is for the carbothermal shock treatment of a carbon paper. A high current pulse is used to pass the carbon paper, and heat the sample up to thousands of Kelvins within tens of milliseconds. For the sample preparation recorded by this picture, it was found that the samples had different electrochemical performance at different location on the carbon strip. Is there any potential issue in this picture?



Answer by OpenAI o3 model:

The photo reveals a pronounced temperature (and therefore treatment) gradient along the strip that stems from the way the sample is clamped and cooled.

What the picture shows	Why it matters	Resulting non-uniformity	How to fix / verify
Large metal binder clips on both ends act as massive heat-sinks. Their jaws are in direct contact with (and even partly cover) the carbon strip.	During the 1- to 10-ms high-current pulse, the clamps rapidly suck heat away from the regions they touch. Meanwhile the free span in the middle has no cooling except radiation → it reaches the highest temperature.	Centre of the strip hits the target 1500–2 000 K; zones a few mm from each clamp may be hundreds of kelvin cooler. Catalytic phase, porosity, dopant distribution, and defect density therefore vary with position → different	① Thermally isolate the clamps: place 0.5 mm thick Al_2O_3 or BN washers between carbon and jaws, or use Mo/NiCr jaw tips only 1–2 mm deep so they touch the strip minimally. ② Use matched, water-cooled bus bars to keep contact temperature constant and

What the picture shows	Why it matters	Resulting non-uniformity	How to fix / verify
		electrochemical activity along the strip.	symmetric.③ Cut out only the central, uniformly heated 5–10 mm segment for electrochemical testing.
Unequal electrical contact pressure: binder clips rarely press exactly the same on left and right; oxidation builds up on one jaw sooner than the other.	Different contact resistances → slightly different current pathways → current crowds toward the lower-resistance end, amplifying the thermal gradient.	One half of the span can be noticeably hotter (or cooler) than the other; catalytic performance then varies left-to-right.	After each run measure the 4-wire resistance of the clamped strip ($< 10\text{ m}\Omega$; $\leq 1\text{ m}\Omega$ difference between ends). Re-polish or replace jaws if higher.
White paper sheet under the fixture	Even in Ar, paper pyrolyzes $\geq 550\text{ }^{\circ}\text{C}$. It locally absorbs radiant heat and releases volatiles; the volatile flow and additional convective cooling are stronger in the center than near the clamps.	Further accentuates the center–edge temperature gap and can introduce C/O/H impurities that change surface chemistry.	Replace paper with ceramic tile or leave the bench bare metal.
Argon nozzle aimed at the middle of the strip (white/red fitting)	Gas jet cools the center more than the ends if not aligned symmetrically or if the flow is turbulent only on one side.	Adds yet another spatially varying heat-loss term.	Either remove the jet (whole glove-box is already Ar) or use a low, laminar side-to-side sweep that covers the entire span uniformly.
No in-situ temperature mapping	Visual glow cannot resolve $\pm 100\text{ K}$ differences; you only discover the gradient when the	Hard to correlate structure–performance relationships.	Mount a two-color pyrometer or IR micro-camera perpendicular to the strip to capture the

What the picture shows	Why it matters	Resulting non-uniformity	How to fix / verify
	electrocatalytic data scatter.		temperature profile during every pulse.

Quick, low-cost mitigation

1. **Slip a narrow (1–2 mm) Mo shim between each clamp face and carbon** so only a small area is heat-sunk, keeping the hot zone more uniform.
2. **Pre-mark and cut out the central 5–8 mm after the shock**, discarding the cooler ends.
3. **Log the contact resistance** before pulsing; re-seat clamps if $> 10 \text{ m}\Omega$ imbalance is seen.
4. **Remove combustible paper pad** and use an alumina slab instead.

Human Rating: Correct

Human Comments on the AI answer:

The AI clearly observes ‘pronounced temperature (and therefore treatment) gradient along the strip’, which is the most important issue of this experiment. The AI further mentions that to ‘pre-mark and cut out the central 5–8 mm after the shock, discarding the cooler ends’, which is exactly what we did in practice. It is in general challenging to ensure uniform heating with this method, and thus we only keep the samples in the middle of the carbon paper, which reaches our target temperature.

Question 6:

The picture below is for using the robotic arm to unload a sample holder back to the red sample stage. This occurs when the robot should have finished unloading the sample, and then leave. The original contact of the gripper (with the sample holder) is silicone pad.



Answer by OpenAI o3 model:

Observation in the frame	Why it causes trouble	Symptom you see (“robot should have left but can’t”)	Remedies / best practice
Silicone pads on the gripper jaws still press against the	<ul style="list-style-type: none">• Static friction rises above the few newtons the gripper	Robot times out, drops the part, or displaces nearby	<i>Swap the pad material → use low-durometer FKM/Viton® or</i>

Observation in the frame	Why it causes trouble	Symptom you see (“robot should have left but can’t”)	Remedies / best practice
<p>PTFE stem when the robot tries to retreat. Silicone is compliant and slightly tacky; after a few seconds under clamping force, it “cold-welds” to smooth PTFE.</p>	<p>releases when it opens. • The holder stays stuck to one pad, so when the arm moves up it lifts the holder out of its nest or drags it across the rack.</p>	<p>samples; sometimes crashes the next motion plan.</p>	<p>polyurethane (Shore A 80–90) pads; coefficient of friction on PTFE < 0.3. <i>Texture the pad surface</i> with shallow (≈ 0.2 mm) cross-hatch grooves so only the ridges contact, cutting the real contact area. <i>Add a release pulse:</i> program the gripper to open + retract 0.5 mm, pause 200 ms, then jog 1 mm up; the micro-jog breaks static adhesion.</p>
<p>No positive mechanical datum between holder and red stage. The holder is resting in a V-groove created by the white forks; any small side load when the pad “lets go” can tip it.</p>	<p>Position repeatability depends entirely on the robot’s placement accuracy and friction between PTFE and stage.</p>	<p>Samples sit at different angles → later pick-up fails or electrochemical cell alignment drifts.</p>	<p>Add a shallow pocket or magnet pin on the stage that mates with a notch/sleeve on the holder tip, so it snaps into a repeatable seat once the robot releases.</p>
<p>Red flying lead routed over the wrist & gripper. During unload it can brush the holder and apply a random side force; it also limits the gripper’s opening stroke.</p>	<p>Unpredictable extra force re-engages the tacky pad > holder sticks again.</p>	<p>Intermittent failures: some cycles clean, some stick.</p>	<p>Route the lead along the upper arm with loose spirals, terminate in a lightweight pigtail that clears the gripper’s work envelope by at least 30 mm.</p>

Observation in the frame	Why it causes trouble	Symptom you see (“robot should have left but can’t”)	Remedies / best practice
No force/torque sensor feedback for ‘release confirmed’. Sequence relies on time-delay assumption that the jaws are empty.	When stick-slip occurs the controller never notices; it immediately issues the next move with the holder still attached.	Collision or dropped part.	Fit an inline capacitive or strain-gauge sensor on the gripper or use the robot’s joint-torque data; abort if measured weight after opening $\neq 0$ N.

Quick test to confirm the root cause

1. **Single-step manual run:** open jaws fully and wiggle the arm ± 0.5 mm laterally; if the holder follows the pad, adhesion is to blame.
2. **Swap one silicone pad for smooth Kapton tape** and repeat; if sticking disappears the material is the issue.

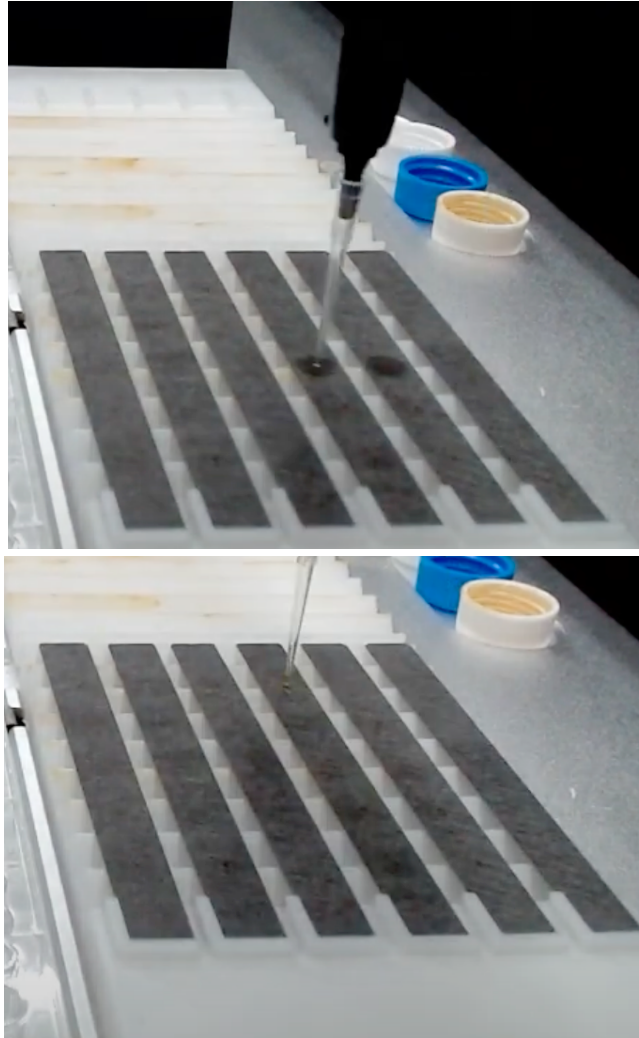
Human Rating: Correct

Human Comments on the AI answer:

The AI answers this question correctly. The AI first realizes that the silicone pad could be too adhesive and lead to cold-welding when clamping the sample holder with force. Then it suggests swapping with a tape, which is exactly what we did in practice. A minor point is that the tape has to be conductive because it will be in contact with the sample holder, but in general, the AI answers in a good way.

Question 7:

The two pictures below are for using the liquid handling robot to dispense the same amount of liquid on to a carbon strip (190 μm thick) at different locations. Both pictures were taken immediately after dispensing the liquid. Please only look at the fourth carbon strip from the left.



Answer by OpenAI o3 model:

Frame	What you can see immediately after dispensing	What it tells you	Root cause
Picture A (upper panel)	A large, almost circular dark patch; a second, paler halo a few mm downstream.	The liquid has already wicked deep into (and partly under) the	The pipette tip actually touched / compressed the strip. The pressure jet and mechanical squeeze opened the fiber network and forced liquid

Frame	What you can see immediately after dispensing	What it tells you	Root cause
		carbon paper, then bled out again.	through to the back side and into the slot beneath.
Picture B (lower panel)	Only a small, faint mark; the rest of the strip is still dry.	Most of the droplet is sitting on top of the strip, still forming a bead. Little fluid has penetrated.	Here the tip stopped ≈ 1 mm higher, so the droplet fell gently; capillary uptake is slow and limited to the immediate landing zone.

Because the carbon paper is only $190\text{ }\mu\text{m}$ thick and highly porous, **a difference of $<0.5\text{ mm}$ in the Z-height or in clamp deflection is enough to switch between “impact-injection” and “gentle deposition.”** That is why you get two completely different wetting footprints with what should be identical doses.

How to fix it

- 1. Hard stop for the pipette Z-axis.**

*Zero the robot to the plastic stage, **not** to the strip surface; dispense from a fixed height (e.g. 1.0 mm) everywhere.*

- 2. Keep the carbon perfectly flat.**

Insert a thin ($\leq 100\text{ }\mu\text{m}$) rigid backing plate under each strip or apply light vacuum under the slots to pull the paper flush.

- 3. Slow, low-momentum dispense.**

Cut the dispense speed to $\leq 3\text{ }\mu\text{L s}^{-1}$ and pause 300 ms before retracting so the droplet relaxes instead of splashing.

- 4. Add a vision / laser height check.**

Measure the actual strip height at each lane once per plate; apply a Z-offset if a strip sits high or low.

- 5. Use a wetting pre-treatment if needed.**

If the liquid is very hydrophobic on the paper, mist a tiny amount of the solvent first so all shots wet similarly.

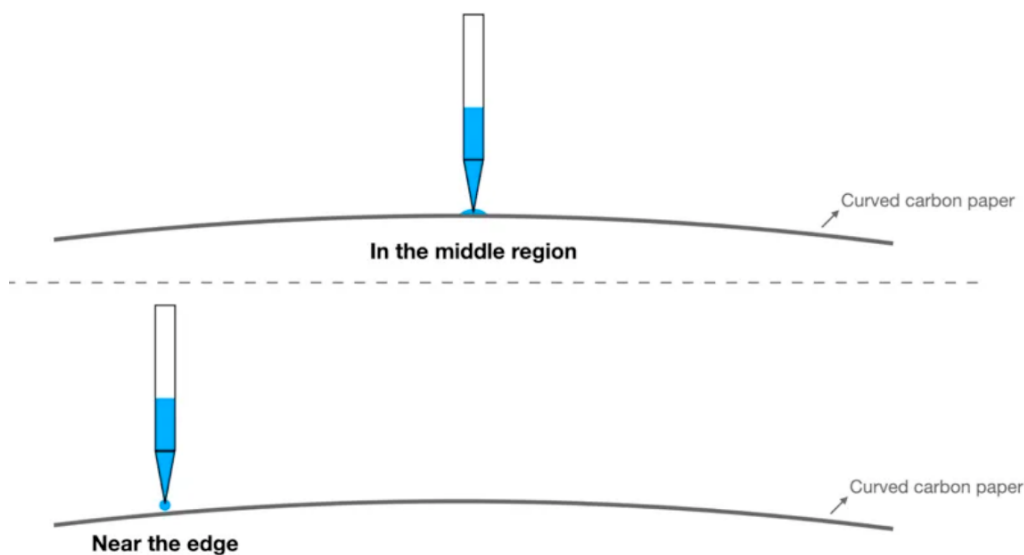
With a controlled stand-off distance and a flat, supported substrate every dispense will land in the same way, giving identical wet-spots and, more importantly, reliable electrochemical performance across the whole strip.

Human Rating: Correct

Human Comments on the AI answer:

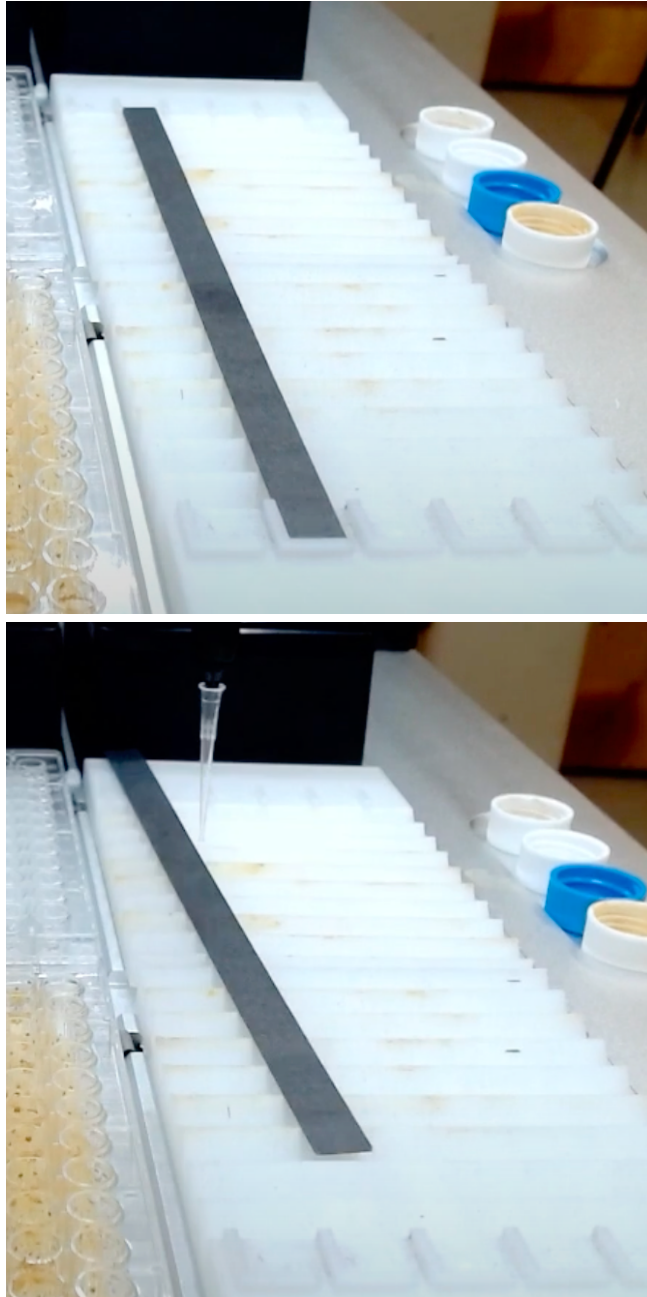
First, the AI could clearly observe that there is strong difference between the wetting behavior in the two pictures. For the upper panel, the AI mentions ‘a large, almost circular dark patch’, while for the lower panel, the AI mentions ‘only a small, faint mark; the rest of the strip is still dry’. Importantly, for the lower panel picture, the AI realizes ‘here the tip stopped ≈ 1 mm higher’. The AI further emphasizes the importance of the carbon height as in ‘Because the carbon paper is only $190\text{ }\mu\text{m}$ thick and highly porous, a difference of $<0.5\text{ mm}$ in the Z-height or in clamp deflection is enough to switch between impact-injection and gentle deposition’. For a strategy to solve this, the AI mentions to ‘keep the carbon perfectly flat’.

After further analysis, we attribute this problem to the ‘curvature’ issue of thin carbon paper (which leads to the height difference between the end and the middle part of the carbon strip), but this issue is much more relieved for a thick carbon substrate, which is what we generally use. If we insist on using thin carbon papers, then we need to ensure the carbon paper to be perfectly flat, as the AI suggested.



Question 8:

The two pictures below are for using the liquid handling robot to dispense a certain amount of liquid on a carbon strip at different locations. The top image was taken before the second image. What has happened and why?



Answer by OpenAI o3 model:

Step	Evidence in the pictures	What actually happened	Underlying reason
Before dispensing (upper photo)	The long carbon strip sits neatly in its lane; its lower end stops flush with the last pocket of the white plastic jig. No pipette is present.	Initial, correct placement.	
During/after dispensing (lower photo)	A pipette tip has descended; the whole strip is now shifted ~1 lane lower ($\approx 6\text{--}7\text{ mm}$) compared with its original position (overlay shows a clear offset).	The robot's tip <i>physically hit</i> the strip, pushing/sliding it down the tray while trying to deliver the droplet.	1. Z-height too low or collision when probing the surface. 2. Strip is only $\sim 190\text{ }\mu\text{m}$ thick and lies loose—very little friction holds it, so a light touch from the tip is enough to displace it. 3. Any solvent already on the jig acts as a lubricant, further lowering friction.

Because the strip is now mis-registered, all subsequent dispense or pick-up operations referenced to the original coordinates will be off, leading to uneven wetting, missed spots, or mechanical jams.

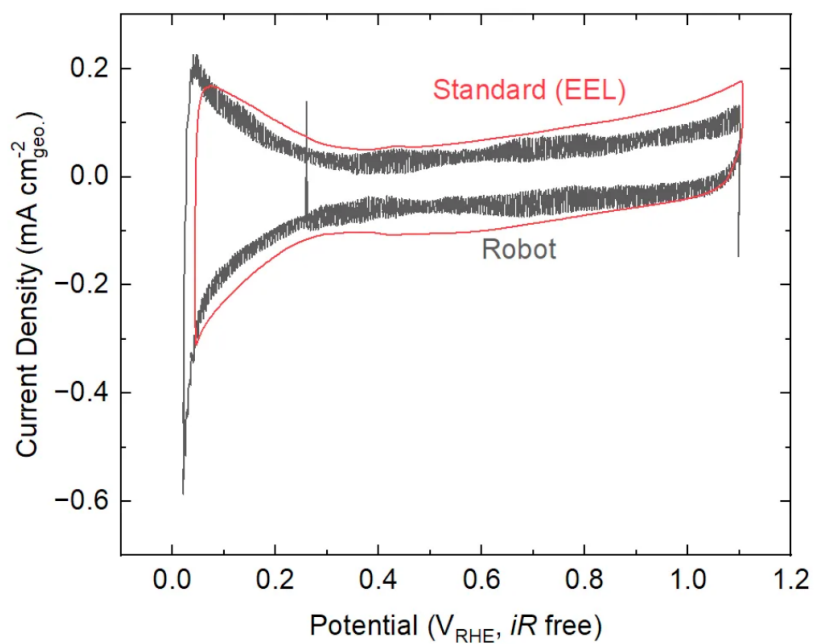
Human Rating: Correct

Human Comments on the AI answer:

For this question, we find that the AI is able to call tools such as ‘measuring distance in an x-y coordinate system’, and thus it successfully found that the carbon paper in the lower panel picture is clearly displaced. The AI further mentions that there may be collision due to Z-height too low, which is exactly correct.

Question 9:

The first picture below shows the electric wiring that we use for electrochemical tests. The second picture shows the data collected with this setup, compared with the standard test.



Answer by OpenAI o3 model:

What I see in the wiring photo	What that does electrically	Effect on the CV trace (grey)	How to cure it
Switch-mode wall wart (black) sharing the same duplex outlet as the potentiostat	SMPS bricks switch at 20-200 kHz; the fast edges couple through the mains and radiate.	High-frequency ripple rides on the current signal → dense “grass” along the whole voltammogram.	Power the potentiostat from a separate, earthed outlet or an isolation transformer ; if the wall wart must stay, add an LC mains filter or move it ≥ 30 cm away.
Unshielded signal leads (yellow, green) draped across mains lines and tubing	Act like antennas; pick up conducted and radiated noise and inject it into the high-impedance reference and sense lines.	Random $\pm 50\text{--}100\ \mu\text{A}$ spikes, most obvious near 0 V where the true current is small.	Route the potentiostat cables separately , twist WE-CE and RE-CE pairs, and keep them at least 5 cm from any 120 V line. Use shielded, twisted-pair patch cords.
No obvious chassis/earth bond on the test cell (green/yellow wire appears to float)	Floating shields invite ground-loop-induced hum and provide no path for common-mode interference to bleed off.	Low-frequency wander and 60 Hz sidebands broaden the baseline.	Bond the potentiostat chassis, test jig and Faraday cage (if present) to a single star-ground tied to the earth pin of the outlet.
Bundle of color wires for pumps/valves running parallel to sense leads	Stepper- or DC-motor drivers generate sharp current spikes; magnetic coupling injects differential noise.	Isolated narrow spikes that vary with pump or valve events.	Run all motor and solenoid wiring in a separate harness and add ferrite beads or RC snubbers at the driver outputs.
Adapters and barrel-plug extensions between wall wart and device	Extra contacts oxidize; intermittent micro-contact causes burst noise.	Occasional tall spikes (seen around 0.15 V in the plot).	Use one solid cable; clean or replace corroded plugs; strain-relieve the lead so nothing moves during a scan.

Quick checklist to reproduce the “standard” smooth baseline

1. **Give the potentiostat its own filtered, earthed outlet** (or isolation transformer).
2. **Separate the harnesses:** high-current 120 V/24 V supply cables on one side of the bench, high-impedance cell leads on the other.
3. **Twist and shield** working-, reference- and counter-electrode leads; connect the shield to earth at the potentiostat end only.
4. **Star-ground** the potentiostat chassis, cell stand and any Faraday cage.
5. **Add ferrite clamps** on every motor/solenoid/pump supply and signal line.
6. After rewiring, run an open-circuit scan: the baseline noise density should drop to the red “Standard” level ($< 1 \mu\text{A cm}^{-2}$ p-p).

Implementing these housekeeping fixes typically improves S/N by an order of magnitude and aligns the robotic system’s electrochemical signatures with those obtained on the dedicated electrochemistry bench.

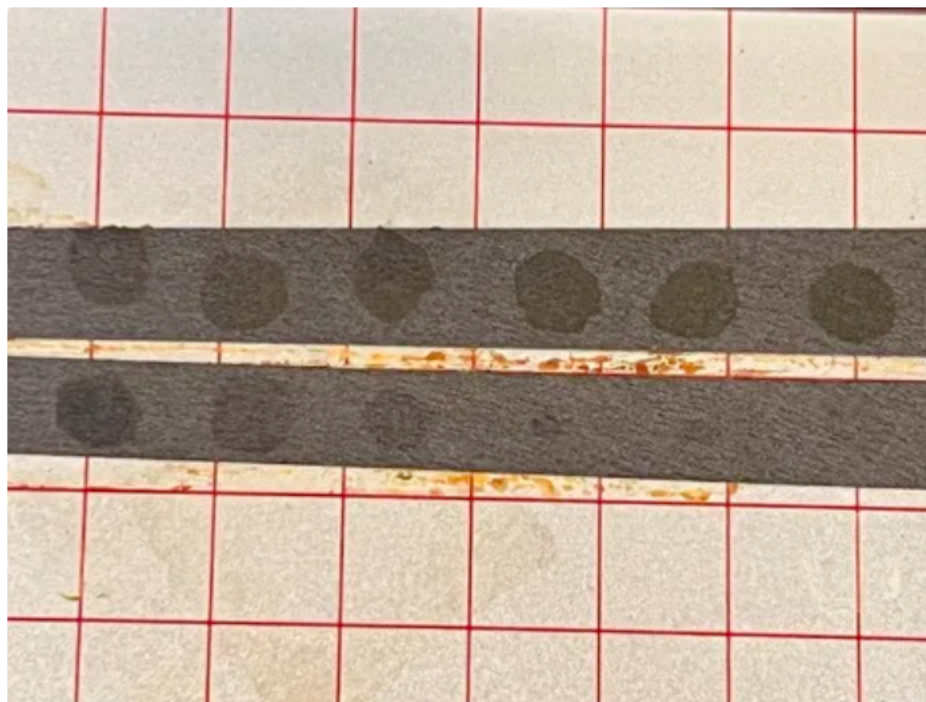
Human Rating: Correct

Human Comments on the AI answer:

This could be a tricky problem for an electrochemist, who is in general not a professional electrician. Encouragingly, the AI clearly realizes a magnetic noise issue coming from the bundle of unshielded wiring and the adapter in the upper panel picture, especially when they are in proximity. In fact, after we separate the wiring away from the power adapter, there is no such current noise issue anymore.

Question 10:

We use a liquid handling robot to dropcast a certain amount of liquid onto the carbon paper. what could be an issue?



Answer by OpenAI o3 model:

Issue spotted (yes, there *is* a problem)

The pale-orange stripe running between the two carbon-paper lanes shows that the liquid you just dispensed has **wicked completely through the 190 μm strip and bled out onto the support plate**. In other words, each drop is *not* staying where the robot put it:

- The dark round spots on top are only part of the story—the rest of the solution has percolated downward and spread laterally under the strip.
- The orange crust (metal-salt residue) is proof that solute is now *below* the working face and even between adjacent lanes, so true loading per spot is lower and poorly controlled.
- If the solute contains Fe, Cu, Ni, etc., those ions can migrate and cross-contaminate neighboring discs or the current-collector foil.

Cause	Explanation
Highly porous, hydrophilic carbon paper (no microporous layer)	Capillaries pull the low-viscosity solvent right through the thickness.

Cause	Explanation
Open slot beneath each lane in the plastic jig	Nothing blocks the liquid once it exits the back side, so it flows onto the tray and along the slot.
Large drop volume / low viscosity solvent	The driving force for through-wicking is stronger than surface tension holding the bead on top.

Human Rating: Correct

Human Comments on the AI answer:

This AI observes the color difference on the sample stage beneath the carbon strips, and attribute this issue to the leakage of the solution from the back side of the carbon strips. To resolve this issue, we customized a sample stage with 3D-printing (see below) which provides no physical contact at the drop-casting spot, ensuring that the precursor solution remains within the carbon substrate, avoiding the problem suggested by the AI.

References

1. Auer, P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research* 397–422 (2002).
2. Balandat, M. *et al.* BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *arXiv* (2019) doi:10.48550/arxiv.1910.06403.
3. Song, C. *et al.* PEM fuel cell reaction kinetics in the temperature range of 23–120°C. *Electrochimica Acta* **52**, 2552–2561 (2007).
4. Zhang, Y. *et al.* Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv* (2025) doi:10.48550/arxiv.2506.05176.
5. Nussbaum, Z., Morris, J. X., Duderstadt, B. & Mulyar, A. Nomic Embed: Training a Reproducible Long Context Text Embedder. *arXiv* (2024) doi:10.48550/arxiv.2402.01613.
6. Wang, W. *et al.* MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *arXiv* (2020) doi:10.48550/arxiv.2002.10957.
7. Xiao, S. *et al.* C-Pack: Packed Resources For General Chinese Embeddings. *Proc. 47th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* 641–649 (2024) doi:10.1145/3626772.3657878.
8. Yang, L. *et al.* Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. *arXiv* (2024) doi:10.48550/arxiv.2401.10891.
9. Ravel, B. & Newville, M. ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT. *J. Synchrotron Radiat.* **12**, 537–541 (2005).
10. Zabinsky, S. I., Rehr, J. J., Ankudinov, A., Albers, R. C. & Eller, M. J. Multiple-scattering calculations of x-ray-absorption spectra. *Phys. Rev. B* **52**, 2995–3009 (1995).
11. Riasse, R. *et al.* Benchmarking proton exchange membrane fuel cell cathode catalyst at high current density: A comparison between the rotating disk electrode, the gas diffusion electrode and differential cell. *J. Power Sources* **556**, 232491 (2023).
12. Lazaridis, T., Stühmeier, B. M., Gasteiger, H. A. & El-Sayed, H. A. Capabilities and limitations of rotating disk electrodes versus membrane electrode assemblies in the investigation of electrocatalysts. *Nat Catal* **5**, 363–373 (2022).
13. Choun, M., Hong, S. & Lee, J. Adsorbed Hydrogen as a Site-Occupying Species in the Electrocatalytic Oxidation of Formate on Pd/C in Alkaline Medium. *J Electrochem Soc* **165**, J3266–J3270 (2018).

14. Müller, S., Feurer, M., Hollmann, N. & Hutter, F. PFNs4BO: In-Context Learning for Bayesian Optimization. *arXiv* (2023) doi:10.48550/arxiv.2305.17535.
15. Rietveld, H. M. A profile refinement method for nuclear and magnetic structures. *J. Appl. Crystallogr.* **2**, 65–71 (1969).
16. Bail, A. L., Duroy, H. & Fourquet, J. L. Ab-initio structure determination of LiSbWO₆ by X-ray powder diffraction. *Mater. Res. Bull.* **23**, 447–452 (1988).
17. Hong, J. *et al.* Metastable hexagonal close-packed palladium hydride in liquid cell TEM. *Nature* **603**, 631–636 (2022).
18. Compound Interest. A periodic table of element prices.
<https://www.compoundchem.com/2019advent/day23/>