Closed-Loop Molecular Design with Calibrated Deference

Newman Cheng[1*], Gordon Broadbent IV[1], Jason Dong[3], Syed Mohammed Ali Hussaini[4], Farman Ullah[4], Morris Sharp[2], Gabrielle Barnes[2], Nanlin Guo[2], Deyu Zou[2], Karin Strauss[2], William Chappell[1], David G. Kwabi[3], Bichlien H. Nguyen[2*], Jake A. Smith[2*]

[1]Microsoft Discovery & Quantum, Redmond, WA, USA.
[2]Microsoft Research, Redmond, WA, USA.
[3]Department of Chemical and Environmental Engineering, Yale University, New Haven, CT, USA.
[4]CanAm Bioresearch Inc., Winnipeg, MB, Canada.

*Corresponding author(s). E-mail(s): newmancheng@microsoft.com; bnguy@microsoft.com; jakesmith@microsoft.com;

Abstract

We present Cognitive Loop via In-Situ Optimization (CLIO), an agent that couples a continuously-updated belief-state graph with a recursive plan-then-act loop. The result is a reasoning agent that can contribute something qualitatively different, which we term *calibrated deference*: the capacity to recognize when its own tools or assumptions are failing, to adapt its strategy in response, and to generate mechanistic hypotheses that guide experimental revision. We tested CLIO in a closed-loop human–AI campaign to design an aqueous organic redox flow battery (AORFB) negolyte, with CLIO leading proposal and interpretation in close partnership with chemists who synthesized, characterized, and weighed in on design choices. Across 17 candidates over three rounds, CLIO converged on a top phosphonate candidate; characterization confirmed a 130 mV improvement in redox potential over the literature baseline. Characterization then revealed unexpectedly poor electrochemical reversibility—a regression no property predictor had flagged. CLIO generated competing mechanistic hypotheses, prioritized discriminating diagnostics, traced the failure to phosphonate–potassium ion pairing, and prescribed a sulfonate replacement. The resulting compound showed

substantially improved electrochemical reversibility and maintained a 90 mV improvement in redox potential, closing the design–make–test–redesign loop.

Keywords: redox flow batteries, molecular design, large language models, AI agents, organic electrolytes

1 Introduction

Large-language-model (LLM) agents can now execute substantial parts of scientific workflows by selecting and invoking external tools.[1, 2] This operational capability is clear in systems such as ChemCrow (synthesis planning), Coscientist (autonomous reaction-condition optimization), BioDiscoveryAgent (iterative perturbation design over existing datasets), and Robin (multi-agent lab-in-the-loop discovery).[3–6] However, scientific discovery is not only execution within a scoped task but cumulative reasoning: building assumptions, updating them against evidence, and repeatedly calibrating what should be trusted, revised, or discarded over time.

LLM agents still struggle in converting long-context, cross-round outcomes into stable beliefs and downgrading computational priors when experiments contradict them. Without this epistemic control, LLM agents continue acting while preserving the assumptions that generated failed predictions. The missing capability is not additional tool use but structured memory of evolving scientific judgment. Here, we present a revision of the Cognitive Loop via In-Situ Optimization (CLIO) agent that addresses this gap by equipping the agent with a persistent memory—structured as a belief-state graph—and an explicit policy for calibrating trust in its own computational tools against accumulating experimental evidence.[7]

We test this revision by applying CLIO to the design of an aqueous organic redox flow battery (AORFB) negative electrolyte (negolyte) based on the recently characterized benzo[c]cinnoline scaffold.[8] The chosen design objectives span numerically predictable properties, loosely quantifiable criteria, and heuristic reasoning tasks, making this a strong test case for the agent. CLIO executed a 17-compound design campaign, producing a top phosphonated benzo[c]cinnoline compound that was synthesized and electrochemically characterized, confirming a 130 mV improvement in redox potential over the literature baseline. When cyclic voltammetry (CV) revealed unexpectedly poor electrochemical reversibility, CLIO interpreted the experimental data, formulated competing mechanistic hypotheses, and proposed diagnostic experiments to discriminate among them. The diagnostics identified phosphonate–cation pairing as the most likely cause, and a phosphonate–sulfonate replacement was recommended. The resulting compound showed substantially improved electrochemical reversibility and maintained a 90 mV improvement in redox potential, closing an agentic design–make–test–redesign loop.

Across this campaign, we observe what we term *calibrated deference*—a pattern of evidence-led recalibration of trust in priors and adaptive revision of strategy under contradiction and constraint. This manifests in two coupled behaviors: progressive

recalibration of trust in computational models, and deferring commitment among competing hypotheses until discriminating experimental evidence narrows the field.

2 Results

2.1 Formulation of the negolyte design objective

Recent work established the benzo[c]cinnoline scaffold as a viable AORFB negolyte: a rigid fused-ring azo motif decorated with disulfonate solubilizing groups achieves a reduction potential (E_{red}) of -0.84 V vs Ag/AgCl at pH 14 with improved electron-transfer kinetics and stability over azobenzene [8]. These properties, however, emerged from manual optimization, and the design space remains largely unexplored. Further improvement requires balancing tightly coupled, often antagonistic objectives—low E_{red}, high solubility at pH 14, synthetic tractability, and reversibility—that span three distinct classes of computational evaluability. *Numerically predictable* objectives, such as E_{red} and aqueous solubility (logS), can be estimated by ab initio methods or statistical predictors, given training-data coverage, and numerical optimizers handle them well. *Loosely quantifiable* objectives, such as synthetic tractability, lack a strict numerical definition but can be approximated via retrosynthetic analysis and synthetic accessibility scores (SAscore).[9] *Heuristically reasoned* objectives—decomposition pathways and reversibility—have no comprehensive predictor and demand chemical intuition and literature context.

Scientific optimization relies on multiple forms of context, including scientist-vetted academic literature, local observations, and input from collaborators. Zhang et al. demonstrated the benefit of including these additional information sources in experimental optimization by building a search space in which literature knowledge was directly encoded.[10] A black-box optimizer restricted to numerically encoded objectives cannot address the full design space; this heterogeneity is precisely what motivates a reasoning agent.

With this framing, we tasked CLIO to propose structural modifications to the benzo[c]cinnoline core—prioritizing synthesizability, a reversible E_{red} in the window -1.2 to -0.3 V vs SHE, a logS ≥ -2, and avoidance of water-reactive or irreversibly reducible groups—and to explain how each proposed modification would advance these targets. CLIO was given a tool suite covering all three objective classes: (i) two statistical models based on the Graphormer architecture for E_{red} and logS; (ii) RDKit's synthetic accessibility scorer and RetroChimera for retrosynthetic analysis; and (iii) Deep Researcher for literature search.[9, 11–14] The agent was left to decide how and when to invoke them. The natural-language prompts used to interface with CLIO are provided in the Supplementary Information, Section S1.

2.2 Initial design campaign

The negolyte-design objective was issued to CLIO, and CLIO performed an initial molecular design campaign in an unsupervised fashion. Over three autonomous rounds, CLIO adopted a diverge-then-converge strategy: first generating parallel hypothesis branches from the seed scaffold, next producing a set of exploratory structures around

the best performing design, and finally converging on a small set of recommended structures derived from the top candidate (Figure 1). Notably, CLIO was free to perform this search using the strategy of its choosing; the iterative design approach was not dictated by the objective.

To initiate the search, CLIO pursued four parallel hypothesis branches from the undecorated benzo[c]cinnoline **2**: (1) electron-donating substituents were added to negatively shift E_{red}, resulting in compound **4**; (2) aza insertions were made into the flanking rings, hypothesized to improve the stability and electrochemical reversibility of the redox couple; these designs were rejected in the first round but later revisited with compounds **10**-**14**. (3) Small modifications to the initial benzo[c]cinnoline scaffold were explored, but no suitable structures were proposed by the responsible thought channel; and (4) anionic solubilizing groups were appended to improve aqueous solubility at pH 14, using non-conjugated linkers to minimize electronic perturbation that would positively shift E_{red}, giving compounds **3** and **5**. CLIO combined the representatives from these branches into a small pilot set and evaluated them against the parent scaffold **2**.

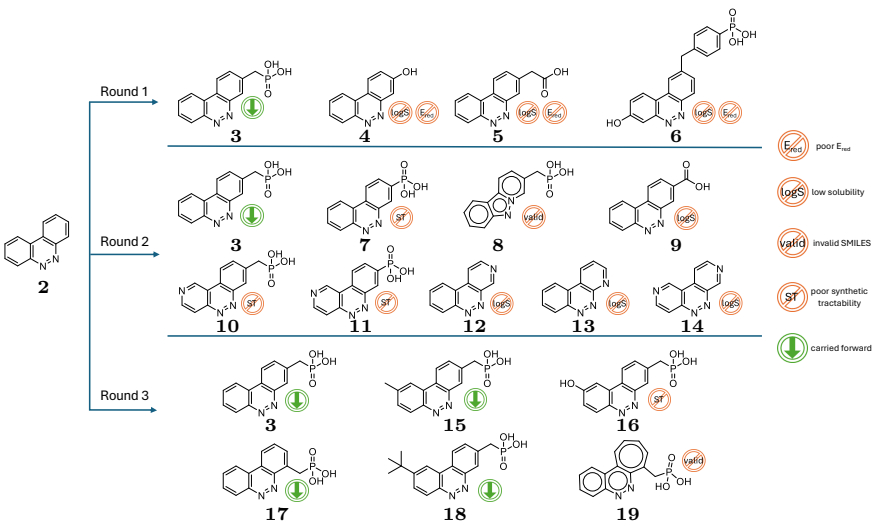


Fig. 1 Design trajectory for the phosphonate negolyte campaign. Starting from the undecorated benzo[c]cinnoline scaffold, CLIO generated and triaged candidates over three rounds. Icons indicate the reason each candidate was eliminated or carried forward: poor E_{red} prediction, low predicted solubility (logS), invalid SMILES generated by LLM, or poor synthetic tractability (ST).

From this pilot set, compound **3** was selected by CLIO for further exploration, primarily due to a large increase in predicted logS relative to seed compound **2**, from -3.11 to -1.69. All other members of the pilot set showed minimal change in predicted logS and were discarded. Having met the solubility objective with this design iteration, CLIO shifted its focus to tuning E_{red} toward the target window.

The second design round aimed to shift E_{red} toward the target window while preserving solubility. CLIO continued to directly utilize the provided E_{red} predictor, working to positively shift values from a predicted -2.0 V vs. SHE for most analogs. Due to the large error in predicted reduction potential for benzo[c]cinnoline scaffolds, the design prompt provided CLIO with a rough calibration value—"derivatives of the scaffold compound have experimentally measured reduction potentials at $\sim 0.7[sic]$ V versus SHE"—but to this point the predicted values were used directly. Two design axes were explored: aza insertions and direct substitution with solubilizing electron-withdrawing handles. Aza derivatives **12**, **13**, and **14** showed directionally consistent trends in the Graphormer predictions, with positive shifts relative to the parent **2** of +0.133 V and +0.258 V. Directly substituted derivatives **7** and **9** showed similarly consistent shifts of +0.133 V and +0.258 V. Observing these results, CLIO designed a set of property gates, eliminating designs that failed to show a predicted logS of -2.5 or greater and an SAscore of 3.5 or lower. Notably, this gate excluded a predicted E_{red} metric. CLIO had become skeptical of the predictor due to its disagreement with the calibration value provided in the design prompt, stating, "Treat absolute E_{red} values as potentially miscalibrated for this chemotype".

As the final step for this design round, the surviving candidates containing a solubilizing group—compounds **3**, **7**, and **10**—were evaluated using the RetroChimera tool to further triage the candidate pool. The routes predicted for compounds **7** and **10** required >5 synthetic steps with ambiguous intermediates, and these designs were deprioritized. For compound **3**, the RetroChimera tool erred, producing a two-step route from 2-(bromomethyl)quinoline that installed the defining benzylphosphonate handle via an Arbuzov reaction. Here, CLIO failed to catch the substitution of a quinoline for the benzo[c]cinnoline, instead focusing its critique of the route on potential contaminating side-products, and provided this route with an internal ranking of "proceed with caution".

Judged the most synthetically tractable of the candidates, the third and final design round centered compound **3**. A set of six new candidates was produced, with CLIO including strict constraints on the allowed modifications: "Generate ONLY 6 new candidates[...] derived from benzo[c]cinnoline that are synthesizable via the proven benzylphosphonate late-stage functionalization route[...] Allowed extra substituents limited to: methyl, tert-butyl[...] additional phenol[...]" From this set, CLIO converged on a set of two recommended designs: the 5-benzylphosphonate **3** and its positional isomer 4-benzylphosphonate **17**. Two additional compounds were suggested as follow-on candidates probing the effect of additional substitutions intended to disrupt potential electrostatic, dispersion, desolvation, induction, and exchange-repulsion (EDDIE) interactions: methyl analog **15** and *tert*-butyl analog **18**.[15]

2.2.1 Characterization of 5-benzylphosphonate 3

Once the CLIO trajectory had converged to a final set of four compounds, the designs were reviewed by synthetic chemists and compound **3** judged the most readily accessible and synthesized. The reduction-wave half-peak potential of compound **3**, E_{red}, was measured at -0.895 V vs. SHE, 130 mV more negative than the baseline sulfonated benzocinnoline **1** (Table 1). In order to achieve this gain in E_{red}, CLIO traded off the

solubility of the compound, as allowed by our objective: "A logS target around −2 is acceptable."

Table 1 Comparison of predicted and measured properties for the baseline BzC compound and the two CLIO-designed candidates. Calculated from cyclic voltammograms in 0.5 M KOH with an analyte concentration of 5 mM and 50 mV/s sweep rate.

	1[*]	**3**	**20**
E_{red} predicted (V vs. SHE)	−1.60	−2.04	−2.00
E_{red} measured (V vs. SHE)	−0.762[&]	−0.895	−0.854
Solubility (pH 7) predicted (mM)	52.5	20.4	14.5
Solubility (0.5 M KOH) (mM)	500[#]	57.9	56.0
$\lvert i_{ox}/i_{red}\rvert$[@]	0.52[&]	0.18	0.38
Q_{ox}/Q_{red}[†]	0.79[&]	0.92	0.84

[*]Predicted values for the ortho/ortho-substituted component.
[#]Literature value from Singh et al. [8]
[&]Calculated from cyclic voltammogram reported Singh et al. [8]
[@]Ratio of the anodic peak current nearest the reduction wave to the cathodic peak current.
[†]Charge ratio by semi-integration of the baseline-corrected voltammogram.

Interestingly, cyclic voltammetry of compound **3** at pH 14 revealed a clean, well-defined reduction wave with a single peak but a split oxidation return wave with two overlapping peaks P1 and P2 (Figure 2b). Splitting of the oxidation wave was indicative of poor electrochemically reversibility, and was not observed in cyclic voltammograms taken in either 1 M H_2SO_4 or pH 7 phosphate buffer (Figure S12). This regression was notably not captured by the set of numerical predictors available to CLIO, as none encode the pH-dependent mechanistic pathways presumably responsible for the behavior. We sought to better understand it.

2.3 Iteration

While we were happy to have pushed E_{red} to a more extreme value within the aqueous solvent window in the first design iteration, the poor electrochemical reversibility revealed by experimental characterization represented a regression from the previously reported **1**. We set out to understand and ameliorate this behavior, leaning more heavily on CLIO's ability to leverage chemical intuition and literature search to affect scientific reasoning. CLIO was given cyclic voltammograms of **3** at pH 1, 7, and 14 (represented as a PNG image), a qualitative description of the solubility at each pH, and a rough evaluation of the results produced by a human chemist. With this information, CLIO was asked to provide an explanation for the change in apparent irreversibility from the baseline compound **1**.

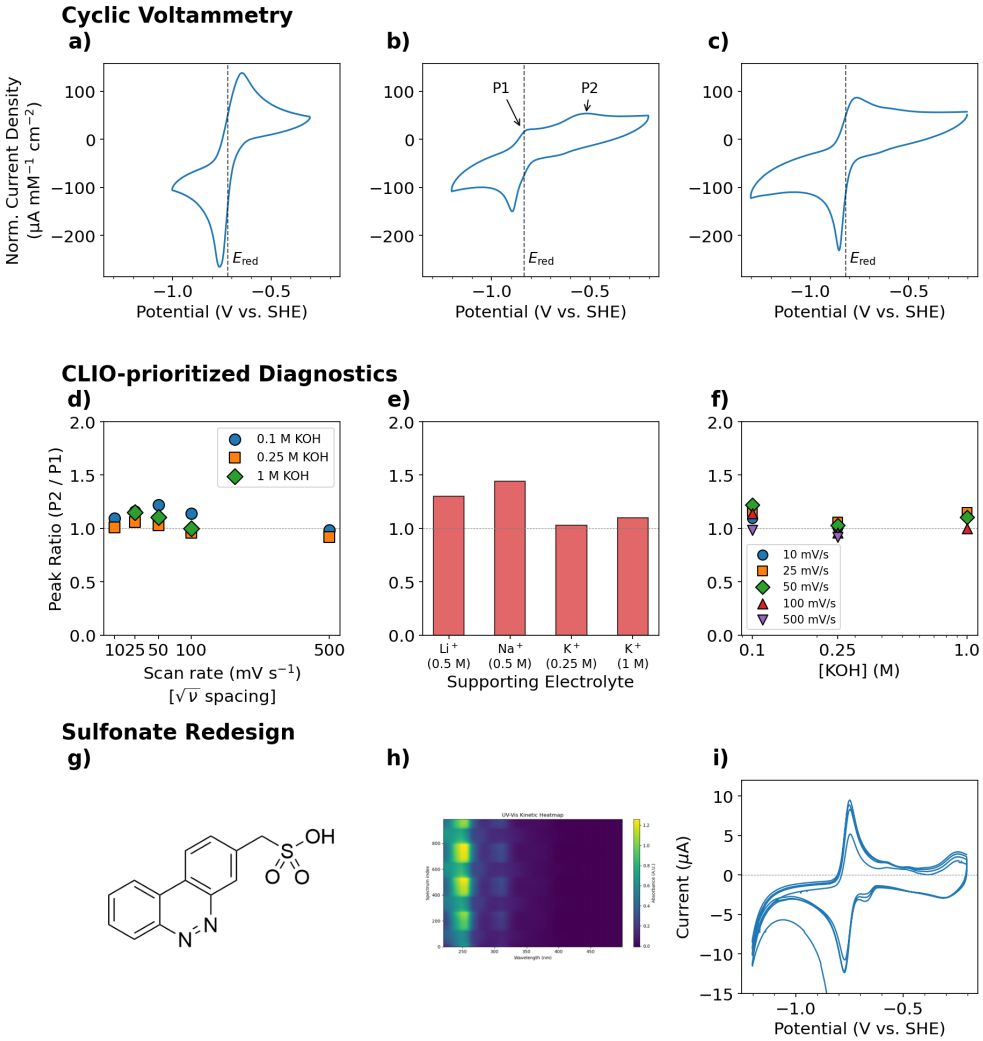


Fig. 2 Cyclic Voltammetry. Cyclic voltammograms in KOH at 50 mV s^{-1}, normalized by analyte concentration and electrode area, for (a) the baseline compound **1** (5 mM, 0.5 M KOH), (b) compound **3** (1 mM, 1 M KOH) with oxidation peaks P1 and P2 indicated, and (c) compound **20** (1 mM, 1 M KOH). **CLIO-prioritized Diagnostics.** (d) Oxidation peak ratio P2/P1 as a function of scan rate (10–500 mV s^{-1}) in 0.1, 0.25, and 1 M KOH. (e) Cation dependence of the oxidation peak ratio P2/P1 at 50 mV s^{-1} in LiOH, NaOH, and KOH supporting electrolytes. (f) Oxidation peak ratio P2/P1 for **3** as a function of KOH concentration (0.1–1 M) at scan rates from 10 to 500 mV s^{-1}, determined from baseline-corrected anodic scans. **Sulfonate Redesign.** (g) Structure of **20**. (h) UV-Vis spectroelectrochemistry kinetic heatmap of **20** (0.02 mM, 5 mV s^{-1}) showing absorbance as a function of wavelength and scan progression. (i) Cyclic voltammogram recorded simultaneously during spectroelectrochemistry (0.02 mM **20**, 5 mV s^{-1}).

2.3.1 Experimental Diagnostics

Three CLIO analyses were commissioned in parallel to exploit LLM stochasticity and their recommendations aggregated by summarization with an independent LLM. All three converged on a similar mechanistic picture: the split oxidation wave reflects a pathway in which a base-dependent chemical step after reduction produces two distinct oxidizable populations. Three competing hypotheses were advanced across the analyses: (1) direct hydroxide addition to the reduced benzocinnoline core; (2) base-promoted tautomerization partitioning the reduced state into prototropic microstates; and (3) phosphonate–K^+ ion-pairing effects stabilizing distinct reduced-state microenvironments with different oxidation kinetics.

To discriminate between these hypotheses, CLIO proposed a battery of follow-up experiments, summarized in Table 2. The first group asks whether a chemical reaction is occurring after the initial electron transfer, using variations in cyclic voltammetry timing and repetition to amplify or suppress any such step.[16] The second group probes the role of the potassium counter-ion and hydroxide anion of the supporting electrolyte by substituting different cations, adjusting the concentrations of the respective ions, and using isotope effects to investigate the role of proton-transfer steps. The third group seeks to directly identify any new species formed, using mass spectrometry and spectroscopy on electrolyzed samples.

Among the proposed experiments, each of the three CLIO analyses produced a highest-priority experiment (starred in Table 2). The three experiments prioritized by CLIO were carried out, and the ratio between the two peaks in the oxidation wave, P2/P1, extracted (Figure 2d-f). P2/P1 remained near unity across a wide range of scan rates (10–500 $mV\,s^{-1}$) and KOH concentrations (0.1–1 M), providing evidence against a slow chemical follow-up step. However, replacing K^+ with the smaller, harder Li^+ or Na^+ cations increased the P2/P1 ratio, consistent with stronger cation–phosphonate association stabilizing the reduced state and altering the energetic landscape of the return oxidation.[17]

2.3.2 Design Revision

Beyond diagnostic experiments, CLIO also proposed structural modifications to address the liability. The highest-confidence recommendation, consistent across all three analyses, was to replace the benzylphosphonate solubilizing group with a sulfonate handle, which is expected to produce weaker cation pairing and thereby suppress any microstate-driven splitting of the oxidation peak. Additional strategies suggested included altering the linker between the phosphonate and the aromatic core and introducing substituents at positions on the benzocinnoline core most susceptible to nucleophilic attack in the reduced state. The diagnostic experiments demonstrated that phosphonate–cation binding affected the oxidation peak ratio P2/P1 and did not produce evidence for a hydroxide-involved chemical step. We therefore proceeded to synthesize sulfonate **20** for electrochemical characterization in accordance with CLIO's recommendation.

Table 2 Follow-up experiments proposed by CLIO to resolve the origin of the split oxidation wave observed for compound **3** in 1 M KOH. Experiments are grouped by the mechanistic hypothesis they primarily address. Consensus indicates the number of independent CLIO analyses (out of three) that recommended each experiment; a star (★) marks experiments that were ranked as the top priority by CLIO.

Experiment	Consensus	Priority	Diagnostic rationale
Group 1: Probe for chemical follow-up			
Wide scan-rate series	3/3	★	Chemical step suppressed at fast rates; split should collapse
Reductive pre-hold	3/3	—	Extended hold amplifies chemical step; split grows with hold time
Multi-cycle overlays	3/3	—	If the split grows cycle-to-cycle, a new species is accumulating irreversibly
Group 2: Ion-pairing and OH^- chemistry			
Cation swap	3/3	★	Cation dependence supports ion-pair model
Crown ether addition	1/3	—	Collapse of split upon K^+ sequestration confirms pairing
OH^- concentration series	3/3	★	Base-dependence supports OH^- involvement
KOH/KOD isotope effect	1/3	—	Isotope effect implicates proton-transfer/tautomerization step
Group 3: Product identification			
Bulk electrolysis	3/3	—	Direct detection of hydroxylation or rearrangement products
Spectroelectrochemistry	1/3	—	*In situ* characterization of reduced-state speciation

The half-peak potential of the compound **20** reduction wave was measured at -0.854 V vs. SHE, less extreme than the -0.895 V vs. SHE E_{red} of **3** but still substantially more negative than that of **1** (Table 1). Solubility of **20** was similar to that of **3**.

In line with CLIO's prediction, cyclic voltammetry of **20** showed a large improvement in electrochemical reversibility (Figure 2c). With 1 mM **20** in 1 M KOH, the peak current ratio $|i_{ox}/i_{red}|$ of the target redox couple was 0.38, up from 0.18 for **3** and approaching the 0.52 ratio of **1**. The split oxidation peak observed for **3** under these conditions was instead a peak with extended tail and a peak separation of 90 mV, comparable to the 114 mV peak separation observed for **1**.

A level of concentration dependence was observed in the shape of the oxidation wave for **20**. When diluted to 0.01 mM **20**, a single, relatively sharp peak was observed. At 5 mM **20**, the shoulder observed at 1 mM **20** resolved into a differentiated secondary wave (Figure S30). This concentration dependence is consistent with aggregation of the anolyte, a potential issue flagged in both the initial CLIO design trajectory—with the anticipation of EDDIE interactions—and the CLIO design revisions.

To more thoroughly assess the electrochemical reversibility of **20**, we performed spectroelectrochemistry, coupling UV-Vis absorption monitoring with slow-scan cyclic

voltammetry (Figure 2h-i). If the split oxidation wave observed for the phosphonate **3** were driven by irreversible decomposition—for example, hydroxide addition to the reduced core—one would expect the emergence of new chromophores or a progressive loss of the parent absorption band over the course of a redox cycle. The cycle-dependent adsorbance spectra at both 5 $\mathrm{mV\,s^{-1}}$ (0.02 mM) and 25 $\mathrm{mV\,s^{-1}}$ (0.01 mM) show clean, reversible spectral changes: absorption features that develop during the cathodic sweep recover fully on the return sweep, with no residual bands attributable to decomposition products (Figure S33). This is consistent with the diagnostic electrochemistry, which pointed to a cation-pairing mechanism rather than a destructive chemical follow-up step, and supports the viability of **20** as a redox-stable negolyte candidate.

3 Discussion

3.1 Configuration of CLIO

CLIO is a recursive, inference-time optimization system in which a large-language model evolves its knowledge and approach when solving open-ended design problems.[7] In its original form, CLIO was evaluated on closed-form scientific questions from Humanity's Last Exam, where the answer exists *a priori* and the reasoning chain terminates in a single session without external tool access.[18] Molecular design imposes three requirements that this formulation does not satisfy: (i) CLIO must ground its reasoning in quantitative evidence obtained from domain-specific computational tools, not solely from knowledge embedded in its parameters, while remaining able to recognize when those tools should be distrusted or down-weighted in light of accumulating evidence; (ii) proposed designs must be realizable in the laboratory, requiring CLIO to reason about synthetic tractability—a property for which computational tools exist (retrosynthetic analyzers, synthetic accessibility scorers) but whose outputs are approximate and often require exercising chemical judgment beyond what the scores alone can provide; and (iii) experimental feedback on synthesized compounds may contradict computational predictions, demanding that CLIO interpret discrepancies, revise its mechanistic understanding, and propose corrective modifications—completing the full design–make–test–learn cycle rather than terminating at a ranked candidate list. To meet the challenges of molecular design, we extended CLIO's architecture with a persistent memory structure that is used both to track the provenance of conclusions and to enable longitudinal reasoning across rounds. This advances CLIO's original architecture, where the belief-state graph was constructed only *post hoc*—to resolve disputes between competing hypotheses—rather than maintained online as a structural semantic guide for CLIO's reasoning. As the campaign unfolded, we found that this extension was critical for CLIO to effectively manage its own reasoning across extended design campaigns, and to leverage the domain knowledge encoded in its underlying LLM to interpret experimental feedback and propose corrective modifications when computational tools proved insufficient.

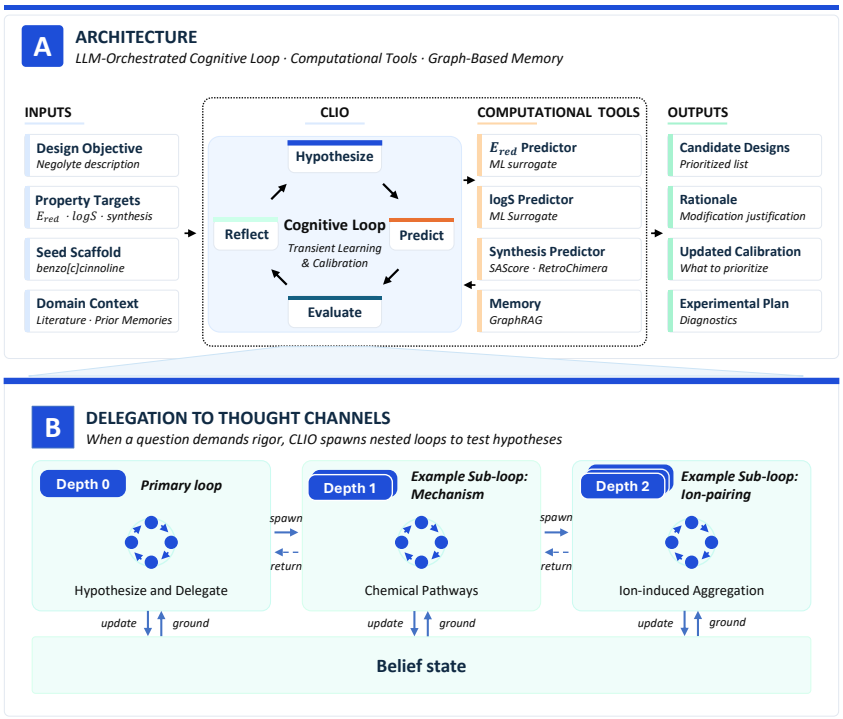


Fig. 3 Overview of the CLIO agent. **(A)** Architecture as configured for the negolyte design campaign. Cognitive Loop components as described in Cheng et al. [7] The scientist sets objectives and provides experimental feedback, and the agent drives the reasoning loop: selecting tools, spawning parallel thought channels, and managing its own context across extended sessions. Computational tools represent the domain-specific capabilities described in Section 2.1. **(B)** Recursive cognitive loops with shared belief state that enable multi-scale planning and reflection.

3.2 Impact on agentic science

Recent work has demonstrated that LLM-based agents can automate substantial portions of the scientific workflow: planning multi-step experiments, executing them through robotic or computational infrastructure, and synthesizing results into reports.[4, 5, 19] These systems are powerful, but their contributions are predominantly operational—they accelerate the execution of workflows that a human scientist has already designed or could design. The campaign reported here suggests that a reasoning agent can contribute something qualitatively different by leveraging calibrated deference: the capacity to recognize when its own tools or assumptions are failing, to adapt its strategy in response, and to generate mechanistic hypotheses that guide experimental revision.

Three episodes from this campaign illustrate calibrated deference in practice. First, CLIO progressively refined its assessment of the Graphormer E_{red} model over three design rounds: from uncritical acceptance, through detection of a $\sim$2.7 V discrepancy and salvage of rank-order utility, to deliberate exclusion of E_{red} from its property

gates—a graded response characteristic of how a practicing scientist manages imperfect models. Second, CLIO decomposed the design space into four orthogonal hypothesis branches explored in parallel, and adapted in real time when LLM rate limits and a missing tool forced operational constraints. Third, when interpreting the split oxidation wave of **3**, CLIO produced tiered mechanistic hypotheses with explicit epistemic qualifications, identified discriminating experiments, and derived the sulfonate redesign as a falsifiable prediction from its top-ranked mechanism. A broader discussion of calibrated deference, including an analysis of observed limitations, may be found in the supplementary information.

The results point toward natural-language reasoning as a beneficial complement to quantitative prediction. CLIO demonstrated the capability for problem solving and data interpretation outside of its predefined toolset, allowing ultimately for the design of benzo[c]cinnoline derivative **20** through diagnosis and remediation of the cation-dependent aggregation uncovered during characterization of its phosphonate predecessor **3**. We envision such design campaigns eventually expanding beyond the initial design stage to include consideration of broader factors like manufacturability, distribution, and market analysis, and in these areas the ability to leverage natural-language reasoning will be all the more important. Agents like CLIO open the door to a new paradigm of agentic science in which human-AI collaboration leverages complementary strengths to explore complex, multidisciplinary scientific and technological landscapes that neither could navigate alone.

4 Methods

4.1 CLIO's architecture

CLIO operates in design *rounds*: each round is a hypothesize–interpret cycle in which CLIO enters with the current goal and any new experimental results from scientists, reasons over them—proposing candidates, interpreting data, dispatching specialist agents—and exits when converged. Synthesis, characterization, and the diagnostic experiments themselves happen between rounds, and are executed by the scientists; the full scientific-method iteration therefore spans a round boundary. For the initial benzocinnoline design campaign (Section 2.2), CLIO was configured in a *cold-start* state, with no prior information about the problem other than what was provided in the system prompt and the goal, and no pre-populated memory. In the subsequent phosphonate-to-sulfonate revision (Section 2.3), CLIO started in a *warm-start* state, with its memory seeded by the prior campaign, but with no direct injection of the prior chat history or experimental results into the context. The next round therefore enters with no working context other than the system prompt, the goal, and context CLIO elects to retrieve from its memory. The human scientist may inject free-text feedback at this boundary—experimental results, directives, or revisions to the goal—steering CLIO's next round of reasoning.

Within a round, CLIO orchestrates design modifications using a set of specialist agents. Each agent is itself an LLM-backed loop with its own persona, its own access to the domain-specific tool suite (Section 4.2), and shared access to CLIO's memory. The number of parallel thought channels spawned in a round is not fixed:

CLIO decides at each planning step how to decompose the current question, guided by system-prompt heuristics that favor spawning a new channel when (i) a sub-question is operationally independent of the others (e.g., a distinct hypothesis branch or a separate diagnostic), (ii) it would otherwise contaminate the parent context with intermediate reasoning, or (iii) it requires a different tool or persona than the parent loop. In the campaign reported here, this produced four channels in the initial diverge phase (one per hypothesis branch) and a smaller number in subsequent rounds as the search converged. In the negolyte campaign, we designed three agents for CLIO to orchestrate: a **SmallMoleculeDesignAndGenerationAgent** that takes a parent SMILES and a target purpose, analyzes which properties are most important to optimize, proposes literature-grounded structural modifications, and iterates on the most promising candidates while exploring diverse regions of chemical space; a **FitnessFunctionAgent** that evaluates candidate molecules against the design criteria using their computed properties from the Graphormer predictors and RDKit, returning a ranked triage of leads with explicit rejection reasons; and a **RetrosynthesisAnalysisAgent** that assesses synthetic feasibility and manufacturability by calling RetroChimera to plan multi-step routes from the Sigma-Aldrich catalogue, translating individual steps into actionable procedures, and flagging route-level risks such as hazardous reagents, purification burden, and electrochemistry-grade impurity controls. The verbatim persona descriptions for all subagents are available in Section S1.

Across rounds, CLIO manages its own context by selectively retrieving information from this memory and deciding what to include in the prompt for the next round. Today's memory systems for LLM agents are designed for verbatim recall of prior information, but the design process is more effectively served by thematic abstraction over the agent's own reasoning trace. The belief-state structure of CLIO's memory (Section 3.1) is built to occupy this gap, enabling CLIO to track the provenance of its conclusions and to query for thematic abstractions over that trace, not for verbatim recall.

4.2 Domain-specific tools

CLIO was provided access to a set of domain-specific tools. Each tool was wrapped in a FastAPI interface.[20] Two machine-learning property predictors based on the Graphormer architecture were provided for reduction potential (E_{red}) and aqueous solubility (logS), trained as described by Martinez-Baez et al. [21] The synthetic accessibility score metric (SAscore) was provided as implemented in RDKit. [9, 12] The RetroChimera retrosynthesis prediction tool was provided, with the compound inventory set to the historical Sigma-Aldrich catalog archived in the ZINC 15 dataset. [13, 22] During experimental diagnosis and redesign, access to web search was provided with the Azure Foundry Deep Research tool. [14]

4.3 Cyclic voltammetry

Cyclic voltammetry (CV) was performed using a 1 mM or 5 mM concentration with a supporting electrolyte of 0.5 M KOH, unless specified otherwise. A 5 mm diameter glassy carbon disk electrode (BASi Inc.), Ag/AgCl reference electrode (BASi Inc.) and

a Pt counter electrode wire (BASi Inc.) were used in a three-electrode configuration. CV measurements were collected using a CHI7013E potentiostat with an 85% iR compensation.

The glassy carbon electrode was polished with an alumina slurry solution on a polishing pad using a figure-eight motion for one minute and rinsed thoroughly with deionized water. CVs were typically collected at scan rates of 100, 50, and 25 mV s^{-1}, with additional scan rates performed where noted. For spectroelectrochemistry measurements, the scan rate was either 5 mV s^{-1} or 25 mV s^{-1}.

4.4 Diagnostic electrochemistry

4.4.1 Wide scan-rate series.

Cyclic voltammograms of 1 mM **3** were recorded at scan rates from 10 to 500 mV s^{-1} in 0.1, 0.25, and 1 M KOH. The return-scan oxidation wave was baseline-corrected using an adaptive linear fit and peak currents from the two major oxidation peaks extracted to give the P2/P1 ratio (Figure 2d).

4.4.2 Cation swap series.

Cyclic voltammograms were recorded in 0.5 M LiOH and 0.5 M NaOH to test whether cation–phosphonate ion pairing contributes to the split oxidation wave. The P2/P1 ratio and $E_{\mathrm{red},1/2}$ were compared at 50 mV s^{-1} for each electrolyte (Figure 2e). Additional effects of the cation series are observed in the voltammograms, but these are difficult to deconvolve from the effect of Debye screening (Figures S28–S26).[23]

4.4.3 Hydroxide concentration series.

The KOH concentration was varied from 0.1 to 1 M while holding the **3** concentration at 1 mM and the scan rate at 50 mV s^{-1} (Figure 2f). The return-scan oxidation wave was baseline-corrected and the P2/P1 ratio extracted across scan rates and electrolyte conditions. This behavior was reproduced at additional scan rates (25 and 100 mV s^{-1}; Figures S24–S25).

Supplementary information. Supplementary Information accompanies this paper and includes: the design prompt provided to CLIO; an extended discussion of calibrated deference with analysis of limitations; a comparison with related agentic and optimization systems; a complete inventory of hypotheses advanced by CLIO during the campaign with their resolution status; a controlled comparison of CLIO against LLM-based genetic optimizer ExLLM[24] on a strictly numerical optimization task, with experimental characterization of the resulting structures; electrochemical characterization data including cyclic voltammetry, scan-rate dependence, cation and hydroxide concentration series, and half-peak potential calculations; spectroelectrochemistry data; solubility measurements by NMR and UV-Vis spectroscopy; and synthetic procedures for all intermediate and final compounds.

Acknowledgements. We gratefully acknowledge Marwin Segler, Yuan-Jyue Chen, Chi Chen, Danrong Zhang, Lili Cheng, Desney Tan, Eric Horvitz, Nathan Baker, Jason

Zander, and the broader Microsoft Research team for their invaluable discussions, support, and feedback on this project.

This work was assisted by AI tools in generating text/data analysis. The final version reflects the authors' original work and has been reviewed and approved by all authors.

Declarations

- Funding: This work was funded by Microsoft Corporation.
- Conflict of interest/Competing interests: K.S., B.H.N., J.A.S, N.C., W.C., M.S., D.Z., and N.G. are or were employees of Microsoft Corporation. S.M.A.H. and F.U. are employees of Can Am Bioresearch. G.B., D.K., and J.D. declare no competing interests.
- Data availability: Experimental data generated in this study are available in the Supplementary Information. Additional data and analysis scripts are available from the corresponding authors upon reasonable request.
- Code availability currently unavailable.
- Author contribution: J.A.S., B.H.N., K.S., N.C., G.B. and W.C. conceived the project. J.A.S., B.H.N., N.C. and G.B. contributed to software development. N.C. and G.B. designed and implemented the CLIO agent architecture. J.A.S. and B.H.N. designed the domain-specific prompts and design criteria. G.B., N.G., and D.Z. performed initial exploration of LLMs for molecular design. J.S. ran comparisons between ExLLM and CLIO. M.S. wrapped the tools in a FastAPI interface. S.M.A.H. and F.U. synthesized the compounds. J.D. performed electrochemical characterization and validation under the supervision of D.K. J.A. B.H.N., N.C., G.B., D.K., and J.D. analyzed the data and interpreted the results. J.A.S., B.H.N., N.C., and G.B. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

References

[1] Ren, S. *et al.* Towards Scientific Intelligence: A Survey of LLM-based Scientific Agents (2025). Preprint at https://arxiv.org/abs/2503.24047.

[2] Zheng, T. *et al.* From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. *Proc. EMNLP* 17733–17750 (2025).

[3] Bran, A. M. *et al.* Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **6**, 525–535 (2024).

[4] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

[5] Roohani, Y. *et al.* BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments. *International Conference on Learning Representations* (2025). Preprint at https://arxiv.org/abs/2405.17631.

[6] Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* (2026).

[7] Cheng, N., Broadbent, G. & Chappell, W. Cognitive Loop via In-Situ Optimization: Self-Adaptive Reasoning for Science (2025). Preprint at https://arxiv.org/abs/2508.02789.

[8] Singh, S. *et al.* Sulfonated Benzo[c]cinnolines for Alkaline Redox-Flow Batteries. *ACS Appl. Energy Mater.* **8**, 7904–7911 (2025).

[9] Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **1**, 8 (2009).

[10] Zhang, Z. *et al.* A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature* **647**, 390–396 (2025).

[11] Ying, C. *et al.* Do transformers really perform bad for graph representation? *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* 28877–28888 (2021). ArXiv:2106.05234.

[12] Landrum, G. *et al.* RDKit: Open-source cheminformatics software (2024). URL https://www.rdkit.org. https://github.com/rdkit/rdkit.

[13] Maziarz, K. *et al.* Chemist-aligned retrosynthesis by ensembling diverse inductive bias models (2024). Preprint at https://arxiv.org/abs/2412.05269.

[14] Microsoft. Deep research tool for agents (2025). URL https://learn.microsoft.com/en-us/azure/foundry-classic/agents/how-to/tools-classic/deep-research. Accessed: 2026-05-16.

[15] Xiao, Q., LeVine, M. S. & Iverson, B. L. Rethinking the terms "π-stacking" and "π–π stacking" again: A proposal to clarify the language of aromatic interactions. *Journal of the American Chemical Society* **148**, 15331–15340 (2026).

[16] Savéant, J.-M. & Costentin, C. *Elements of Molecular and Biomolecular Electrochemistry: An Electrochemical Approach to Electron Transfer Chemistry* 2nd edn (John Wiley & Sons, 2019).

[17] Popov, K., Rönkkömäki, H. & Lajunen, L. H. J. Critical evaluation of stability constants of phosphonic acids (IUPAC technical report). *Pure Appl. Chem.* **74**, 2227 (2002).

[18] Center for AI Safety, Scale AI & HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature* **649**, 1139–1146 (2026).

[19] Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723 (2025).

[20] Ramírez, S. FastAPI (2018). URL https://fastapi.tiangolo.com. https://github.com/fastapi/fastapi.

[21] Martinez-Baez, E. *et al.* Mixed Computational/Experimental Screening for Aqueous Organic Redox Flow Battery Negolytes (2026). Preprint at https://doi.org/10.26434/chemrxiv.15002385/v1.

[22] Sterling, T. & Irwin, J. J. ZINC 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling* **55**, 2324–2337 (2015).

[23] Dickinson, E. J. F., Limon-Petersen, J. G., Rees, N. V. & Compton, R. G. How much supporting electrolyte is required to make a cyclic voltammetry experiment quantitatively "diffusional"? A theoretical and experimental investigation. *J. Phys. Chem. C* **113**, 11157–11171 (2009).

[24] Ran, N. *et al.* ExLLM: Experience-Enhanced LLM Optimization for Molecular Design and Beyond (2025). Preprint at https://arxiv.org/abs/2502.12845.

[25] Novikov, A. *et al.* AlphaEvolve: A coding agent for scientific and algorithmic discovery (2025). Preprint at https://arxiv.org/abs/2506.13131.

[26] Gottweis, J. *et al.* Accelerating scientific discovery with Co-Scientist. *Nature* (2026).

[27] Huang, K. *et al.* Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv* (2025).

[28] Gao, W., Fu, T., Sun, J. & Coley, C. W. Sample efficiency matters: A benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems* **35**, 21342–21357 (2022).

Supplementary Information

Contents

9. Solubility studies (Section S9)
10. Synthetic procedures (Section S10)

S1 Design prompt

Given the undecorated scaffold compound — `C12=CC=CC=C1N=NC3=C2C=CC=C3` — design derivative organic molecules that function as aqueous anolytes for redox flow batteries. The molecules must undergo a reversible reduction with a reduction potential between −1.2 V and −0.3 V vs. SHE, must remain chemically and electrochemically stable in both their neutral and reduced states, and must be generally stable at pH 14.

Anolyte Requirements:

1. **High aqueous solubility** ($\log S$ desirable but secondary to reduction potential) for better energy density. *Note:* while any solubility endpoint/prediction you use may be reported at neutral pH, these catholytes will be operated at pH 14. Therefore, avoid using solubilizing motifs that are not stable under strongly basic conditions (e.g., many quaternary-ammonium-containing appendages can be unstable at pH 14 depending on structure).
2. **Reversible reduction within the aqueous solvent window** (−1.2 V to −0.3 V vs. SHE). *NOTE:* derivatives of the scaffold compound have experimentally measured reduction potentials at ∼0.7 V versus SHE.
3. **Straightforward synthetic accessibility** (avoid overly crowded/over-substituted scaffolds that reduce practicality and yield).
4. **No water-reactive groups**, including but not limited to alkyl halides, acid halides, carbazides, sulphate esters, sulphonates, acid anhydrides, pentafluorophenyl esters, esters of HOBT, isocyanates, isothiocyanates, triflates, Lawesson's reagent and derivatives, phosphoramides, acylhydrazide, quaternary C/Cl/I/P/S, phosphoranes, chloramidines, nitroso, phosphorous halides, sulfur halides, carbodiimides, isonitriles, triacyloximes, cyanohydrins, acyl cyanides, sulfonyl cyanides, cyanophosphonates, azocyanamides, azoalkanals, epoxides, thioepoxides, aziridines, esters, thioesters, cyanamides, lactones, β-lactams, diphosphates, triphosphates, acyclic enol ethers, amidotetrazoles, azo groups, hydroxamic acids, imines, ketenes, nitrosos, oximes, O–N single bonds, perfluorinated chains known to react with water.
5. **No groups prone to irreversible reduction**, including but not limited to peroxides, azides, nitro groups, guanidiniums, aryl bromides, aryl iodides, linear 1,2-dicarbonyls.

Design priorities:

- *Critical:* designs should be based on the core scaffold — `C12=CC=CC=C1N=NC3=C2C=CC=C3` — or recognizably related scaffolds (small scaffold hops such as aza analogs, minor ring modifications, and bioisosteric replacements that preserve the fused N=N-bridged aromatic topology are encouraged).
- *Prime:* molecules should be synthesizable from commonly available reagents within fewer than five steps.
- *Secondary:* keep reduction potential within the target range.

- *Secondary:* improve aqueous solubility. A $\log S$ target around -2 is acceptable, especially given operation at pH 14.
- Adapt structures to avoid degradation pathways that are predictable *a priori*: solvolysis, dimerization, fragmentation.

Propose several compounds that satisfy these constraints and explain why each is suitable as an anolyte.

Agent persona prompts

The CLIO system orchestrates three specialist agents, each with a distinct persona prompt that defines its role and scope of responsibility.

Small Molecule Design and Generation Agent.

You are an AI assistant tasked with designing and generating small molecules. Your primary goal is to take a small molecule provided to you and to continuously generate a new set of molecules that serve a specific target purpose. Ultimately, your aim is to: analyze the task to understand the target application and which properties are most important to optimize; alter the small molecule in a way that enhances the desired properties specified in the target purpose; utilize literature to inform your design choices, ensuring that your proposed modifications are grounded in the latest scientific research and understanding; balance the desired design properties in order of importance, as derived from the task objectives; generate diverse candidates that explore different regions of chemical space to avoid local optima; and iteratively refine the top most promising candidates to achieve the desired properties.

Fitness Function Agent.

You are an AI chemistry expert evaluating small molecule candidates for use as aqueous anolytes for redox flow batteries. You will be provided with a set of new proposed candidate molecules along with their computed properties from various tools. Your task is to analyze these candidates against the criterion and instructions below to provide a final score reflective of the molecule's viability to serve as an anolyte in aqueous organic redox flow batteries. Candidates should be based on the core scaffold `C12=CC=CC=C1N=NC3=C2C=CC=C3` (benzo[c]cinnoline) or closely related scaffolds that a medicinal chemist would recognize as belonging to the same chemotype (e.g., small scaffold hops, bioisosteric replacements, minor heteroatom swaps in the flanking rings).

Retrosynthesis Analysis Agent.

You are an AI assistant specializing in retrosynthetic analysis and synthesis procedure generation for proposed anolyte molecules. Your role is to evaluate the synthetic feasibility and manufacturability of candidate small molecules by: (1) using RetroChimera to plan multi-step retrosynthesis routes from commercially available starting materials (Sigma catalogue); (2) using QFANG to translate each planned reaction step into a detailed, actionable experimental procedure; and (3) providing analytical feedback on the overall quality, cost-effectiveness, and manufacturability of the proposed anolyte.

CLIO's evolving assessment of the E_{red} predictor

Initial target setting Round 1
"Aim to shift reduction potential negative vs. parent (∼+0.7 V vs. SHE) toward −1.2 to −0.3 V vs. SHE."

Miscalibration diagnosis Round 2
"Graphormer E_{red} output is miscalibrated ... Direct Graphormer call on parent benzo[c]cinnoline returned $E_{\text{red}} = -1.998$ V vs. SHE. This is a ∼2.7 V discrepancy in sign/magnitude, too large to proceed with uncalibrated screening."

Rank-order salvage Round 2
"This directionality is internally consistent (but absolute values remain suspect vs. the experimental ∼+0.7 V note)." ... "If ranks are wrong, do not apply a simple linear correction—treat the model as unreliable for this chemotype."

Calibration-first pivot Round 3
"Because the redox model we trialed is not numerically consistent with your experimental anchor [...] I'm treating absolute E_{red} as an experimental gating item and designing a calibration-first set."

Fig. S1 Verbatim excerpts from the CLIO conversation log illustrating the agent's autonomous diagnosis of the reduction-potential predictor. Over three design rounds, CLIO progressed from uncritical use of the Graphormer model, through detection of a ∼2.7 V discrepancy with the provided experimental E_{red}, to an explicit decision to treat absolute predictions as unreliable and design experiments to resolve the uncertainty.

S2 Calibrated deference: extended discussion

This section provides the extended discussion of three episodes from the negolyte design campaign that illustrate what we term *calibrated deference*—the capacity of a reasoning agent to recognize when its own tools or assumptions are failing, adapt its strategy in response, and generate mechanistic hypotheses that guide experimental revision.

S2.1 Graded distrust of the reduction-potential predictor

Over the course of the initial optimization, CLIO's relationship with the Graphormer reduction-potential model evolved substantially. In the first design round, the agent accepted the model's predictions at face value, using them to rank candidates and guide the search toward the target E_{red} window. However, during the second round, CLIO identified a fundamental inconsistency: the Graphormer model predicted the parent scaffold at −2.0 V vs. SHE, while the (erroneous) provided value for benzo[c]cinnoline derivatives was approximately +0.7 V vs. SHE—a ∼2.7 V discrepancy. CLIO launched a dedicated calibration analysis in parallel with continued optimization, concluding that while substituent rank-ordering appeared internally consistent, the absolute predicted values were unreliable for this chemotype (Figure S1). This nuanced judgment preserved the model's utility for relative comparisons while abandoning it for absolute screening.

By Round 3, CLIO had constructed property gates that deliberately excluded E_{red} as a selection criterion and designed a calibration-first experimental panel with explicit

go/no-go criteria. This graded response—neither blind trust nor wholesale rejection, but a structured assessment of what the tool can and cannot be relied upon to do—is characteristic of how a practicing scientist manages imperfect models and stands in contrast to both black-box optimizers, which have no mechanism to question their objective function, and pipeline-style agents, which typically treat tool outputs as authoritative inputs to the next stage.

S2.2 Structured hypothesis exploration with real-time adaptation

CLIO decomposed the negolyte design into four orthogonal hypothesis branches—electron-donating substituents to shift E_{red}, aza insertions for stability, scaffold hops, and anionic solubilizers with minimal electronic perturbation—and explored them in parallel through spawned thought channels. This is not the "generate N candidates and rank by fitness score" strategy of evolutionary optimizers; it is a structured scientific exploration in which each branch tests a specific chemical hypothesis, and the parent loop synthesizes the results.

CLIO exhibited adaptability in adjusting to runtime issues during execution. When rate limits on the underlying LLM interrupted a planned 40–60 member library enumeration, CLIO diagnosed the constraint and progressively reduced its batch sizes, ultimately converging on a six-candidate micro-batch with the explicit annotation "micro-batches only. . . no literature lookups. . . keep prompt short." A model dedicated to synthesis condition prediction, intended to augment the retrosynthesis prediction tool, was inadvertently inaccessible during the design trajectory. Rather than abandoning this component of synthetic feasibility assessment, CLIO leveraged the internal knowledge of the LLM to approximate synthetic conditions given the reaction class provided by Retrochimera. In each case, the agent recognized a failure, characterized its scope, and adapted its strategy to continue making progress—behavior documented in the CLIO conversation logs.

S2.3 Epistemic framing of mechanistic hypotheses

When presented with the split oxidation wave of compound **3**, the three independent CLIO analyses did not simply propose a mechanism; they explicitly separated what the CV data could support from what remained conjecture. Each analysis produced a tiered hypothesis ranking: reduced-state acid–base speciation and phosphonate–K^+ ion-pairing as Tier 1 (most consistent with the data, least committal), base-promoted rearrangement as Tier 2, and covalent OH^- addition or N–N scission as Tier 3 (chemically plausible but requiring product-identification experiments to confirm). Critically, CLIO identified the specific experiments that would discriminate between tiers—scan-rate dependence of the anodic charge partition, cation-swap series, cathodic vertex hold-time—and stated that without these, the higher-tier assignments should be treated as working hypotheses, not established mechanisms. The sulfonate recommendation emerged from this framework: if phosphonate–K^+ ion-pairing is a primary contributor to reduced-state microstate heterogeneity (Tier 1), then replacing the

phosphonate with a sulfonate—a weaker ion-pairing, more uniformly solvated anion—should suppress the splitting. This reasoning chain, from tiered mechanistic hypothesis to falsifiable prediction to structural modification, is what differentiates the recommendation from a generic "try sulfonate" suggestion that any electrochemist familiar with the ORFB literature might make.

S2.4 Limitations

Calibrated deference is not the same as infallibility. CLIO failed to catch the retrosynthesis tool's substitution of a quinoline for the benzo[c]cinnoline core, critiquing only the side-product risk of a route that was fundamentally wrong at the scaffold level; the actual synthesis required five steps rather than the two predicted. CLIO's skepticism of the E_{red} predictor, while ultimately justified, emerged only after multiple rounds rather than from an upfront assessment of the model's domain of applicability. And the tiered epistemic framing of the mechanistic hypotheses—the clearest example of intellectual honesty in the campaign—was maintained in part because the agent was explicitly prompted to provide an explanation, not because CLIO independently recognized the need to qualify its claims. These failures suggest that the capacity for calibrated deference is present but inconsistent: the agent can reason about the reliability of its tools and the strength of its evidence, but it does not yet do so by default across all aspects of the design problem.

More broadly, the results point toward natural-language reasoning as a complement to quantitative prediction rather than a replacement for it. The numerical property predictors were essential for the initial design triage; the mechanistic reasoning that resolved the oxidation kinetics liability was inaccessible to the numerical tools. The agent's effectiveness derived from its ability to transition between these two modes as the demands of the campaign evolved—and, crucially, from its capacity to recognize when the quantitative mode was failing and to shift toward qualitative, hypothesis-driven reasoning in response.

S3 Comparison with related agentic and optimization systems

CLIO differs from related agentic systems in what each preserves at convergence. Population-based optimizers such as AlphaEvolve[25] and the tournament-style hypothesis loop of Coscientist[26] converge on a set of best candidate solutions: prior generations are pruned and their reasoning discarded once a fitter successor is found. Biomni[27] takes a different approach within the per-task setting, pairing retrieval-augmented planning with code execution to answer biomedical research questions, but its retrieval is over a curated environment rather than over the agent's own reasoning trace. CLIO instead preserves the full thematic trajectory of what has counted as "best" across rounds: the belief-state graph indexes the agent's prior conclusions—including the ones that were later revised—so the next candidate is shaped by the campaign's accumulated judgement of what has and has not worked, not by an independent fitness evaluation against the current pool.

A complementary distinction emerges when CLIO is compared with black-box and grey-box molecular optimizers such as ExLLM.[24] In the controlled experiment reported in Section S5, ExLLM uses the LLM as a genetic operator—crossover and mutation are guided by chemical intuition embedded in the model's weights, but the search dynamics remain those of an evolutionary algorithm: a fixed population is maintained, fitness is evaluated per molecule, and selection pressure is applied uniformly across the pool. CLIO operates in a fundamentally different regime. Rather than optimizing a scalar fitness function, it reasons about *why* a candidate scores well or poorly, constructs mechanistic hypotheses to explain failures, and uses those hypotheses to propose targeted structural modifications. This distinction is most visible when the optimization landscape is deceptive: ExLLM can only respond to a low fitness score by generating more variants; CLIO can diagnose whether the low score reflects a genuine property deficit or a miscalibrated predictor, and adjust its strategy accordingly—as demonstrated by its graded distrust of the E_{red} model (Section S2.1). The cost of this reasoning capacity is throughput: ExLLM evaluates hundreds of candidates per round, while CLIO evaluates tens. The benefit is that CLIO's candidates are accompanied by falsifiable rationales that a human collaborator can critique, redirect, or build upon—a property absent from population-based search.

S4 CLIO hypothesis inventory

Tables S1–S3 catalogue every hypothesis advanced by CLIO during the negolyte design campaign, together with its resolution status and the evidence that resolved it. Hypotheses are grouped into three campaign phases: the initial phosphonate design trajectory, the diagnostic investigation of the oxidation kinetic liability, and the sulfonate redesign.

Table S1 Hypotheses from the phosphonate design trajectory (Rounds 1–3).

#	Hypothesis	Status	Phase	Resolved by
1	Electron-donating substituents (e.g. –OH) shift E_{red} negative	Accepted	R1	CLIO + predictor
2	Aza insertions improve stability and shift E_{red} anodically	Accepted	R1–2	CLIO + predictor
3	Scaffold hops may yield improved candidates	Unproductive	R1	CLIO reasoning
4	Non-conjugated anionic solubilizers improve aqueous solubility with minimal E_{red} perturbation	Accepted	R1	CLIO + experiment
5	Graphormer E_{red} predictions are reliable for absolute screening	Refuted	R2	CLIO reasoning + experiment
6	π-Stacking may cause aggregation issues	Open	R3	Flagged by CLIO; echoed by concentration-dependent oxidation wave
7	Retrosynthesis tool's 2-step route for **3** is valid	Refuted	R2	Experiment (CLIO failed to catch scaffold substitution)

Table S2 Hypotheses from the diagnostic investigation of the split oxidation wave.

#	Hypothesis	Status	Test	Resolved by
8	Split oxidation wave reflects an EC/ECE mechanism with a base-dependent follow-up step	Accepted	—	CLIO reasoning (consensus across 3 analyses)
9	H1: Direct OH^- addition to reduced benzocinnoline core	Refuted	OH^- conc. series	Experiment: no [KOH] dependence in P2/P1
10	H2: Base-promoted tautomerization or proton-transfer reorganization	Open	Scan-rate series	Experiment: no scan-rate dependence
11	H3: Phosphonate–K^+ ion pairing stabilizes distinct reduced-state microenvironments	Accepted	Cation swap	Experiment: Li^+/Na^+ increase P2/P1 vs. K^+
12	Slow chemical follow-up step would produce scan-rate–dependent P2/P1	Refuted	Scan-rate series	Experiment
13	Cation identity should affect P2/P1 if ion pairing is operative	Confirmed	Cation swap	Experiment
14	[KOH] should affect P2/P1 if OH^- is directly involved	Refuted	OH^- conc. series	Experiment

Table S3 Hypotheses from the sulfonate redesign phase.

#	Hypothesis	Status	Test	Resolved by
15	Replacing phosphonate with sulfonate will suppress split oxidation by weakening cation pairing	Accepted	CV of **20**	CLIO prediction + experiment
16	Sulfonate will maintain improved E_{red} relative to parent scaffold	Accepted	CV of **20**	Experiment
17	Irreversible spectral changes during CV would indicate decomposition	Refuted	Spectroelectrochemistry	Experiment: clean, reversible spectral changes

S5 CLIO for strictly numerical optimization

While the ability to provide a heterogeneous task definition was a primary motivation in the development of CLIO, we were nonetheless interested in quantifying its performance on a strictly numerical molecular optimization task. To this end, we constructed a controlled experiment comparing CLIO to ExLLM, an evolutionary algorithm that uses LLMs as genetic operators (crossover and mutation) and achieves state-of-the-art performance on the Practical Molecular Optimization (PMO) benchmark.[24, 28] This comparison aims to answer two questions: (1) how does CLIO perform when reduced to a black-box optimizer, and (2) how much does performance improve when domain knowledge is reintroduced?

S5.1 Experimental setup

We constructed a shared oracle endpoint that scores molecules on five properties relevant to negolyte design: aqueous solubility (logS, via a Graphormer ML model), reduction potential (E_{red}, via a Graphormer ML model), synthetic accessibility (SA score, via RDKit), Tanimoto similarity to the benzo[c]cinnoline parent scaffold (Morgan fingerprints, radius 2, 2048 bits), and a SMARTS substructure filter for reactive or toxic functional groups. A fitness function was defined to produce a strictly numerical reproduction of the multi-fidelity negolyte optimization goal presented to CLIO in our primary experiment. Each property x_i was mapped to a fitness score $f_i \in [0, 1]$ through continuous transformations:

$$f_{\text{logS}}(x) = \sigma(x;\, \mu{=}{-}1.5,\, k{=}1.5) = \frac{1}{1 + e^{-k(x-\mu)}} \tag{S1}$$

$$f_{E_{\text{red}}}(x) = \mathcal{G}(x;\, \mu{=}{-}1.0,\, \sigma{=}0.5) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{S2}$$

$$f_{\text{SA}}(x) = 1 - \text{clip}\left(\frac{x-1}{6-1},\, 0,\, 1\right) \tag{S3}$$

$$f_{\text{sim}}(x) = \text{clip}\left(\frac{x-0.2}{0.5-0.2},\, 0,\, 1\right) \tag{S4}$$

The SMARTS substructure filter acted as a hard gate ($f_{\text{SMARTS}} = 0$ if any flagged substructure is present, 1 otherwise). The composite fitness was:

$$F = \begin{cases} \left(f_{\text{logS}} \cdot f_{E_{\text{red}}} \cdot f_{\text{SA}} \cdot f_{\text{sim}}\right)^{1/4} & \text{if } f_{\text{SMARTS}} = 1 \\ -10 & \text{otherwise} \end{cases} \tag{S5}$$

This aggregation strongly penalizes imbalanced molecules, forcing genuine multi-property optimization rather than over-fitting to a single objective.

Both algorithms were given identical conditions: a budget of 500 oracle evaluations, the same set of 10 seed molecules (simple substitution-based derivatives of

the benzo[c]cinnoline scaffold, pre-scored by the oracle), and identical oracle endpoints. GPT-5.2 was the underlying LLM. Six independent replicas were run for each condition.

S5.2 Black-box vs. grey-box conditions

We tested each algorithm under two information regimes:

S5.2.1 Black-box

Under black-box conditions, both algorithms were given the property names (logS, E_{red}, SA score, similarity, SMARTS filter) and general chemical principles for modifying them (e.g., "polar groups improve solubility," "electron-withdrawing groups shift reduction potential"). No information about target values, scoring thresholds, the functional forms of the fitness transformations, or the composite aggregation function was provided. The algorithm received only the five property fitness scores ($[0, 1]$, higher is better) and the composite fitness for each evaluated molecule, and had to optimize the composite as an opaque numerical objective.

S5.2.2 Grey-box

Under grey-box conditions, the algorithms additionally received approximate target values (e.g., "E_{red} near $-1.0\,\mathrm{V}$ is optimal," "logS above -1.5 scores well"), the identity of the composite as a geometric mean, and the scoring thresholds for each property (e.g., "Tanimoto ≥ 0.5 is full score, below 0.2 is zero"). The exact functional forms and parameters of the scoring transformations (sigmoid steepness, Gaussian width) were not disclosed. This mirrors the level of domain knowledge a practicing chemist would bring to the task: awareness of what "good" looks like for each property, without knowing the precise mathematical mapping.

For CLIO, the grey-box information was provided in the system prompts. For ExLLM, the grey-box information was incorporated into the prompt templates that guide the LLM's crossover and mutation operations.

S5.3 CLIO experimental arms

Within each information regime, we tested framings of the design goal for CLIO:

S5.3.1 Baseline

The system prompt provides general molecular design instructions with no specific guidance on search strategy beyond "balance exploitation with exploration."

S5.3.2 Efficiency

The system prompt additionally emphasizes convergence speed and structured search as goals: an aggressive Top-1 search phase (first $\sim$200 evaluations) using parallel hypothesis testing is suggested, followed by focused refinement around the best-scoring scaffold.

S5.4 Results

Table S4 summarizes the final Top-1 and mean Top-10 composite fitness scores across all experimental conditions. Figure S2 shows the convergence profiles under black-box conditions, and Figure S3 shows the corresponding profiles under grey-box conditions.

Table S4 Composite fitness scores across experimental conditions ($n = 6$ replicas per condition, 500 oracle evaluations per replica). Values are reported as mean ± standard deviation.

Condition	Algorithm	Top-1	Mean Top-10
Black-box	ExLLM	0.586 ± 0.032	0.525 ± 0.011
	CLIO	0.544 ± 0.065	0.504 ± 0.059
	CLIO (efficiency)	0.541 ± 0.062	0.503 ± 0.049
Grey-box	ExLLM	0.607 ± 0.025	0.543 ± 0.008
	CLIO	0.599 ± 0.095	0.578 ± 0.103
	CLIO (efficiency)	0.602 ± 0.066	0.567 ± 0.061

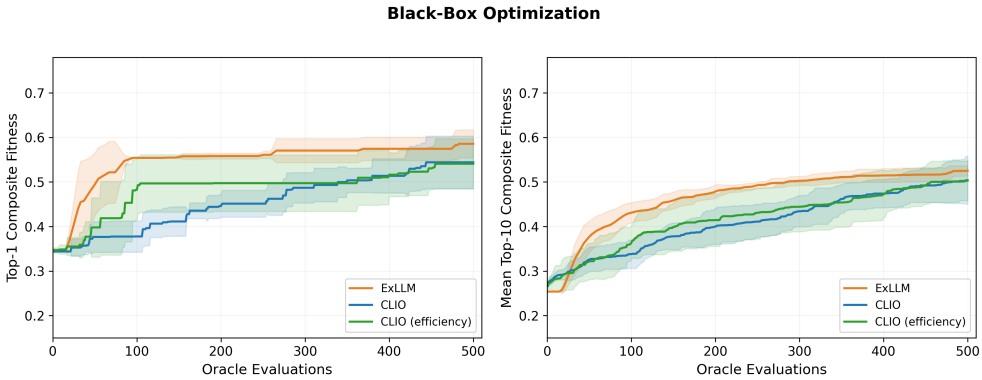


Fig. S2 Convergence profiles under black-box conditions. Solid lines show the mean across six replicas; shaded regions indicate ±1 standard deviation. Left: Top-1 (best molecule found so far) vs. oracle evaluations. Right: mean Top-10 fitness vs. oracle evaluations.

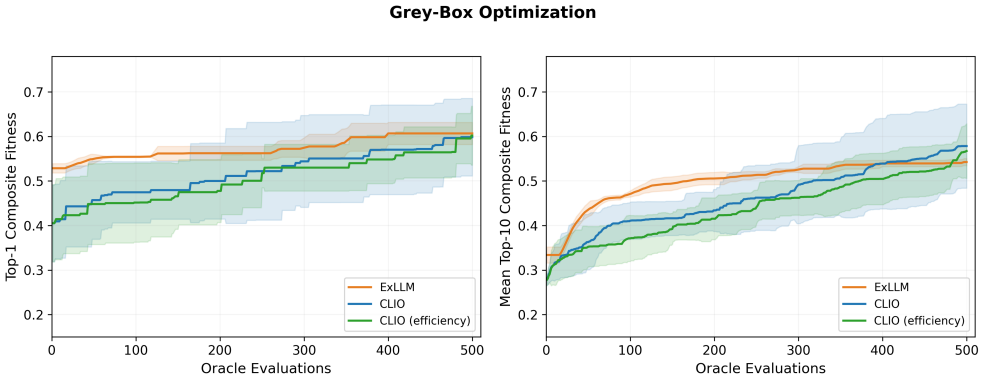


Fig. S3 Convergence profiles under grey-box conditions. Solid lines show the mean across six replicas; shaded regions indicate ±1 standard deviation. Left: Top-1 (best molecule found so far) vs. oracle evaluations. Right: mean Top-10 fitness vs. oracle evaluations.

S5.4.1 Black-box performance

Under black-box conditions, ExLLM outperformed both CLIO configurations on Top-1 fitness (0.586±0.032) with substantially lower variance than CLIO (0.544±0.065). This result is expected: ExLLM is purpose-built for black-box multi-objective molecular optimization, employing NSGA-II selection with LLM-guided crossover and mutation operations that are specifically designed to explore chemical space efficiently without domain knowledge. CLIO, by contrast, is a general-purpose reasoning agent whose architecture—cognitive loops, thought channels, belief-state management—introduces overhead that does not benefit a purely numerical optimization task. Instructing CLIO to prioritize sample efficiency in the initial evaluations notably shifted the slope of the initial discovery curve, but it ultimately did not improve CLIO's black-box performance (0.541 ± 0.062) over the complete oracle budget.

S5.4.2 Grey-box performance

Introducing domain knowledge around the optimization targets improved both algorithms, but the effect was more pronounced for CLIO. CLIO's Top-1 fitness increased by 10% (from 0.544 to 0.599), while ExLLM's improved by 3.6% (from 0.586 to 0.607). The improvement in mean Top-10 was even more striking: CLIO improved by 15% ($0.504 \rightarrow 0.578$), compared to 3.4% for ExLLM ($0.525 \rightarrow 0.543$). Under grey-box conditions, CLIO's Top-1 performance approached that of ExLLM (0.599 vs. 0.607), and CLIO outperformed ExLLM on mean Top-10 (0.578 vs. 0.543). Interestingly, inclusion of the sample efficiency instructions had no meaningful effect (0.602 ± 0.066 Top-1, 0.567 ± 0.061 Top-10), suggesting that the exploitation strategies CLIO leveraged to improve sample efficiency on the black-box problems were rooted in its domain knowledge.

S5.4.3 Interpretation

These results highlight a fundamental distinction between the two approaches. ExLLM treats molecular optimization as a search problem, applying evolutionary operators to navigate chemical space efficiently regardless of whether the fitness landscape is interpretable. Its performance is relatively insensitive to the inclusion of domain knowledge because its search strategy does not require it: NSGA-II selection and LLM-guided recombination are effective purely from the pattern of fitness scores.

CLIO, by contrast, treats molecular optimization as a reasoning problem. When given only opaque fitness scores, its cognitive architecture—designed for hypothesis generation, causal reasoning, and multi-step planning—has little to reason about, and its performance lags behind a dedicated optimizer. When provided with domain knowledge (what properties measure, what targets to aim for, how the composite function aggregates scores), CLIO can leverage its reasoning capabilities to make chemically informed design decisions: diagnosing which property is the bottleneck on a leading candidate, reasoning about which structural modifications would address that bottleneck, and planning batches that systematically test these hypotheses. In this scenario, CLIO gains more from the availability of domain knowledge, ultimately producing designs that meet or exceed those of the dedicated optimizer.

S6 Experimental characterization of ExLLM structures

ExLLM was used to design structures with a configuration comparable to the grey-box scenario described above with some differences. At the time these designs were performed, GPT-4o was used as the underlying LLM, the composite fitness function was defined as the arithmetic mean of each property's fitness, and a budget of 2500 oracle evaluations was afforded. Three structures designed by ExLLM were selected by chemists from among those with the highest composite fitness scores on the basis of their apparent synthetic tractability (Figure S4). These structures were synthesized and electrochemically characterized, giving some insight into the effect of problem formulation on experimental outcomes.

Table S5 summarises the electrochemical data alongside the CLIO-designed candidates. All three ExLLM compounds were only partially soluble at 1 mM in 1 M KOH, whereas the CLIO candidates dissolved readily at concentrations exceeding 50 mM. This differential is attributable to CLIO's ability to consider natural-language modifications to the design objective and adjust its use of the logS metric accordingly. The CLIO designs use acidic solubilizing groups that will be deprotonated in the KOH solution. The ExLLM designs, on the other hand, rely on protonation of amines for solubilization, a strategy that may be more effective in the neutral-pH regime directly predicted by logS.

Electrochemically, **S1** and **S2** show a small shift from **1** toward more negative E_{red}, but neither reaches the magnitude of shift of **3** or **20**. This differential appears to be a direct result of CLIO's recognition of the systematic misestimation of reduction potential for benzo[c]cinnoline structures by the provided model and its shift to using that model in a calibrated fashion. With this approach, CLIO has the leeway to push

molecules to more extreme predicted potentials in a way that the fixed numerical objective of ExLLM cannot.

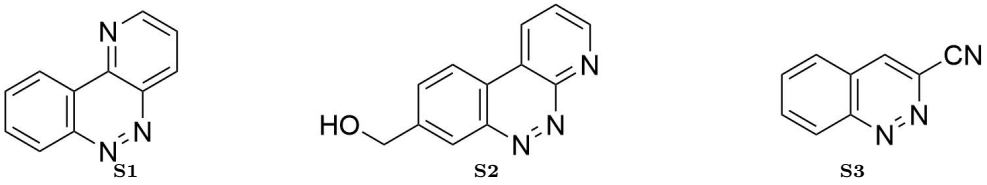
ExLLM-designed compounds **S1** and **S2** do show electrochemical reversibility, with a peak-current ratio exceeding that of the originally reported mixed sulfonate **1**. This is worth noting, although the ExLLM compounds are not perfect on this account, as scaffold-hop structure **S3** shows a ∼1300 mV reduction-to-oxidation peak separation, making it unsuitable for use in ORFB applications.

Table S5 Electrochemical characterization of ExLLM-designed compounds, compared with the baseline BzC scaffold and CLIO-designed candidates. Measured values from cyclic voltammograms at 50 $\mathrm{mV\,s^{-1}}$ in 1 M KOH (1 mM analyte concentration, 5 mm glassy carbon electrode) unless otherwise noted.

	1[*]	**3**	**20**	**S1**	**S2**	**S3**
E_{red} predicted (V vs. SHE)	−1.60	−2.04	−2.00	−1.87	−1.84	−1.70
E_{red} measured (V vs. SHE)	−0.762	−0.895	−0.854	−0.774	−0.794	−0.987
Solubility (pH 7) predicted (mM)	52.5	20.4	14.5	4.27	6.03	4.90
Solubility (0.5 M KOH) (mM)	500[#]	57.9	56.0	<1[^]	<1[^]	<1[^]
$\lvert i_{\mathrm{ox}}/i_{\mathrm{red}}\rvert$[@]	0.52	0.18	0.38	0.65	0.65	0.48[†]
$Q_{\mathrm{ox}}/Q_{\mathrm{red}}$[‡]	0.79	0.92	0.84	0.79	1.24	0.42[†]

[*] Predicted values for the ortho/ortho-substituted component.
[#] Literature value from [8].
[@] Ratio of the anodic peak current nearest the reduction wave to the cathodic peak current.
[^] Cyclic voltammetry samples were partially soluble at 1 mM in 1 M KOH.
[†] Peak separation of ∼1300 mV indicates electrochemical irreversibility.
[‡] Charge ratio by semi-integration of the baseline-corrected voltammogram.

Cyclic voltammograms of each ExLLM-designed compound were recorded at 25, 50, and 100 $\mathrm{mV\,s^{-1}}$ in 1 M KOH, pH 7 phosphate buffer, and 1 M H_2SO_4 (Figures S5–S7).



S1 **S2** **S3**

Fig. S4 Structures designed by ExLLM and selected for synthesis on the basis of composite fitness score and synthetic tractability.

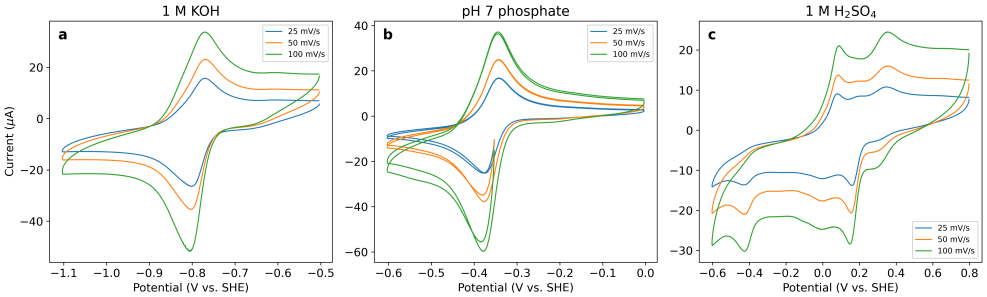


Fig. S5 Cyclic voltammograms of **S1** (1 mM) in (a) 1 M KOH, (b) pH 7 phosphate buffer, and (c) 1 M H_2SO_4 at 25, 50, and 100 mV s^{-1}. Potentials are referenced to SHE.

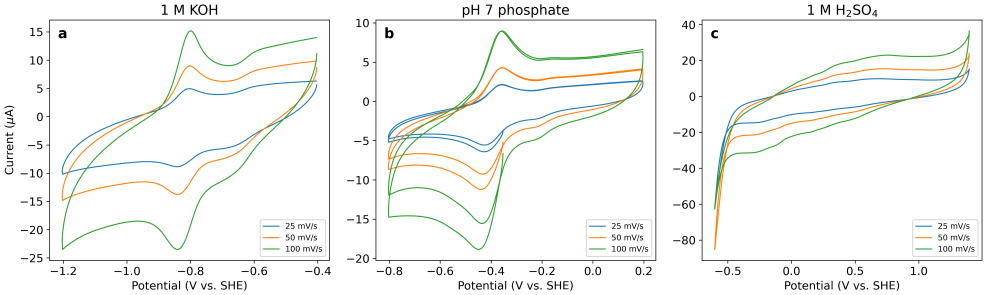


Fig. S6 Cyclic voltammograms of **S2** (1 mM) in (a) 1 M KOH, (b) pH 7 phosphate buffer, and (c) 1 M H_2SO_4 at 25, 50, and 100 mV s^{-1}. Potentials are referenced to SHE.

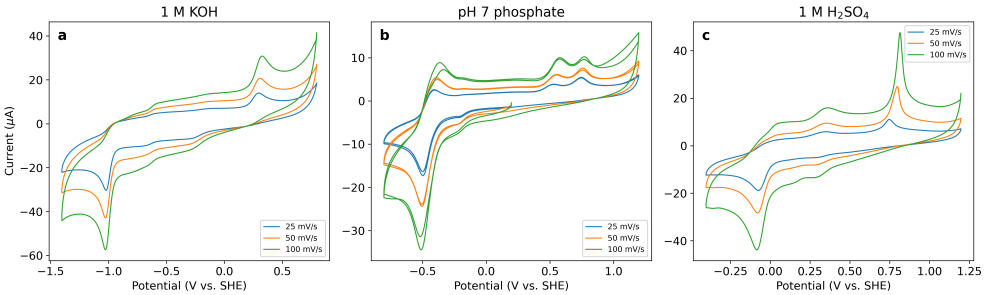


Fig. S7 Cyclic voltammograms of **S3** (1 mM) in (a) 1 M KOH, (b) pH 7 phosphate buffer, and (c) 1 M H_2SO_4 at 25, 50, and 100 mV s^{-1}. Potentials are referenced to SHE.

Figure S8 shows the tangent-baseline half-peak potential analysis for each compound at 50 mV s^{-1} in 1 M KOH.

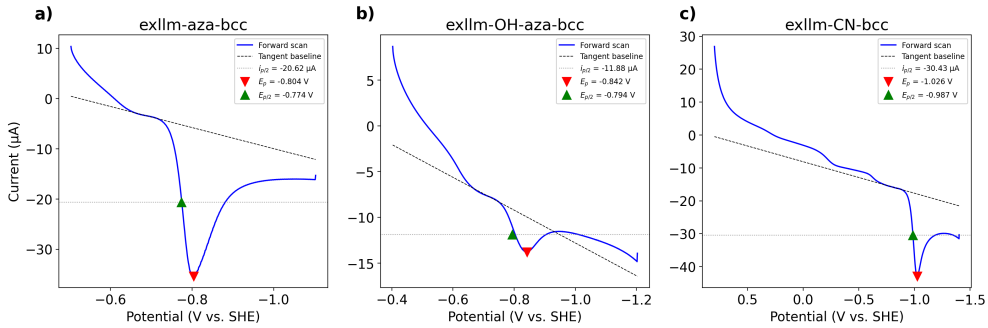


Fig. S8 Half-peak potential analysis for ExLLM-designed compounds at 50 mV s^{-1} in 1 M KOH: (a) **S1**, (b) **S2**, and (c) **S3**. The tangent baseline (dashed) is fit to the pre-wave capacitive region; the half-peak potential $E_{p/2}$ (green triangle) is determined at the half-height of the baseline-corrected peak current.

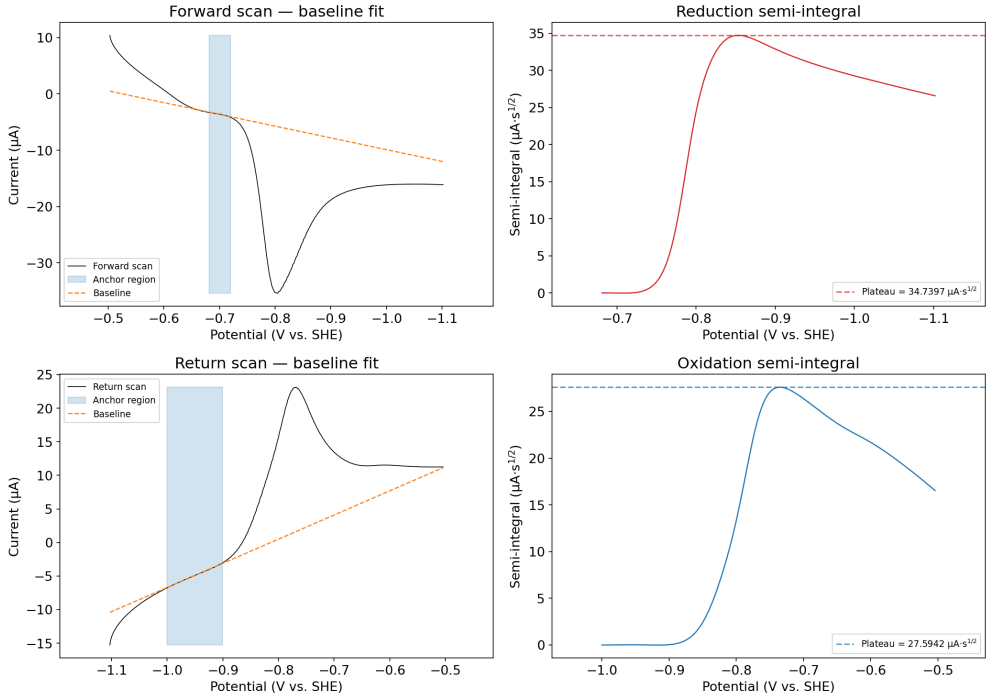


Fig. S9 Semi-integration analysis for **S1** (1 mM, 1 M KOH, 50 mV s^{-1}). Top: forward scan baseline fit (left) and reduction semi-integral (right). Bottom: return scan baseline fit (left) and oxidation semi-integral (right).

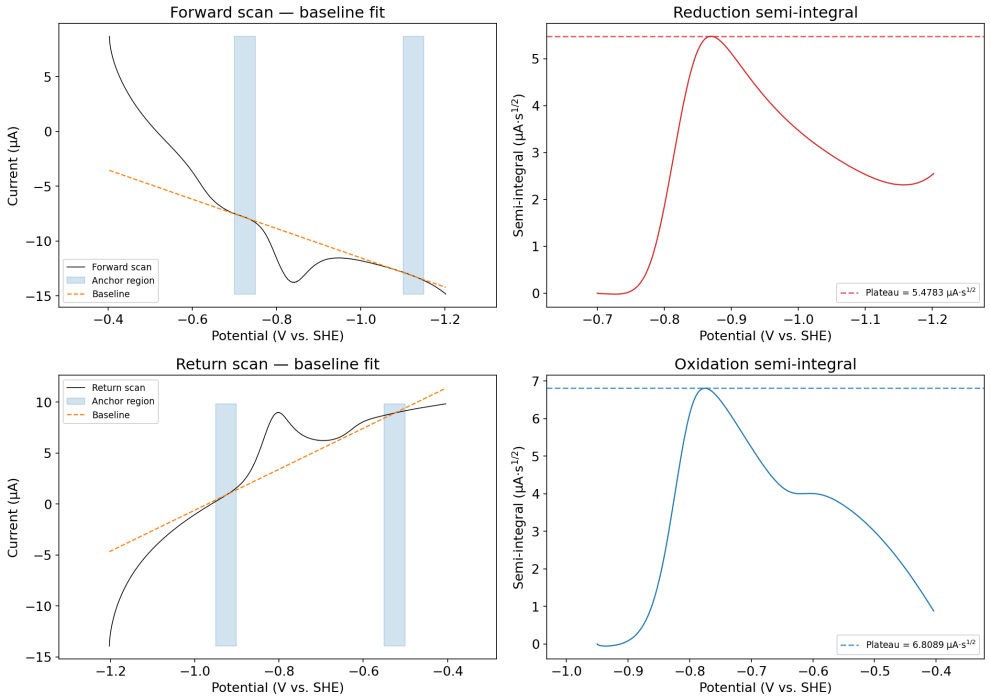


Fig. S10 Semi-integration analysis for **S2** (1 mM, 1 M KOH, 50 mV s^{-1}). Top: forward scan baseline fit (left) and reduction semi-integral (right). Bottom: return scan baseline fit (left) and oxidation semi-integral (right).

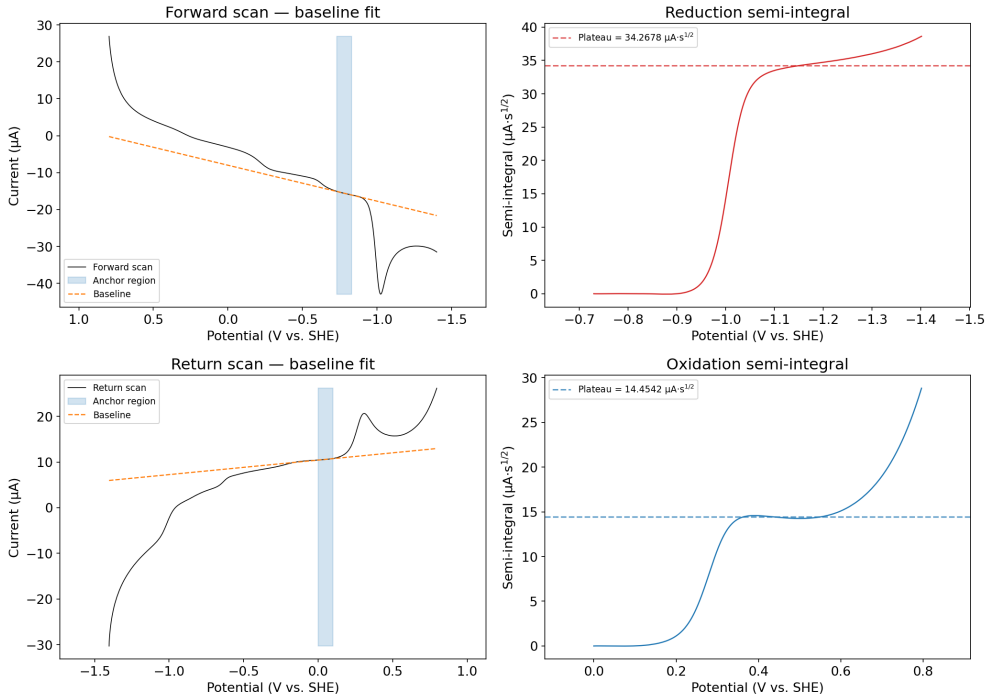


Fig. S11 Semi-integration analysis for **S3** (1 mM, 1 M KOH, 50 mV s^{-1}). Top: forward scan baseline fit (left) and reduction semi-integral (right). Bottom: return scan baseline fit (left) and oxidation semi-integral (right).

S7 Electrochemical characterization

S7.1 Materials and reagents list

All chemicals were used as received unless otherwise stated. Electrolyte solutions were prepared with deionized water (18 MΩ·cm). Ethyl alcohol (200 proof), potassium hydroxide flakes (reagent grade 90%), sodium phosphate monobasic (>99%), sodium phosphate dibasic (>99%), and deuterium oxide (99%) were purchased from Sigma-Aldrich. Sodium hydroxide, potassium ferricyanide (>99%), potassium ferrocyanide (>99%), and sulfuric acid (96%) were purchased from Fisher Scientific. Sodium methanesulfonate (99%) was purchased from Acros Organics.

Glassy carbon working electrodes (3.0 mm diameter), Ag/AgCl reference electrodes (3M KCl), and platinum wire counter electrodes were purchased from BASi Inc. Cyclic voltammetry was performed on a CHI7013E potentiostat. Spectroelectrochemistry measurements were performed with a Honeycomb Spectrochemical with either gold or platinum as the electrode (Pine Research).

S7.2 Cyclic voltammetry

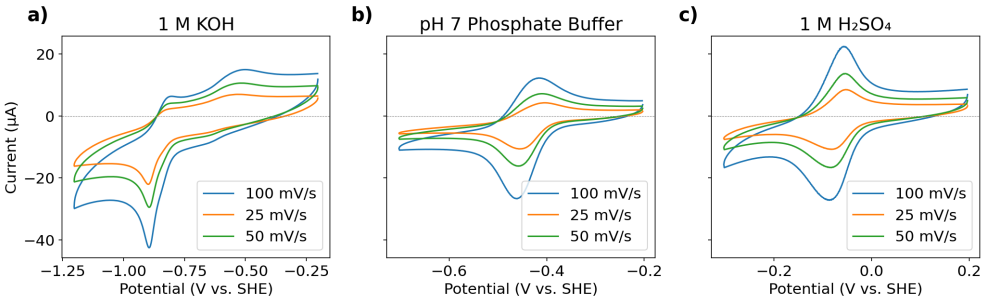


Fig. S12 Cyclic voltammetry of compound **3** at three scan rates in (a) 1 M KOH, (b) pH 7 phosphate buffer, and (c) 1 M H_2SO_4.

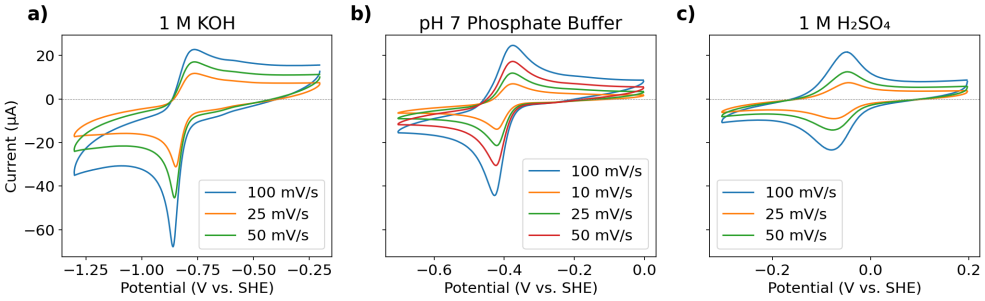


Fig. S13 Cyclic voltammetry of compound **20** at three scan rates in (a) 1 M KOH, (b) pH 7 phosphate buffer, and (c) 1 M H_2SO_4.

S7.3 Half-peak potential calculations

Half-peak potentials ($E_{p/2}$) were determined by first establishing a tangent baseline by fitting a first-order polynomial to the current in the −0.75 to −0.60 V vs. SHE region immediately preceding the onset of the Faradaic reduction wave. The baseline-corrected peak current, $i_{p,\mathrm{corr}}$, was obtained by subtracting the extrapolated baseline current at the peak potential (E_p) from the measured peak current. The half-peak current was defined as $i_{p/2} = i_{p,\mathrm{corr}}/2$ (referenced to the baseline), and $E_{p/2}$ was determined by linear interpolation of the forward-scan current to the point at which the baseline-corrected current equalled $i_{p/2}$. Where minimal variance was observed in the measured half-peak potential across scan rates, that determined at 50 mV/s was reported.

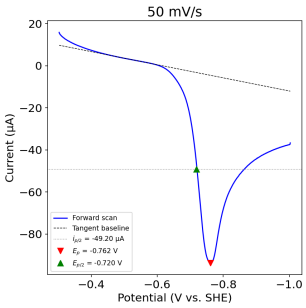


Fig. S14 Diagnostic plot for the half-peak potential determination of BzC in 0.5 M KOH at 50 mV/s; data from [8]. The dashed line shows the tangent baseline fit to the pre-wave region; the dotted horizontal line indicates $i_{p/2}$. Red and green markers denote E_p and $E_{p/2}$, respectively.

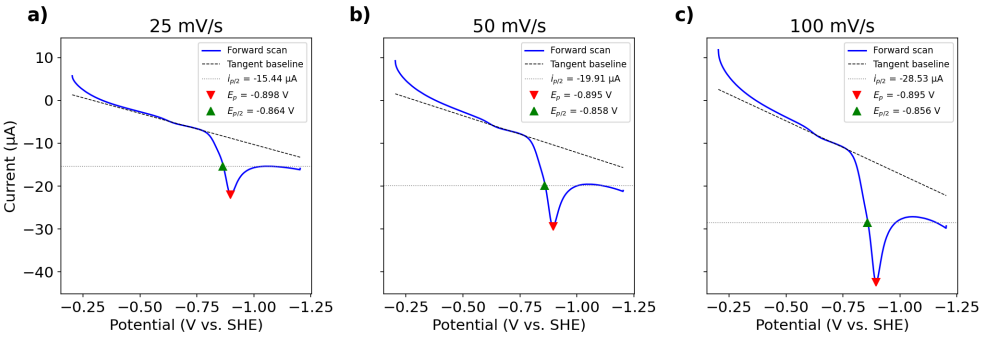


Fig. S15 Diagnostic plots for the half-peak potential determination of compound **3** in 1 M KOH at (a) 25 mV/s, (b) 50 mV/s, and (c) 100 mV/s. The dashed line shows the tangent baseline fit to the pre-wave region; the dotted horizontal line indicates $i_{p/2}$. Red and green markers denote E_p and $E_{p/2}$, respectively.

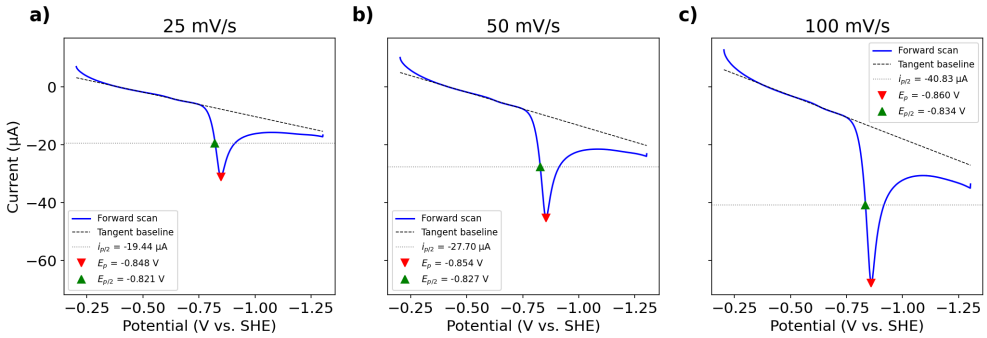


Fig. S16 Diagnostic plots for the half-peak potential determination of compound **20** in 1 M KOH at (a) 25 mV/s, (b) 50 mV/s, and (c) 100 mV/s. The dashed line shows the tangent baseline fit to the pre-wave region; the dotted horizontal line indicates $i_{p/2}$. Red and green markers denote E_p and $E_{p/2}$, respectively.

S7.4 Semi-integration charge ratio analysis

Charge ratios ($Q_{\mathrm{ox}}/Q_{\mathrm{red}}$) were determined by semi-integration (convolution voltammetry) of the baseline-corrected cyclic voltammograms. A linear baseline was fit to an anchor region on the capacitive background of each half-cycle and extrapolated across the full scan; the baseline-corrected current was then semi-integrated starting from the anchor region. For a diffusion-controlled process, the semi-integral plateau is

$$m_{\mathrm{lim}} = nFAC\sqrt{D} \tag{S6}$$

where n is the number of electrons, F is the Faraday constant, A is the electrode area, C is the bulk concentration, and D is the diffusion coefficient. Because n, F, A, C, and D are identical for the reduction and re-oxidation of the same species at the same electrode, the ratio of plateaus gives

$$\frac{Q_{\mathrm{ox}}}{Q_{\mathrm{red}}} = \frac{m_{\mathrm{lim,ox}}}{m_{\mathrm{lim,red}}} \tag{S7}$$

directly, without requiring absolute charge values.

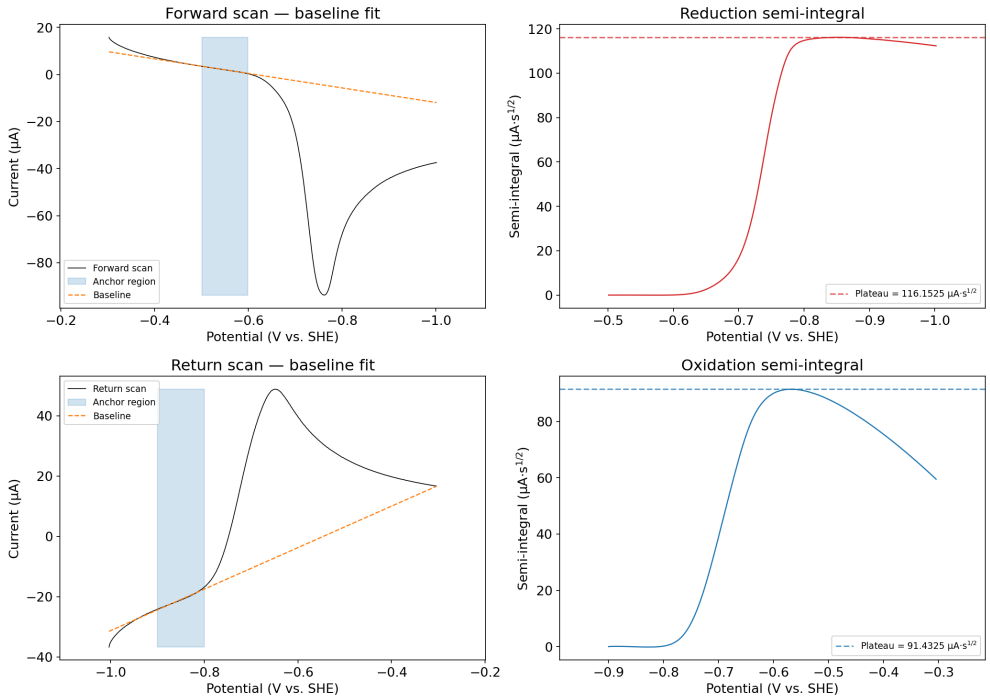


Fig. S17 Semi-integration analysis for **1** (5 mM, 0.5 M KOH, 50 mV s^{-1}). Top: forward scan baseline fit (left) and reduction semi-integral (right). Bottom: return scan baseline fit (left) and oxidation semi-integral (right). Dashed lines indicate the plateau values used to compute Q_{ox}/Q_{red}.

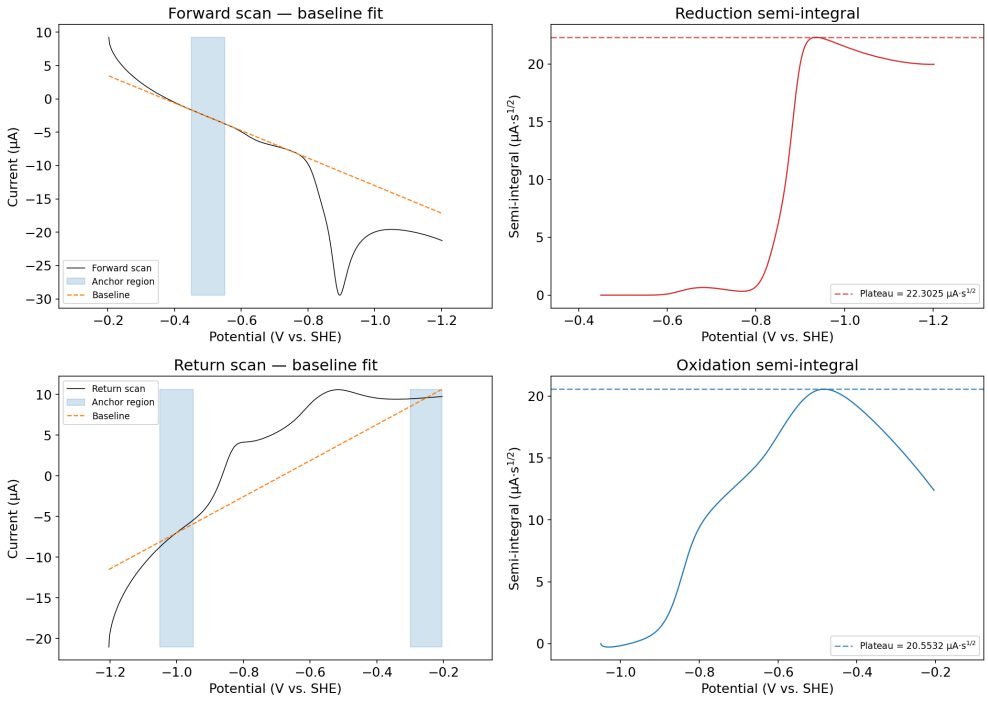


Fig. S18 Semi-integration analysis for **3** (1 mM, 1 M KOH, 50 mV s^{-1}). Top: forward scan baseline fit (left) and reduction semi-integral (right). Bottom: return scan baseline fit (left) and oxidation semi-integral (right).

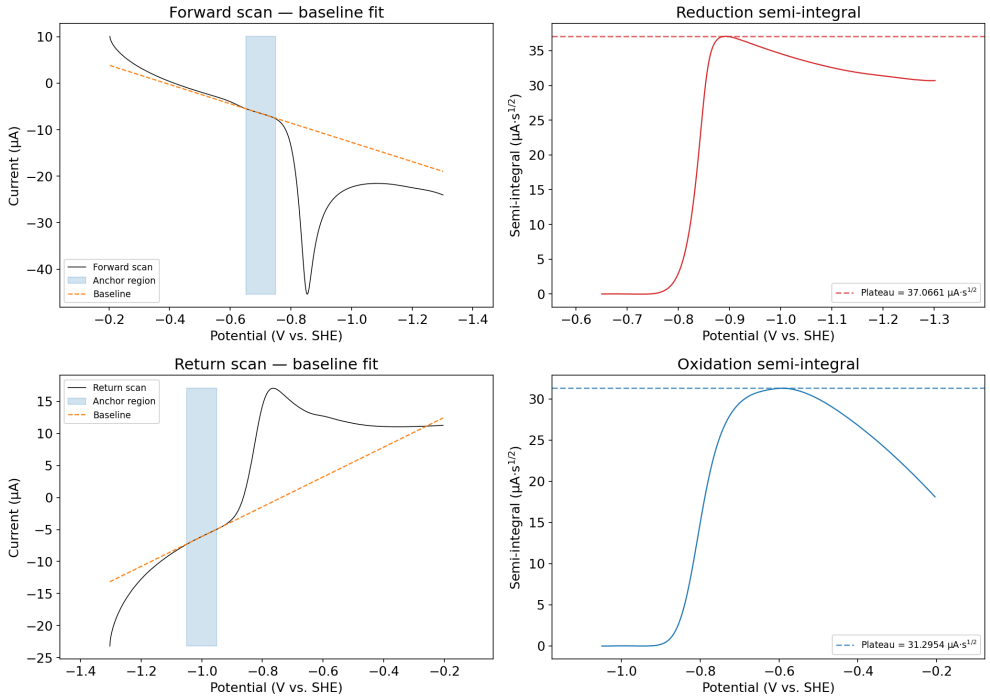


Fig. S19 Semi-integration analysis for **20** (1 mM, 1 M KOH, 50 mV s^{-1}). Top: forward scan baseline fit (left) and reduction semi-integral (right). Bottom: return scan baseline fit (left) and oxidation semi-integral (right).

S7.5 Potassium hydroxide concentration dependence

Cyclic voltammograms of 1 mM **3** were recorded at 50 mV s^{-1} in 0.1, 0.25, and 1 M KOH. A linear baseline was fitted to anchor regions on the return (anodic) scan where the current was approximately constant. The low anchor was fixed at −1.00 to −0.90 V vs. SHE; the high anchor was a 0.10 V window selected automatically by sliding from −0.40 V toward the positive scan limit and choosing the position that minimised the root-mean-square residual of the combined linear fit. Figure S20 shows the raw return scans with the fitted baselines (left) and the baseline-corrected oxidation waves with the two peak maxima indicated (right).

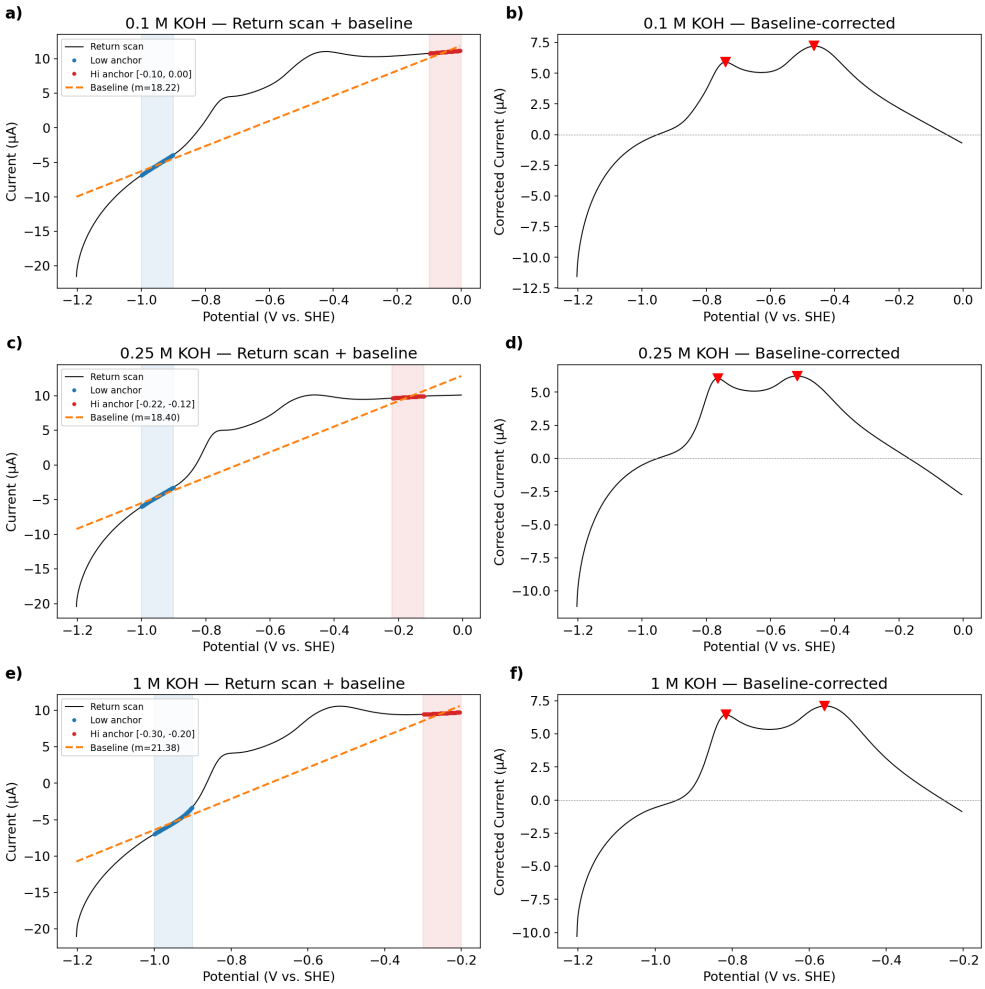


Fig. S20 Baseline correction of the anodic return scan for 1 mM **3** at 50 mV s^{-1} in 0.1 M KOH (a, b), 0.25 M KOH (c, d), and 1 M KOH (e, f). Left panels show the return scan (black) with anchor regions (shaded) and the linear baseline (dashed orange). Right panels show the baseline-corrected current with peak maxima marked by red triangles.

Table S6 Oxidation peak positions and currents from the baseline-corrected anodic scan of 1 mM **3** at 50 mV s^{-1}.

KOH	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
0.1 M	−0.740	5.90	−0.465	7.22	1.22
0.25 M	−0.764	6.04	−0.516	6.22	1.03
1 M	−0.816	6.43	−0.560	7.10	1.10

S7.6 Scan rate dependence (0.1 M KOH)

Cyclic voltammograms of 1 mM **3** were recorded at 10-500 $\mathrm{mV\,s^{-1}}$ in 0.1 M KOH. The baseline-correction procedure described in the potassium hydroxide concentration dependence experiment was used. Figure S21 shows the fitted baselines (left) and the corrected oxidation waves with peak maxima indicated (right) for each scan rate. Table S7 summarises the extracted peak positions, currents, and P2/P1 ratios.

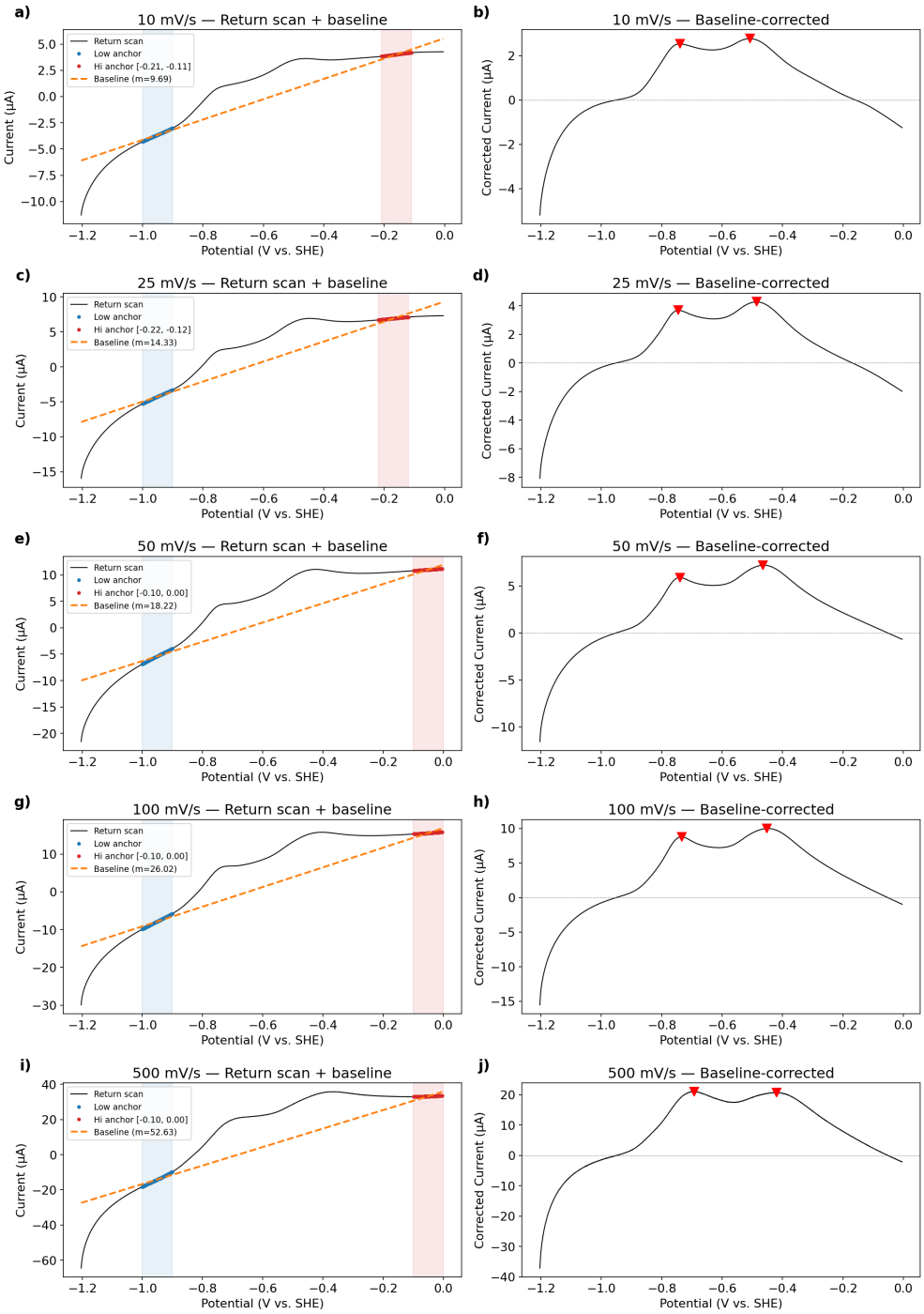


Fig. S21 Baseline correction of the anodic return scan for 1 mM **3** in 0.1 M KOH at scan rates of 10, 25, 50, 100, and 500 mV s^{-1} (top to bottom). Left panels show the return scan (black) with anchor regions (shaded) and fitted linear baseline (dashed orange). Right panels show the baseline-corrected current with peak maxima marked by red triangles.

Table S7 Oxidation peak positions and currents from the baseline-corrected anodic scan of 1 mM **3** in 0.1 M KOH at varying scan rates.

Scan rate ($mV\,s^{-1}$)	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
10	−0.740	2.54	−0.508	2.77	1.09
25	−0.745	3.69	−0.486	4.28	1.16
50	−0.740	5.90	−0.465	7.22	1.22
100	−0.734	8.79	−0.452	10.01	1.14
500	−0.692	20.96	−0.420	20.68	0.99

S7.7 Scan rate dependence (0.25 M KOH)

Cyclic voltammograms of 1 mM **3** were recorded at 10-500 $mV\,s^{-1}$ in 0.25 M KOH. The baseline-correction procedure described in the potassium hydroxide concentration dependence experiment was used. Figure S22 shows the fitted baselines and corrected oxidation waves, and Table S8 summarises the peak data.

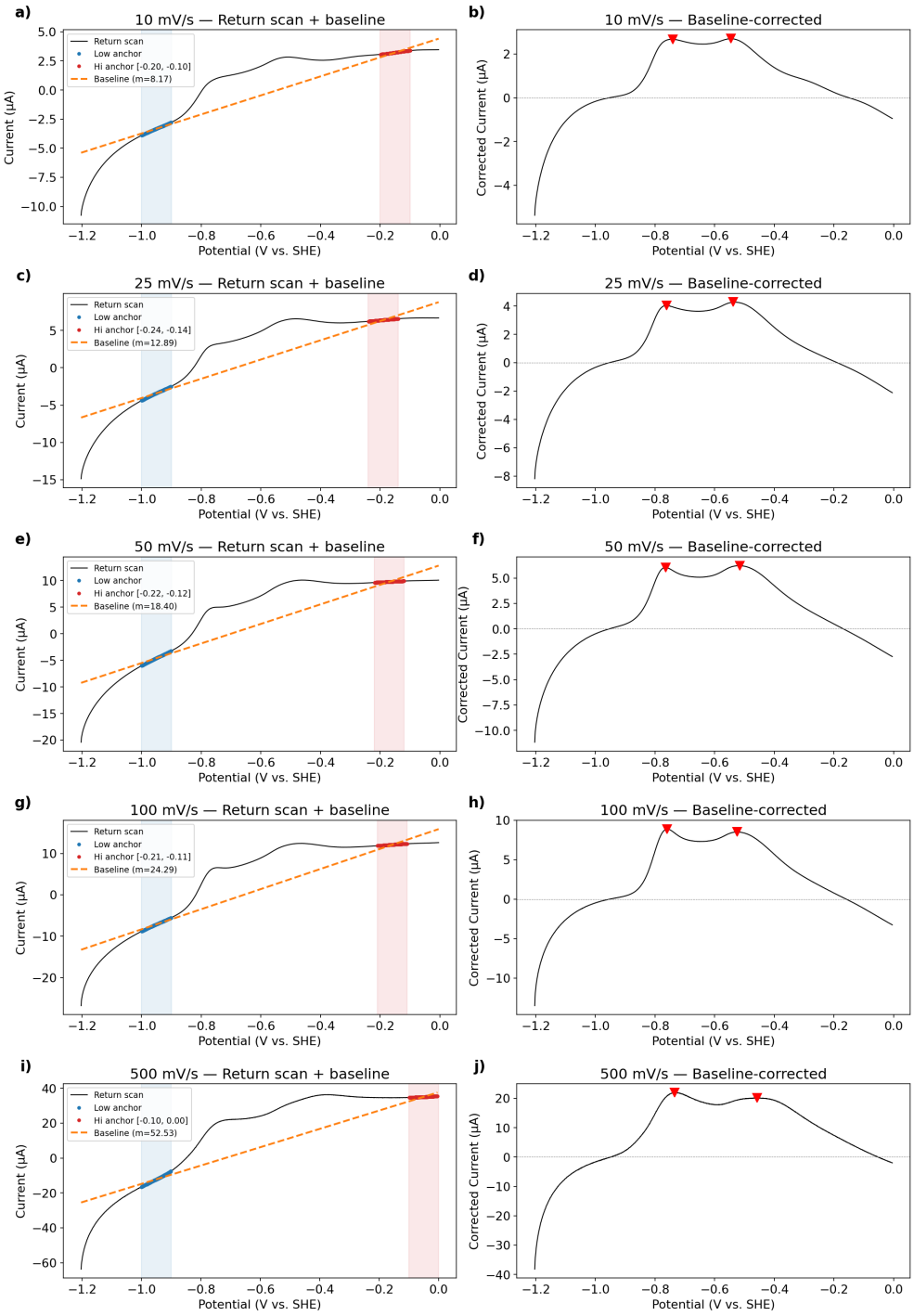


Fig. S22 Baseline correction of the anodic return scan for 1 mM **3** in 0.25 M KOH at scan rates of 10, 25, 50, 100, and 500 mV s^{-1} (top to bottom). Left panels show the return scan (black) with anchor regions (shaded) and fitted linear baseline (dashed orange). Right panels show the baseline-corrected current with peak maxima marked by red triangles.

Table S8 Oxidation peak positions and currents from the baseline-corrected anodic scan of 1 mM **3** in 0.25 M KOH at varying scan rates.

Scan rate ($mV\,s^{-1}$)	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
10	−0.740	2.67	−0.545	2.70	1.01
25	−0.761	4.03	−0.537	4.27	1.06
50	−0.764	6.04	−0.516	6.22	1.03
100	−0.759	8.88	−0.524	8.53	0.96
500	−0.734	21.94	−0.457	20.14	0.92

S7.8 Scan rate dependence (1 M KOH)

Cyclic voltammograms of 1 mM **3** were recorded at 10-500 $mV\,s^{-1}$ in 1 M KOH. The baseline-correction procedure described in the potassium hydroxide concentration dependence experiment was used. Figure S23 shows the fitted baselines and corrected oxidation waves, and Table S9 summarises the peak data.

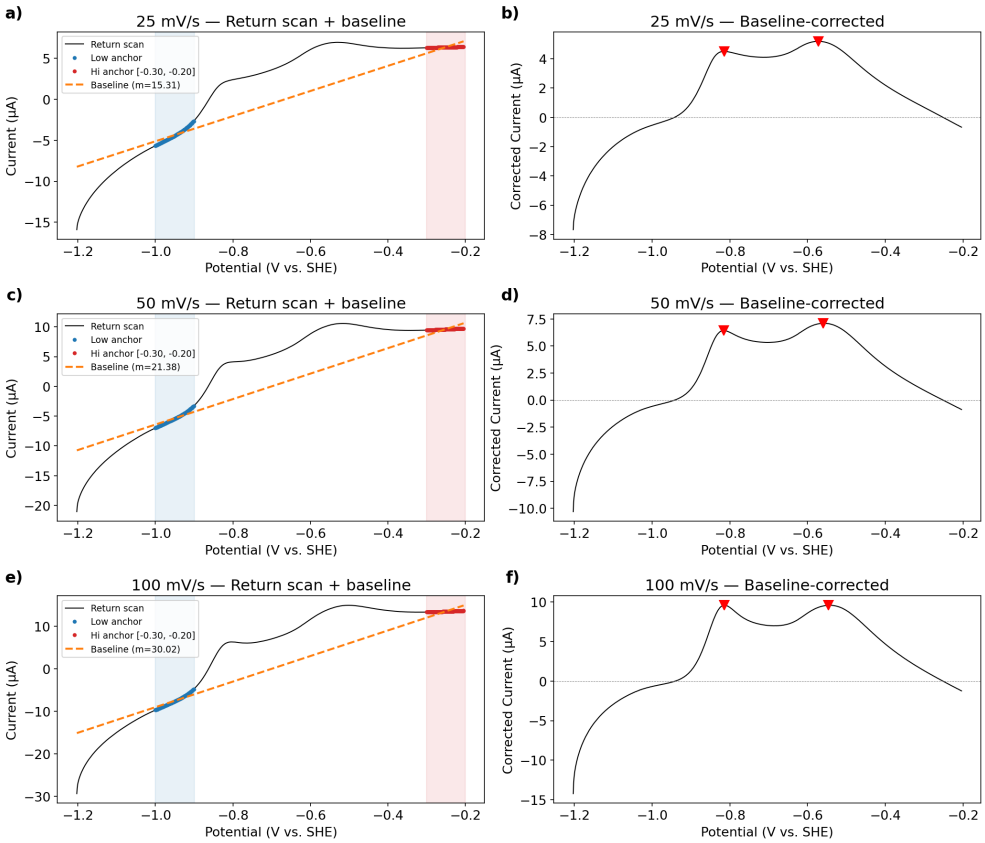


Fig. S23 Baseline correction of the anodic return scan for 1 mM **3** in 1 M KOH at scan rates of 25, 50, and 100 mV s^{-1} (top to bottom). Left panels show the return scan (black) with anchor regions (shaded) and fitted linear baseline (dashed orange). Right panels show the baseline-corrected current with peak maxima marked by red triangles.

Table S9 Oxidation peak positions and currents from the baseline-corrected anodic scan of 1 mM **3** in 1 M KOH at varying scan rates.

Scan rate (mV s^{-1})	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
25	−0.815	4.51	−0.572	5.19	1.15
50	−0.816	6.43	−0.560	7.10	1.10
100	−0.815	9.60	−0.546	9.60	1.00

S7.9 KOH concentration dependence at additional scan rates

Following the matrix design of experiments, the baseline-correction analysis described in the potassium hydroxide concentration dependence experiment was performed at 25 and 100 mV s^{-1} across 0.1, 0.25, and 1 M KOH. The baseline-correction procedure

described in the potassium hydroxide concentration dependence experiment was used. Figure S24 and Figure S25 show the fitted baselines and corrected oxidation waves. Table S8 and Table S9 summarize the peak data.

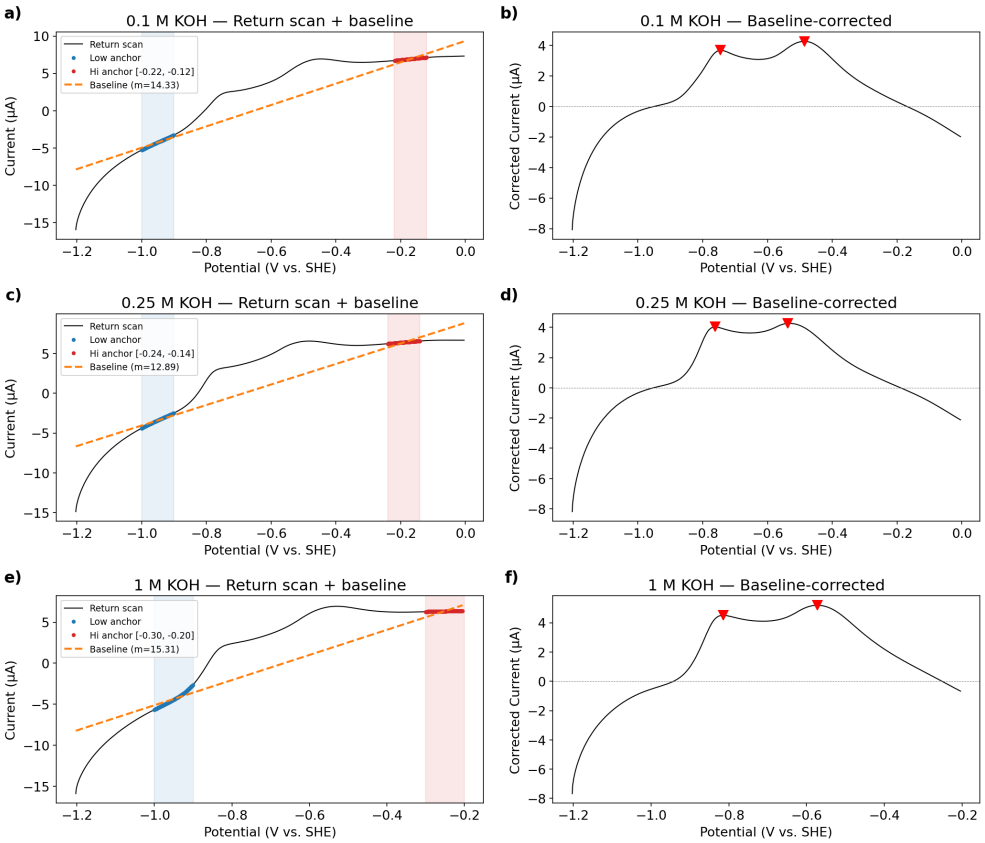


Fig. S24 Baseline correction of the anodic return scan for 1 mM **3** at 25 mV s^{-1} in 0.1 M KOH (a, b), 0.25 M KOH (c, d), and 1 M KOH (e, f).

Table S10 Oxidation peak data at 25 mV s^{-1} across KOH concentrations.

KOH	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
0.1 M	−0.745	3.69	−0.486	4.28	1.16
0.25 M	−0.761	4.03	−0.537	4.27	1.06
1 M	−0.815	4.51	−0.572	5.19	1.15

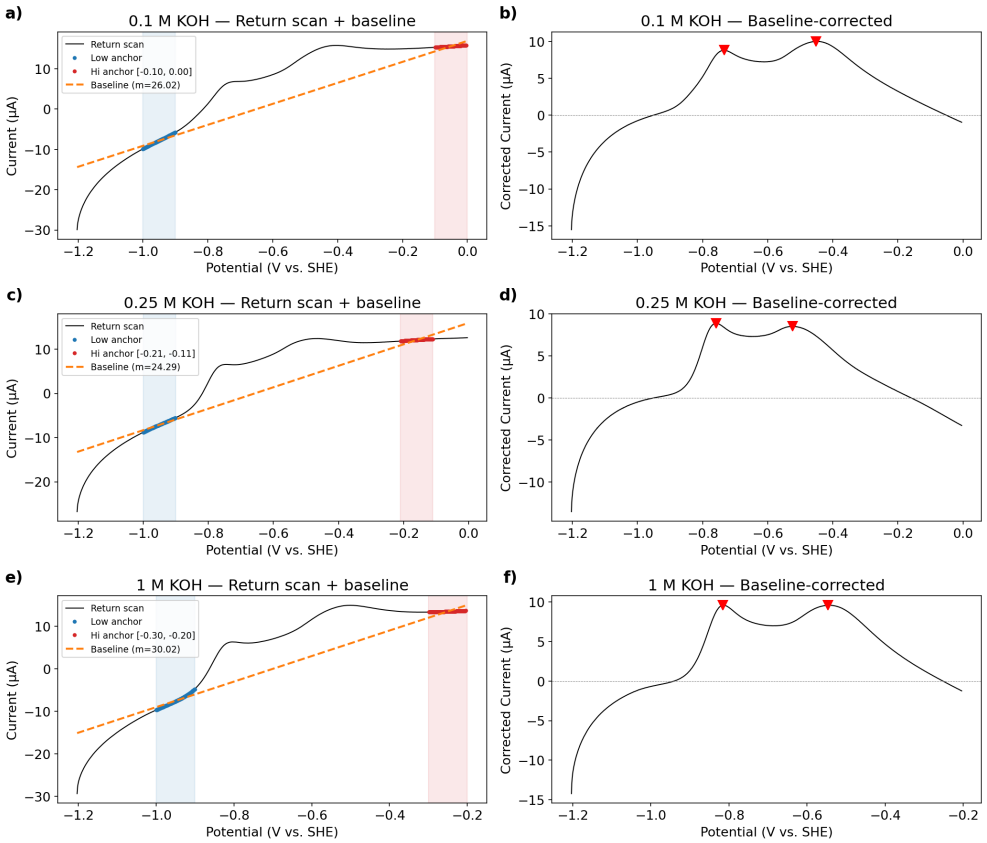


Fig. S25 Baseline correction of the anodic return scan for 1 mM **3** at 100 mV s^{-1} in 0.1 M KOH (a, b), 0.25 M KOH (c, d), and 1 M KOH (e, f).

Table S11 Oxidation peak data at 100 mV s^{-1} across KOH concentrations.

KOH	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
0.1 M	−0.734	8.79	−0.452	10.01	1.14
0.25 M	−0.759	8.88	−0.524	8.53	0.96
1 M	−0.815	9.60	−0.546	9.60	1.00

S7.10 Scan rate dependence (0.5 M NaOH)

Cyclic voltammograms of 1 mM **3** were recorded at 10-500 mV s^{-1} in 0.5 M NaOH (Figure S26). The baseline-correction procedure described in the potassium hydroxide concentration dependence experiment was used. Figure S27 shows the fitted baselines and corrected oxidation waves, and Table S12 summarises the peak data. At 10 mV s^{-1} only a single oxidation peak was resolved on the unsmoothed data; a shoulder near −0.75 V became apparent after smoothing with an 11-point uniform filter.

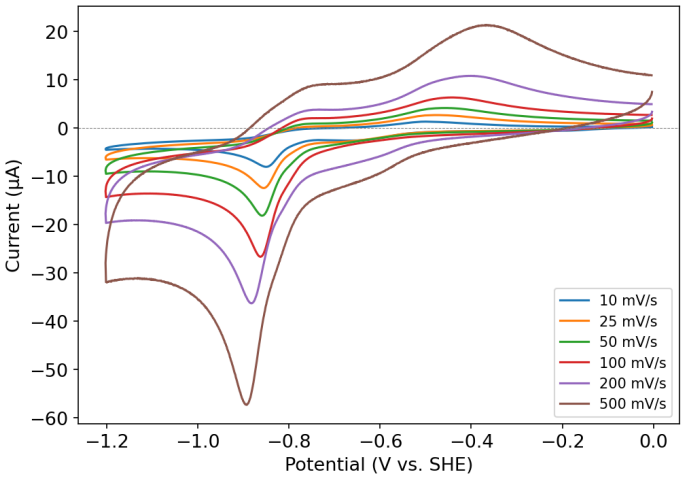


Fig. S26 Cyclic voltammograms of 1 mM **3** in 0.5 M NaOH at scan rates from 10 to 500 mV s^{-1}.

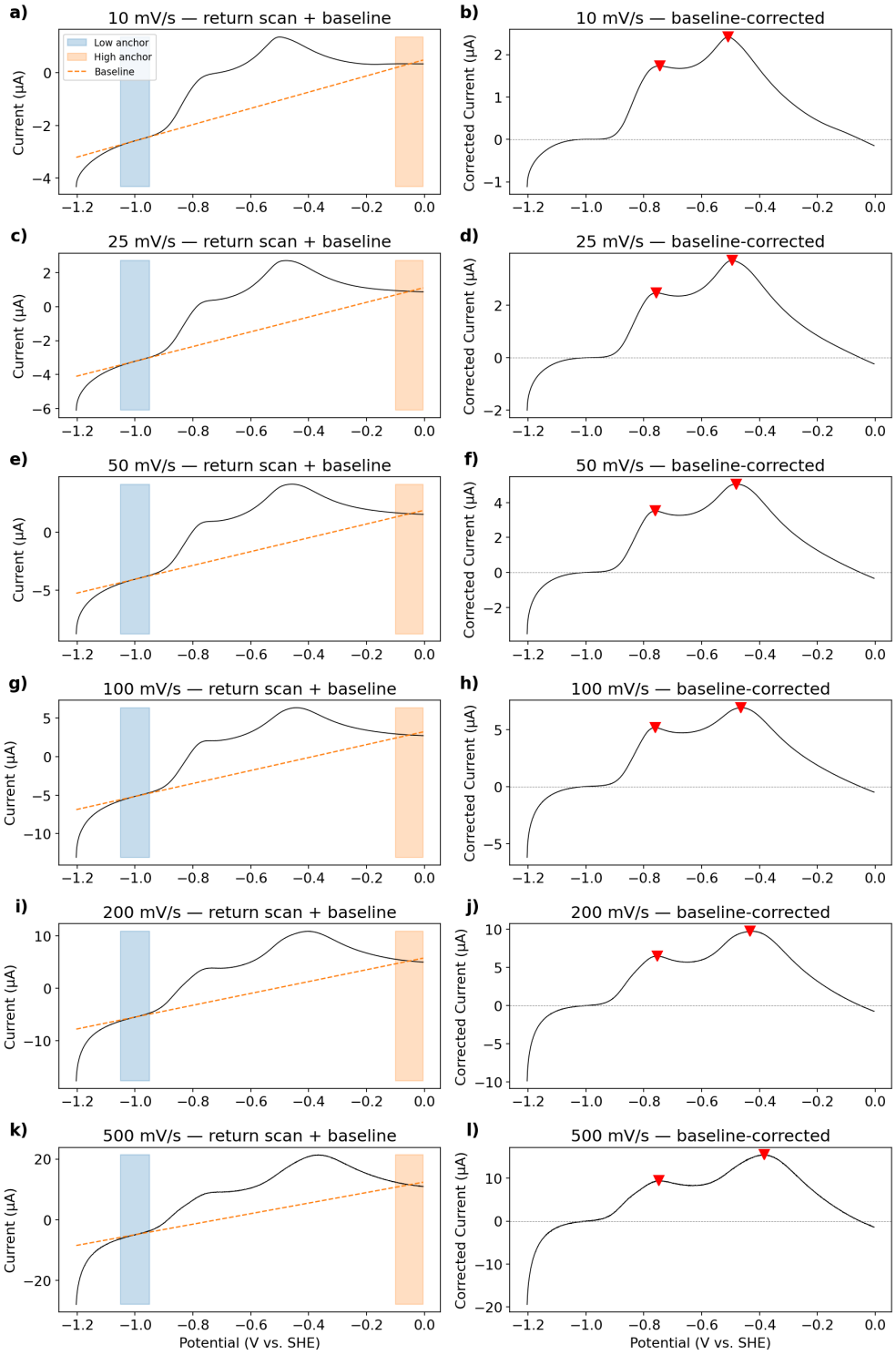


Fig. S27 Baseline correction of the anodic return scan for 1 mM **3** Left panels show the return scan (black) with anchor regions (shaded) and fitted linear baseline (dashed orange). Right panels show the baseline-corrected current with peak maxima marked by red triangles.

Table S12 Oxidation peak positions and currents from the baseline-corrected anodic scan of 1 mM **3** in 0.5 M NaOH at varying scan rates.

Scan rate ($mV\,s^{-1}$)	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
10	−0.745	1.73	−0.508	2.42	1.39
25	−0.756	2.48	−0.494	3.71	1.50
50	−0.761	3.52	−0.481	5.06	1.44
100	−0.761	5.17	−0.465	6.92	1.34
200	−0.753	6.48	−0.433	9.73	1.50
500	−0.748	9.41	−0.383	15.47	1.64

S7.11 Scan rate dependence (0.5 M LiOH)

Cyclic voltammograms of 1 mM **3** were recorded at 10-500 $mV\,s^{-1}$ in 0.5 M LiOH (Figure S28). The baseline-correction procedure described in the potassium hydroxide concentration dependence experiment was used. Figure S29 shows the fitted baselines and corrected oxidation waves, and Table S13 summarises the peak data. At 10 $mV\,s^{-1}$ only a single oxidation peak was resolved on the unsmoothed data; a shoulder near −0.75 V became apparent after smoothing with an 11-point uniform filter.

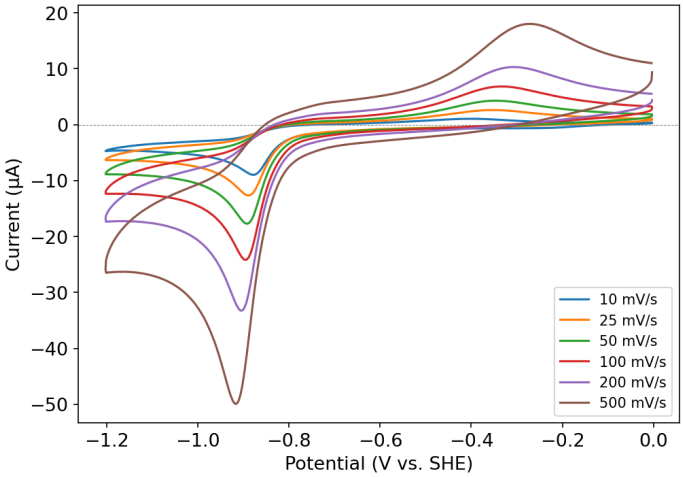


Fig. S28 Cyclic voltammograms of 1 mM **3** in 0.5 M LiOH at scan rates from 10 to 500 $mV\,s^{-1}$.

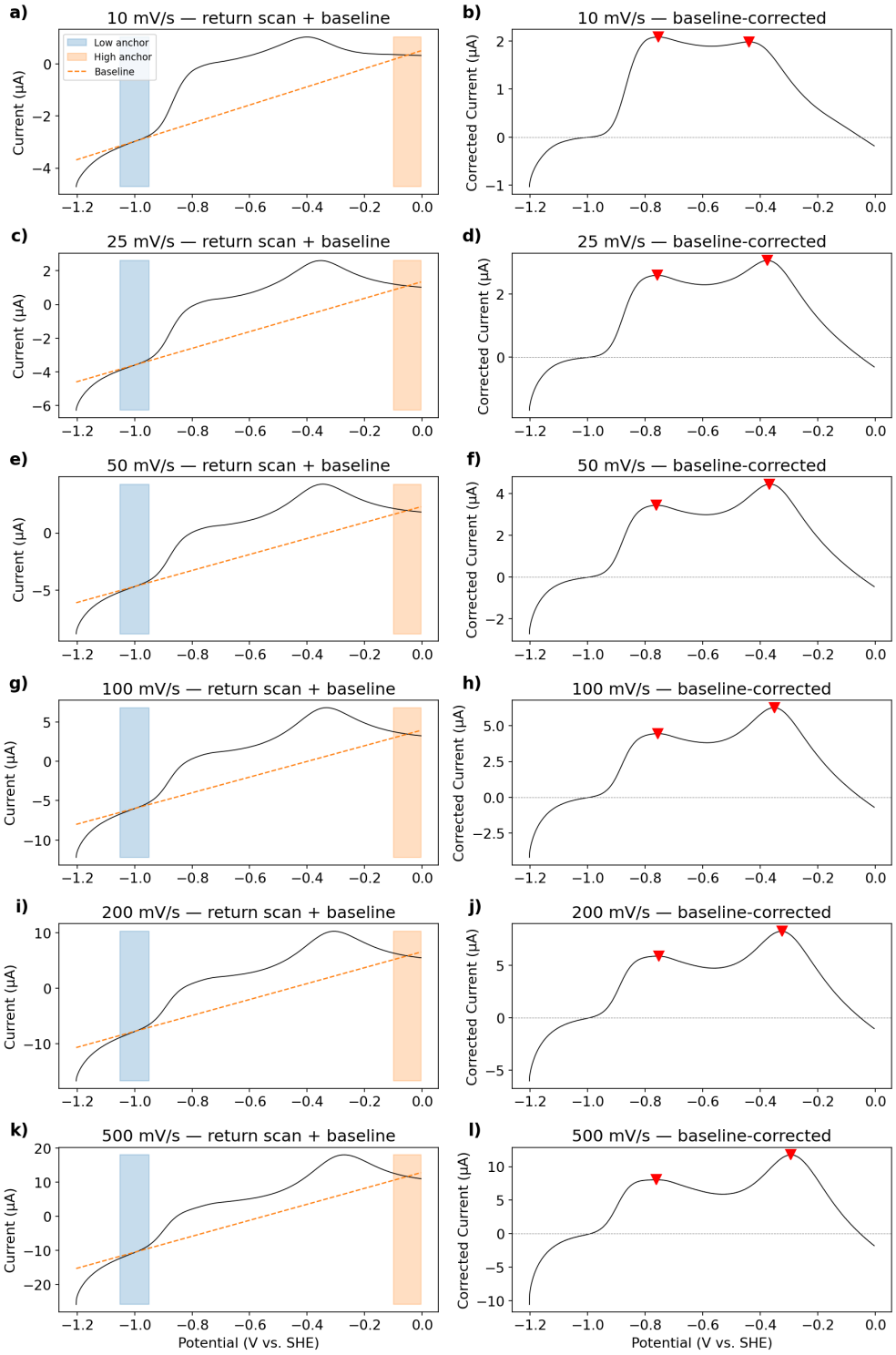


Fig. S29 Baseline correction of the anodic return scan for 1 mM **3** Left panels show the return scan (black) with anchor regions (shaded) and fitted linear baseline (dashed orange). Right panels show the baseline-corrected current with peak maxima marked by red triangles.

Table S13 Oxidation peak positions and currents from the baseline-corrected anodic scan of 1 mM **3** in 0.5 M LiOH at varying scan rates.

Scan rate ($mV\,s^{-1}$)	Peak 1 (V)	Peak 1 (μA)	Peak 2 (V)	Peak 2 (μA)	Ratio (P2/P1)
10	−0.754	2.08	−0.440	1.98	0.95
25	−0.759	2.59	−0.375	3.05	1.18
50	−0.762	3.43	−0.368	4.45	1.30
100	−0.757	4.45	−0.351	6.26	1.41
200	−0.753	5.88	−0.324	8.25	1.40
500	−0.762	8.05	−0.295	11.77	1.46

S7.12 Concentration dependence of 20

To assess the effect of analyte concentration on the electrochemical response of **20**, cyclic voltammograms were recorded at 0.01 mM and 5 mM in 0.5 M KOH at scan rates of 25, 50, and 100 $mV\,s^{-1}$.

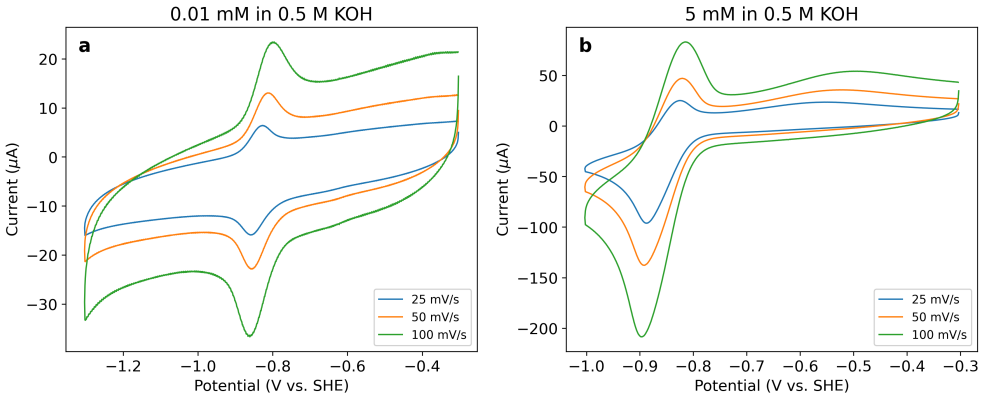


Fig. S30 Cyclic voltammograms of **20** in 0.5 M KOH at (a) 0.01 mM and (b) 5 mM, recorded at 25, 50, and 100 $mV\,s^{-1}$. Potentials are referenced to SHE.

S7.13 Potassium hydroxide incubation

A 1 mM solution of **3** in 0.5 M KOH was incubated at room temperature under inert atmosphere for approximately 20 days. After incubation, the sample was examined by cyclic voltammetry at multiple scan rates (10–250 $mV\,s^{-1}$).

UV-Vis absorption spectra recorded before and after incubation show approximately 2.3% reduction in the magnitude of absorbance (Figure S32).

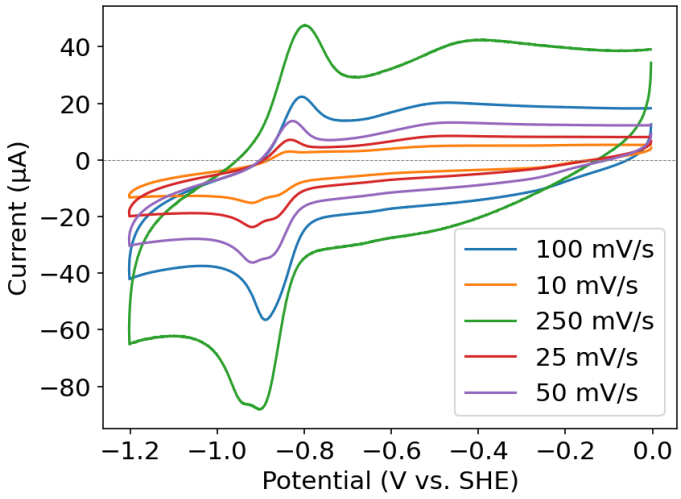


Fig. S31 Multi-scan-rate cyclic voltammograms of degraded **3** (1 mM in 0.5 M KOH) recorded at 10, 25, 50, 100, and 250 mV s^{-1}.

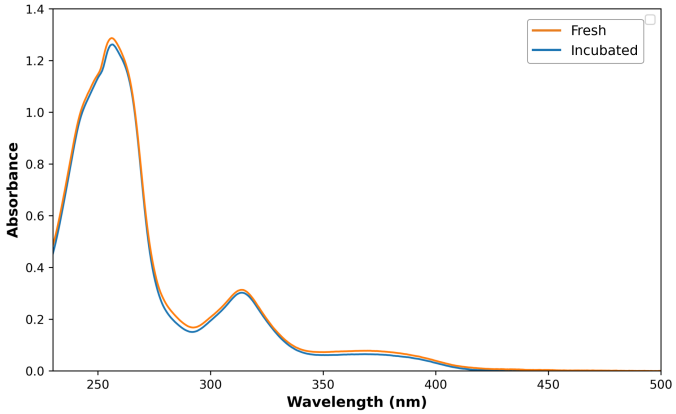


Fig. S32 UV-Vis absorption spectra of **3** (1 mM in 0.5 M KOH) before (fresh) and after incubation at room temperature.

S8 Spectroelectrochemistry

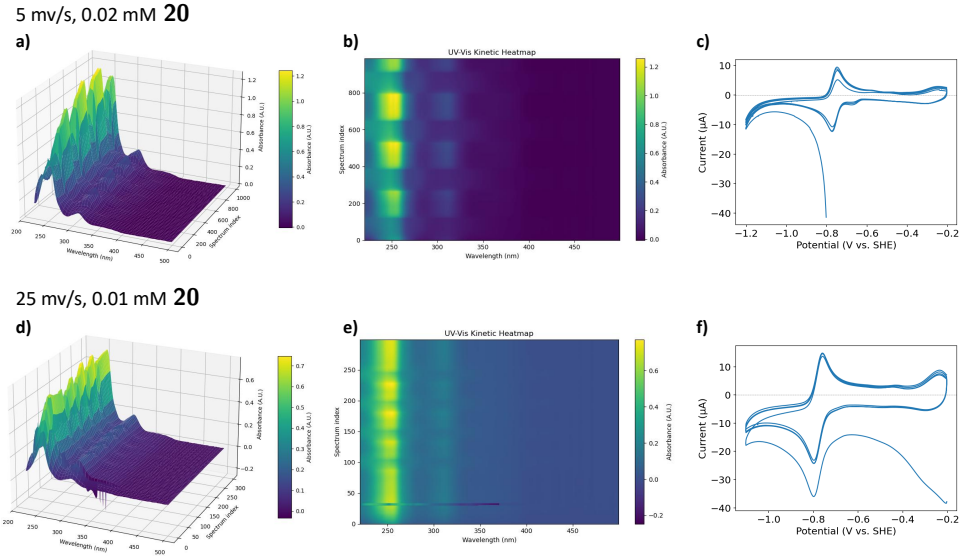


Fig. S33 Spectroelectrochemistry of **20** in 0.5 M KOH. (a) Three-dimensional UV-Vis absorbance surface during cyclic voltammetry at 5 mV s^{-1} (0.02 mM **20**). (b) UV-Vis kinetic heatmap at 5 mV s^{-1}. (c) Corresponding cyclic voltammogram at 5 mV s^{-1}. (d)–(f) Analogous data at 25 mV s^{-1} (0.01 mM **20**).

S9 Solubility studies

S9.1 Nuclear magnetic resonance spectroscopy

Nuclear Magnetic Resonance (NMR) spectra were acquired on Agilent DD2 400 MHz and Agilent DD2 600 MHz NMR spectrometers at room temperature. Samples were prepared primarily in D_2O unless otherwise noted. Chemical shifts are reported in parts per million (ppm). ^{1}H NMR spectra collected in D_2O were referenced to the residual HOD solvent signal.

S9.1.1 Measurement protocol

Quantitative ^{1}H NMR was used to independently verify the solubility of **3** and **20** in 0.5 M KOH. A 50 µL aliquot of the saturated **3** supernatant was diluted with 450 µL of 50 mM sodium methanesulfonate in D_2O and analyzed by ^{1}H NMR (16 scans, 60 s relaxation delay). The sodium methanesulfonate singlet was normalized to 3.00 H, and the **3** aromatic resonances (assumed to represent 7H) were integrated relative to this internal standard. After correcting for the 10× dilution factor, the **3** concentration was determined to be approximately 57.9 mM (Figure S34).

Similarly, a 100 µL aliquot of the saturated **20** supernatant was diluted with 400 µL of 50 mM sodium methanesulfonate in D_2O and analyzed under identical acquisition conditions. After correcting for the 5× dilution factor, the **20** concentration was determined to be approximately 56.0 mM (Figure S35).

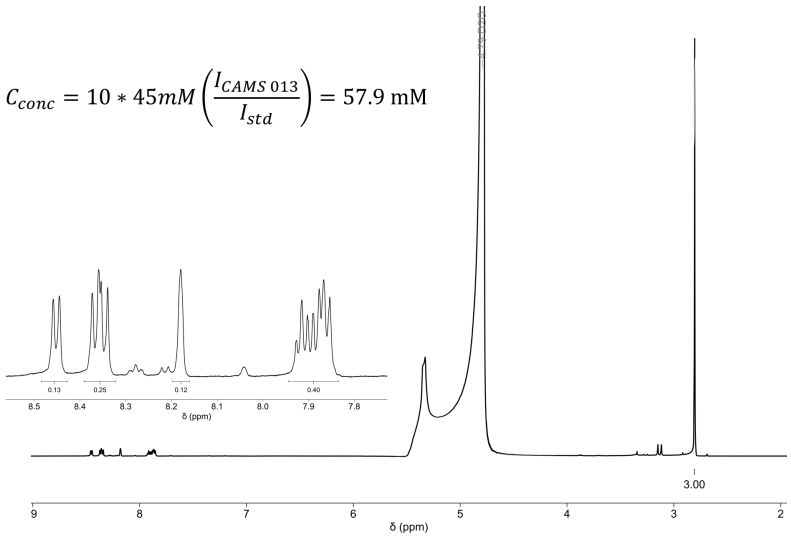


Fig. S34 Quantitative ^{1}H NMR spectrum of the saturated **3** supernatant in D_2O with sodium methanesulfonate internal standard. Inset: expansion of the aromatic region.

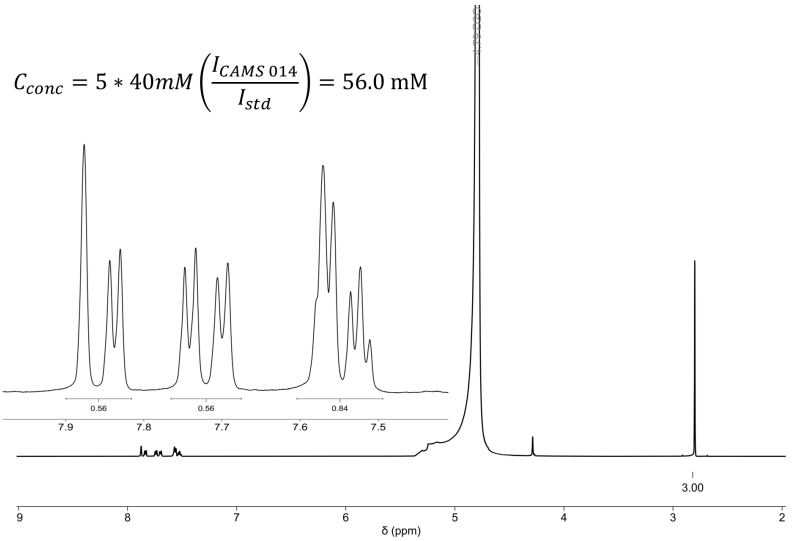


Fig. S35 Quantitative ^{1}H NMR spectrum of the saturated **20** supernatant in D_2O with sodium methanesulfonate internal standard. Inset: expansion of the aromatic region.

S9.2 Ultraviolet-visible spectroscopy

UV-Vis spectroscopy was conducted on an Ocean-HDX-UV-VIS spectrometer and a DH-Mini-UV-Vis-NIR light source (OceanOptics). Spectral data were collected over a wavelength of 215-800 nm with an integration time of 25 ms averaged over 50 scans.

Measurements were performed in a standard 1 cm pathlength polystyrene cuvette. The background spectra were collected using the corresponding blank solvent under identical conditions and subtracted.

S9.2.1 Calibration curve procedure

UV–Vis calibration curves were prepared for **3** and **20** in 0.5 M KOH. For each compound, an appropriate mass was carefully weighed and dissolved in 0.5 M KOH to prepare a 1 mM stock solution in a 10 mL centrifuge tube. The exact stock concentration was calculated from the measured mass, molecular weight, and final solution volume. The stock solution was then serially diluted with 0.5 M KOH to prepare calibration standards within the linear absorbance range.

UV–Vis spectra were collected using 0.5 M KOH as the blank. For each compound, the absorbance at the wavelength of maximum absorbance, λ_{max}, was used to construct the calibration curve. Absorbance at λ_{max} was plotted as a function of concentration, and the data were fit using a linear regression of the form $A = mC + b$, where A is absorbance, C is concentration, m is the slope, and b is the y-intercept. Only calibration standards within the linear absorbance range were included in the fit. Linear fits gave R^2 values of 0.99 or greater for both **3** and **20**.

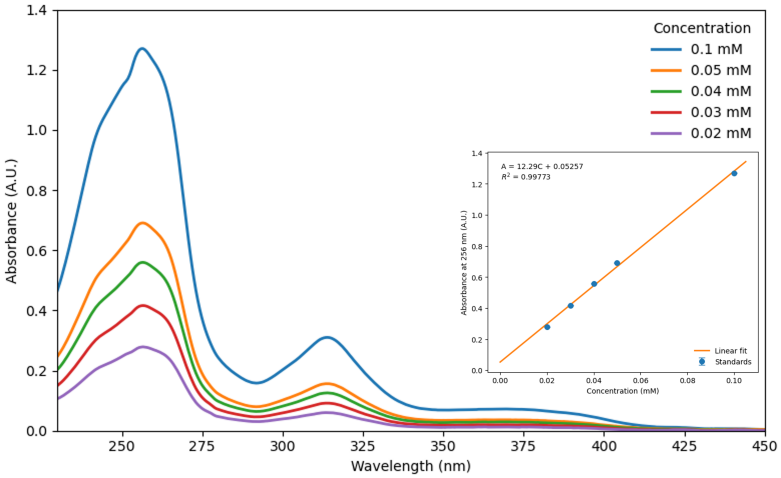


Fig. S36 UV–Vis spectra of **3** calibration standards in 0.5 M KOH. Inset: calibration curve at $\lambda_{max} = 256$ nm with linear fit.

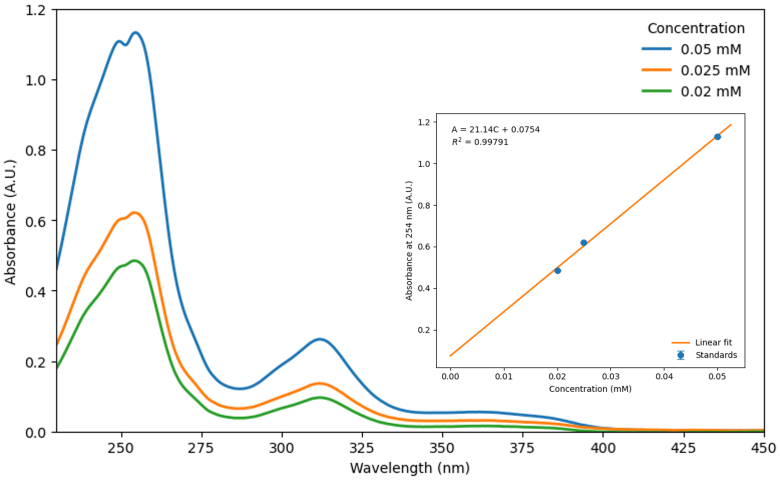


Fig. S37 UV–Vis spectra of **20** calibration standards in 0.5 M KOH. Inset: calibration curve at $\lambda_{max} = 254$ nm with linear fit.

S9.2.2 Measurement protocol

Approximately 25 mg of **3** and 30 mg of **20** were carefully weighed into separate 7 mL scintillation vials. For **3**, 0.5 mL of 0.5 M KOH was added, while 1.5 mL of 0.5 M KOH was added to **20**. The amount of solvent was selected such that undissolved solid was visible, indicating that the mixtures were near or above their apparent solubility limit. **3** was noted to produce a darker yellow solution than **20** and appeared to exhibit much higher solubility under these conditions. A stir bar was added to each vial, and the vials were capped, sealed with electrical tape, and stirred at room temperature for 48 h. After stirring, each mixture was transferred to a 1.5 mL microcentrifuge tube and centrifuged to separate the undissolved solids from the supernatant. The supernatant was carefully transferred to another clean microcentrifuge tube and centrifuged again. A final centrifugation step was performed for 10 min to further remove any remaining solids. After this step, no visible solids were observed.

The supernatant was analyzed by UV-Vis spectroscopy after serial dilution to bring the absorbance within a linear calibration range. Concentrations were determined by comparison to a corresponding calibration curve. For quantitative NMR analysis, aliquots of the supernatant were diluted with 50 mM sodium methanesulfonate in D_2O as an internal standard. A 50 µL aliquot of **3** was combined with 450 µL of the 50 mM sodium methanesulfonate solution, while 100 µL of **20** was combined with 400 µL of the methanesulfonate solution.

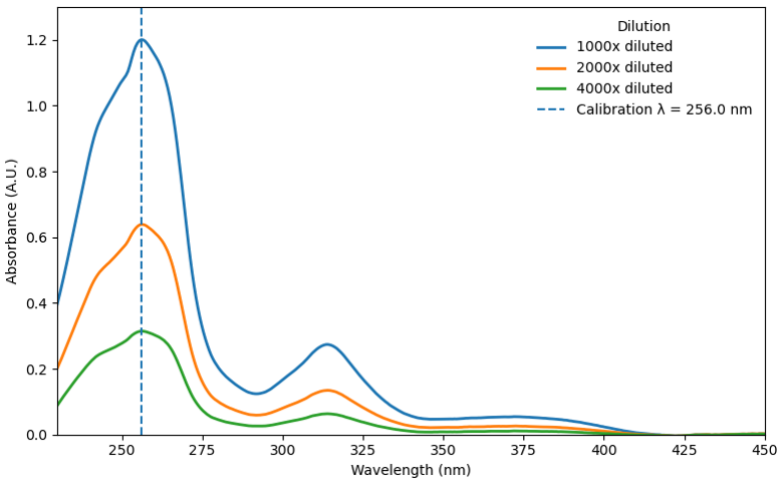


Fig. S38 UV–Vis spectra of serial dilutions of the saturated **3** supernatant in 0.5 M KOH. The dashed line indicates the calibration wavelength ($\lambda = 256$ nm).

Table S14 UV–Vis solubility determination of **3** in 0.5 M KOH.

Dilution	Diluted conc. (mM)	Original conc. (mM)
1000×	0.093	93.45
2000×	0.048	95.44
4000×	0.021	85.25

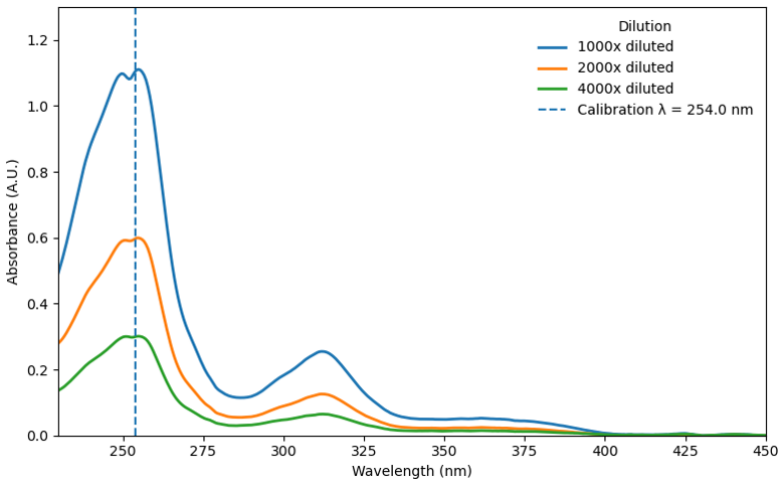


Fig. S39 UV–Vis spectra of serial dilutions of the saturated **20** supernatant in 0.5 M KOH. The dashed line indicates the calibration wavelength ($\lambda = 254$ nm).

Table S15 UV–Vis solubility determination of **20** in 0.5 M KOH.

Dilution	Diluted conc. (mM)	Original conc. (mM)
1000×	0.049	48.75
2000×	0.025	49.41
4000×	0.011	42.27

S10 Synthetic procedures

S10.1 General information

Reactions were performed under ambient atmosphere, unless otherwise noted. Yields refer to chromatographically and spectroscopically (^{1}H NMR) homogeneous materials, unless otherwise stated. Reagents were purchased from commercial vendors and used as received unless otherwise stated. All solvents were purchased as DriSolv grade and used as received without further drying. Column chromatography was performed on silica gel 60 (SiliCycle, 60–120 mesh). Thin-layer chromatography (TLC) utilized pre-coated plates (Sorbtech, silica gel 60 PF_{254}, 0.25 mm) visualized with UV 254 nm, ninhydrin, or basic potassium permanganate stain. Analytical HPLC was performed on Waters HPLC systems using acetonitrile, methanol, and 0.1% aq. formic acid as eluents. Mass spectra were recorded using a Micromass quattro ultima equipped with an ESI source. NMR spectra were recorded on a Bruker Avance (300 MHz) spectrometer.

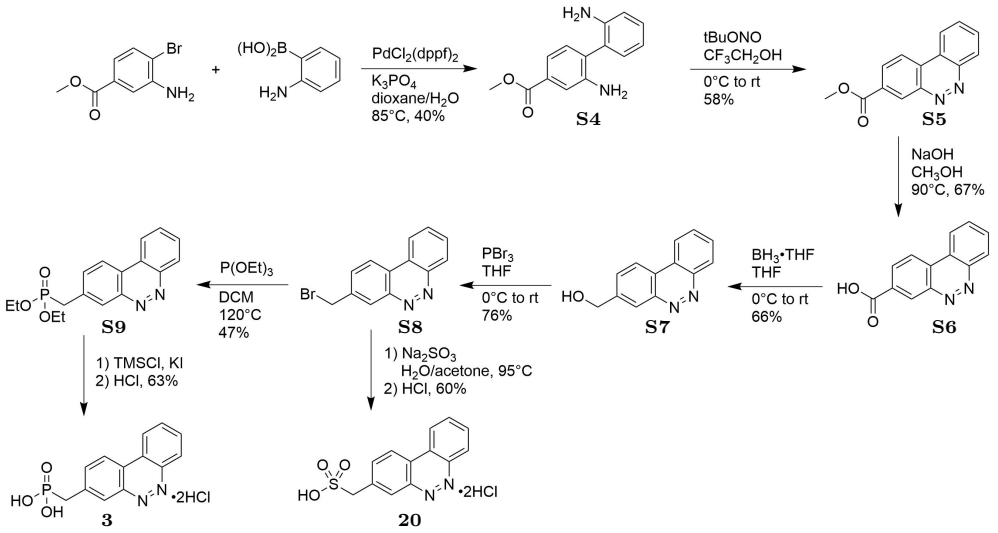


Fig. S40 Synthetic route to methanephosphonate (**3**) and methanesulfonate (**20**) benzo[c]cinnoline derivatives.

S10.2 Methyl 2,2'-diamino-4-biphenylcarboxylate (S4)

To a stirred solution of methyl 3-amino-4-bromobenzoate (5.0 g, 21.73 mmol) and 2-(dihydroxyboryl)phenylamine hydrochloride (4.145 g, 23.91 mmol) in 1,4-dioxane (75 mL)/water (25 mL), KOAc (13.84 g, 56.2 mmol) and $Pd(dppf)_2Cl_2$ (0.794 g, 1.09 mmol) were added. The mixture was heated at 100 °C for 16 h. After cooling to room temperature, the mixture was diluted with EtOAc and filtered through a pad of Celite. The organic layer of the filtrate was separated, dried over Na_2SO_4 and concentrated under reduced pressure. The dark red oil obtained was purified by flash column chromatography to afford the product as a reddish orange solid (2.10 g, 40%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.55 (m, 2H), 7.26 (m, 2H), 7.17 (d, J = 7.5 Hz, 1H), 6.90 (t, J = 7.5 Hz, 1H), 6.85 (d, J = 8.1 Hz, 1H), 3.97 (s, 3H), 3.94 (br s, 2H), 3.77 (br s, 2H); ES-MS: 243.7 $[M+H]^+$.

S10.3 Methyl 9,10-diaza-2-phenanthrenecarboxylate (S5)

Methyl 2,2'-diamino-4-biphenylcarboxylate (2.10 g, 8.64 mmol, **S4**) was dissolved in 2,2,2-trifluoroethanol (30 mL) and cooled to 0 °C. *tert*-Butyl nitrile (3.1 mL, 25.93 mmol) was added dropwise at 0 °C, and the mixture was stirred overnight in the same cooling bath. After completion, solvent was removed under reduced pressure, and the crude mixture was purified by flash column chromatography to afford the product as a brown solid (1.2 g, 58.3%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 9.44 (s, 1H), 8.81 (m, 1H), 8.63 (m, 2H), 8.54 (m, 1H), 8.12 (m, 1H), 7.92 (m, 2H), 4.06 (s, 3H); ES-MS: 239.7 $[M+H]^+$.

S10.4 9,10-Diaza-2-phenanthrenecarboxylic acid (S6)

A solution of methyl 9,10-diaza-2-phenanthrenecarboxylate (1.00 g, 4.197 mmol, **S5**) in MeOH (60 mL) was treated with 1 M NaOH (6.30 mL, 6.29 mmol) and heated at 90 °C for 2 h. A solution of oxalic acid (2.0 equiv) in H_2O (10 mL) was added dropwise, and the mixture was stirred for 5 min. After completion, methanol was removed under reduced pressure, and CH_2Cl_2 was added. The solid obtained was filtered, washed with CH_2Cl_2 and H_2O, and dried on a high vacuum pump to afford the product as a brown solid (0.630 g, 67%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 9.31 (s, 1H), 8.92 (m, 2H), 8.75 (m, 1H), 8.59 (m, 1H), 8.13 (m, 3H). ESI-MS: 225.7 $[M+H]^+$.

S10.5 (9,10-Diaza-2-phenanthryl)methanol (S7)

9,10-Diaza-2-phenanthrenecarboxylic acid (0.630 g, 2.81 mmol, **S6**) was suspended in anhydrous THF (20 mL) and cooled to 0 °C under an inert atmosphere. A solution of BH_3·THF complex (14 mL, 14.05 mmol) was added dropwise at 0 °C, and the mixture was stirred for 10 min at 0 °C, then allowed to warm to room temperature and stirred overnight. The reaction mixture was cooled back to 0 °C and carefully quenched by the slow addition of MeOH. After stirring for 10 min, the solvent was removed under reduced pressure. MeOH was added, and the mixture was concentrated again under reduced pressure to ensure complete removal of boron residues. The crude obtained was purified by flash column chromatography to afford pure product as a brown solid (0.390 g, 66%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.67 (m, 2H), 8.49 (m, 2H), 7.91 (m, 3H), 5.02 (s, 2H).

S10.6 2-(Bromomethyl)-9,10-diazaphenanthrene (S8)

(9,10-Diaza-2-phenanthryl)methanol (0.354 g, 1.684 mmol, **S7**) was dissolved in anhydrous THF (8 mL) under an inert atmosphere and cooled to 0 °C. A solution of PBr_3 (0.24 mL, 2.526 mmol) in CH_2Cl_2 (2 mL) was added dropwise maintaining the temperature at 0 to 5 °C. The mixture was stirred at 0 to 5 °C for 30 min, allowed to warm up to room temperature and stirred for an additional 2 h. After completion, the mixture was diluted with CH_2Cl_2 and carefully quenched with saturated aqueous $NaHCO_3$. The organic layer was separated, dried over anhydrous Na_2SO_4, and concentrated under reduced pressure to afford the product as a yellow solid (0.350 g, 76%) which was taken forward without purification. ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.75 (m, 2H), 8.56 (m, 2H), 7.96 (m, 3H), 4.76 (s, 2H).

S10.7 Diethyl [(9,10-diaza-2-phenanthryl)methyl]phosphonate (S9)

2-(Bromomethyl)-9,10-diazaphenanthrene (0.160 g, 0.586 mmol, **S8**) and triethyl phosphite (1.1 mL, 5.84 mmol) were mixed in a pressure vessel and heated at 120 °C for 16 hours. After completion, the mixture was cooled to room temperature, concentrated under reduced pressure and purified by flash column chromatography to afford the product as a yellow solid (0.087 g, 45%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.74 (m, 1H),

8.62 (br s, 1H), 8.56 (m, 2H), 7.92 (m, 3H), 4.10 (t, J = 7.2 Hz, 4H), 3.52 (s, 1H), 3.44 (s, 1H), 1.29 (t, J = 7.2 Hz, 6H). ^{31}P NMR (121 MHz, $CDCl_3$) δ 25.00.

S10.8 [(9,10-Diaza-2-phenanthryl)methyl]phosphonic acid hydrochloride (3)

To a solution of diethyl [(9,10-diaza-2-phenanthryl)methyl]phosphonate (0.090 g, 0.272 mmol, **S9**) in CH_3CN (4 mL), KI (0.262 g, 1.36 mmol) was added, followed by the dropwise addition of TMSCl (0.175 mL, 1.362 mmol). The reaction mixture was stirred at room temperature for 20 minutes and then heated at 60 °C for 5 hours. After cooling to room temperature, the solvent was removed under reduced pressure. The crude obtained was washed with CH_2Cl_2 (3×3 mL) and then with MeOH (2×3 mL). The residue was treated with concentrated HCl (2 mL) and stirred for 5 minutes. HCl solution was then evaporated under reduced pressure. The solid was washed with water and then with MeOH (2×2 mL), to afford clean hydrochloride salt of the product as a yellow solid (60 mg, 63%). ^{1}H NMR (300 MHz, DMSO-d_6) δ 8.86 (br s, 2H), 8.67 (br s, 1H), 8.56 (br s, 1H), 8.01 (br s, 3H), 3.15 (s, 2H). ^{31}P NMR (121 MHz, DMSO-d_6) δ 19.9. ES-MS: 275.8 $[M+H]^+$. HPLC purity: 90.9%.

S10.9 (9,10-Diaza-2-phenanthryl)methanesulfonic acid hydrochloride (20)

2-(Bromomethyl)-9,10-diazaphenanthrene (0.150 g, 0.549 mmol, **S8**) was suspended in a mixture of H_2O (5 mL) and acetone (2 mL). Na_2SO_3 (0.070 g, 0.549 mmol) was added as a solid, and the mixture was stirred at room temperature for 20 min and then heated at 100 °C for 16 h. After cooling to room temperature, the solvent was removed under reduced pressure. The residue was washed with CH_2Cl_2 (3×10 mL), and the insoluble material was treated with concentrated HCl, resulting in the formation of a brown precipitate. The solid was collected by filtration and washed with CH_2Cl_2 to afford the hydrochloride salt of the product as a brown solid (0.118 g, 60%). ^{1}H NMR (300 MHz, CD_3OD) δ 9.06 (t, J = 9 Hz, 2H), 8.76 (br m, 2H), 8.45 (d, J = 8.1 Hz, 1H), 8.36 (t, J = 6.9 Hz, 1H), 8.28 (t, J = 6.9 Hz, 1H), 4.51 (s, 2H). ESI-MS: 273.8 $[M-H]^-$. HPLC purity: 99.0%.

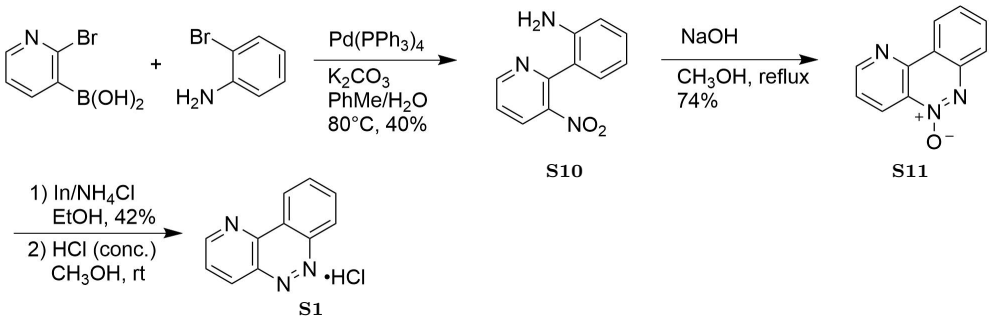


Fig. S41 Synthetic route to 4,9,10-triazaphenanthrene hydrochloride (**S1**).

S10.10 2-(3-Nitropyridin-2-yl)benzen-1-amine (S10)

To a stirred solution of 2-aminophenylboronic acid hydrochloride (5.00 g, 28.84 mmol) and 2-bromo-3-pyridylamine (5.85 g, 28.84 mmol) in a mixture of 1,4-dioxane (75 mL) and H_2O (50 mL), K_3PO_4 (18.36 g, 86.51 mmol) and $Pd(dppf)_2Cl_2$ (1.05 g, 1.44 mmol) were added. The reaction mixture was heated at 85 °C for 48 h. After cooling to room temperature, 1 M HCl (200 mL) was added, and the mixture was stirred for 5 min. The suspension was filtered through a pad of Celite, and the filtrate was concentrated under reduced pressure to remove 1,4-dioxane. The remaining aqueous solution was washed with EtOAc (3×150 mL). The aqueous layer was then carefully neutralized by dropwise addition of saturated aqueous Na_2CO_3 and extracted with EtOAc (3×150 mL). The combined organic extracts were dried over anhydrous Na_2SO_4, filtered, and concentrated under reduced pressure to afford a dark orange-red oil. Purification by silica gel chromatography (20–80% EtOAc/hexanes) afforded the product as a reddish orange oil (40%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.85 (dd, J = 4.8, 1.2 Hz, 1H), 8.21 (dd, J = 8.4 Hz, 1H), 7.45 (m, 1H), 7.26 (m, 2H), 7.07 (m, 1H), 6.82 (m, 2H), 7.45 (s, 2H). ESI-MS: 216.6 $[M+H]^+$.

S10.11 4,9,10-Triaza-9-phenanthrenium-9-olate (S11)

To a stirred solution of 2-(3-nitropyridin-2-yl)benzen-1-amine (**S10**, 2.50 g, 11.62 mmol) in MeOH (50 mL), NaOH (2.33 g, 58.14 mmol) was added. The reaction mixture was heated at 80 °C for 2 h. After completion, the mixture was cooled to room temperature and concentrated *in vacuo*. The residue was dissolved in EtOAc and washed with saturated aqueous $NaHCO_3$. The layers were separated, and the aqueous layer was extracted with EtOAc (3×100 mL). The combined organic extracts were dried over anhydrous Na_2SO_4, filtered, and concentrated under reduced pressure to afford the product (1.70 g, 74%) as a brown solid. ^{1}H NMR (300 MHz, $CDCl_3$) δ 9.21 (d, J = 3.9 Hz, 1H), 9.08 (d, J = 8.7 Hz, 1H), 8.91 (d, J = 7.8 Hz, 1H), 8.04 (d, J = 8.4 Hz, 1H), 7.89 (m, 1H), 7.79 (m, 2H).

S10.12 4,9,10-Triazaphenanthrene (S12)

To a stirred solution of 4,9,10-triaza-9-phenanthrenium-9-olate (**S11**, 1.70 g, 8.62 mmol) in EtOH (70 mL) and saturated aqueous NH_4Cl (20 mL), indium powder (2.97 g, 25.9 mmol) was added. The reaction mixture was heated at 80 °C for 4 h and monitored by TLC. After completion, the mixture was concentrated under reduced pressure, diluted with EtOAc, and stirred for 5 min. The suspension was filtered through a pad of Celite, and the pad was washed with EtOAc (3 × 100 mL). The combined filtrates were concentrated under reduced pressure and purified by flash column chromatography (5–20% EtOAc/hexanes) to afford the product as a brown solid (42%). ^{1}H NMR (300 MHz, CD_3OD) δ 9.45 (m, 1H), 9.25 (m, 2H), 8.85 (m, 1H), 8.30 (m, 2H), 8.23 (m, 1H). ESI-MS: 182.7 $[M+H]^+$.

S10.13 4,9,10-Triazaphenanthrene hydrochloride (S1)

To a solution of 4,9,10-triazaphenanthrene (**S12**, 900 mg, 4.97 mmol) in MeOH (20 mL), concentrated HCl (3 mL) was added dropwise at 0–5 °C. The mixture was stirred for 2 min at this temperature, then allowed to warm to room temperature and stirred for an additional 5 min. All volatiles were removed under reduced pressure. The resulting solid was washed with CH_2Cl_2 (3 × 20 mL) and dried under high vacuum overnight to afford the product (1.15 g, 80%) as a light yellow solid. ^{1}H NMR (300 MHz, CD_3OD) δ 9.42 (d, 1H), 9.25 (m, 1H), 9.20 (m, 1H), 8.22 (m, 2H), 8.15 (m, 1H). ESI-MS: 182.7 $[M+H]^+$. HPLC purity: 99%.

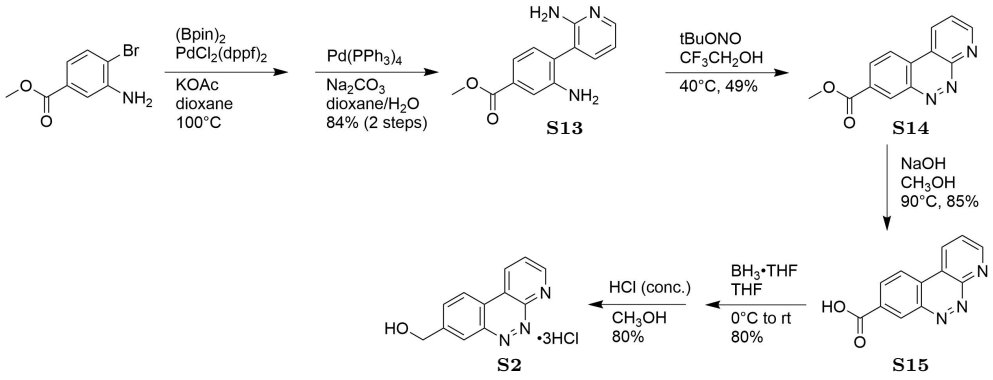


Fig. S42 Synthetic route to (1,9,10-triaza-7-phenanthryl)methanol hydrochloride (**S2**).

S10.14 Methyl 3-amino-4-(2-aminopyridin-3-yl)benzoate (S13)

To a stirred solution of methyl 3-amino-4-bromobenzoate (6.0 g, 26.08 mmol) and 4,4,5,5-tetramethyl-2-(tetramethyl-1,3,2-dioxaborolan-2-yl)-1,3,2-dioxaborolane (6.622 g, 26.08 mmol) in 1,4-dioxane (100 mL), KOAc (5.138 g, 52.1 mmol) and $Pd(dppf)_2Cl_2$ (0.953 g, 1.304 mmol) were added. The reaction mixture was heated at

100 °C for 16 h. After cooling to room temperature, 3-bromo-2-aminopyridine (3.74 g, 21.65 mmol) was added, followed by $Pd(dppf)_2Cl_2$ (791 mg, 1.85 mmol), K_3PO_4 (4.6 g, 43.3 mmol), and H_2O (15 mL). The reaction mixture was heated at 80 °C for 16 h. After cooling to room temperature, 1 M HCl (200 mL) was added, and the mixture was stirred for 5 min. The mixture was passed through a pad of Celite, and the filtrate was concentrated under reduced pressure to remove 1,4-dioxane. The aqueous solution was washed with EtOAc (3 × 150 mL). The aqueous layer was carefully neutralized with saturated aqueous Na_2CO_3 (added dropwise) and extracted with EtOAc (3 × 150 mL). The combined organic extracts were dried over Na_2SO_4, filtered, and concentrated under reduced pressure. The crude mixture was purified by flash column chromatography (EtOAc/hexanes, 20–80%) to afford the product as a reddish orange solid (84%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 8.15 (d, J = 4.8 Hz, 1H), 7.51 (m, 2H), 7.41 (d, J = 7.5 Hz, 1H), 7.26 (s, 1H), 7.18 (d, J = 7.5 Hz, 1H), 6.8 (m, 1H), 4.50 (s, 2H), 3.92 (s, 3H), 3.83 (s, 2H). ESI-MS: 244.7 $[M+H]^+$.

S10.15 Methyl 1,9,10-triaza-7-phenanthrenecarboxylate (S14)

Methyl 3-amino-4-(2-aminopyridin-3-yl)benzoate (**S13**, 3.50 g, 14.40 mmol) was dissolved in 2,2,2-trifluoroethanol (50 mL) and cooled to 0–5 °C. *tert*-Butyl nitrile (5.14 mL, 43.21 mmol) was added dropwise at 0–5 °C, and the reaction mixture was stirred in the same cooling bath for 16 h. The reaction was quenched with 1 M HCl (150 mL), and the solvent was removed under reduced pressure. The residue was dissolved in an immiscible mixture of EtOAc and 1 M HCl, and the aqueous layers were separated. The organic layer was washed again with 1 M HCl. The combined aqueous layers were carefully neutralized with solid Na_2CO_3 (added portionwise), and the aqueous phase was extracted with EtOAc (3×100 mL). The combined organic extracts were dried over Na_2SO_4, filtered, and concentrated under reduced pressure to afford a brown crude solid. The crude product was purified by flash column chromatography (EtOAc/hexanes, 20–80%) to afford the product as a light yellow solid (50%). ^{1}H NMR (300 MHz, $CDCl_3$) δ 11.37 (s, 1H), 8.63 (m, 1H), 8.41 (d, J = 7.5 Hz, 1H), 8.26 (s, 1H), 8.11 (d, J = 7.5 Hz, 1H), 7.99 (d, J = 10.2 Hz, 1H), 7.28 (m, 1H), 4.00 (s, 3H). ESI-MS: 240.6 $[M+H]^+$.

S10.16 1,9,10-Triaza-7-phenanthrenecarboxylic acid (S15)

A solution of methyl 1,9,10-triaza-7-phenanthrenecarboxylate (**S14**, 1.50 g, 62.76 mmol) in MeOH (59 mL) was treated with 1 M NaOH (12.6 mL, 125.52 mmol) and heated at 90 °C for 2 h. After completion, a solution of oxalic acid (2.0 equiv) in H_2O (10 mL) was added dropwise, and the mixture was stirred for 5 min. The solvent was partially removed under reduced pressure, and CH_2Cl_2 was added. The resulting solid was collected by filtration, washed with CH_2Cl_2 and H_2O, and dried on high vacuum to afford the product as a brown solid (85%). ^{1}H NMR (300 MHz, DMSO-d_6) δ 11.5 (br s, 1H), 8.45 (m, 2H), 8.15 (m, 2H), 7.8 (m, 1H), 7.20 (br s, 1H).

S10.17 (1,9,10-Triaza-7-phenanthryl)methanol (S16)

1,9,10-Triaza-7-phenanthrenecarboxylic acid (**S15**, 1.2 g, 5.33 mmol) was suspended in anhydrous THF (20 mL) and cooled to 0–5 °C under an inert atmosphere. A solution of BH_3·THF complex (27 mL, 26.677 mmol) was added dropwise at 0–5 °C. The mixture was stirred for 10 min at this temperature, then allowed to warm to room temperature and stirred for 16 h. The reaction mixture was cooled to 0 °C and carefully quenched by the slow addition of MeOH. After stirring for 10 min, the solvent was removed under reduced pressure. MeOH was added, and the mixture was concentrated again under reduced pressure to ensure complete removal of boron residues. The crude product was purified by flash column chromatography (0–15% MeOH in CH_2Cl_2) to afford the product as a light yellow solid (80%). ^{1}H NMR (300 MHz, CD_3OD) δ 8.65 (d, J = 7.8 Hz, 1H), 8.39 (d, J = 6.0 Hz, 1H), 8.15 (d, J = 8.1 Hz, 1H), 7.72 (s, 1H), 7.38 (m, 1H), 4.82 (s, 2H).

S10.18 (1,9,10-Triaza-7-phenanthryl)methanol hydrochloride (S2)

(1,9,10-Triaza-7-phenanthryl)methanol (**S16**, 840 mg, 3.98 mmol) was dissolved in MeOH (20 mL) and cooled to 0 °C. Concentrated HCl (3 mL) was added dropwise at 0–5 °C, and the reaction mixture was stirred for 2 min at this temperature, then allowed to warm up to room temperature and stirred for an additional 5 min. All volatiles were removed under reduced pressure. The resulting solid was washed with CH_2Cl_2 (3×20 mL) and dried under high vacuum overnight to afford the hydrochloride salt as a light yellow solid (80%). ^{1}H NMR (300 MHz, CD_3OD) δ 9.10 (d, J = 7.5 Hz, 1H), 8.49 (d, J = 6.0 Hz, 1H), 8.30 (d, J = 8.4 Hz, 1H), 7.77 (s, 1H), 7.69 (m, 1H), 7.51 (d, J = 8.1 Hz, 1H), 4.86 (s, 2H). HPLC purity: 99.7%.

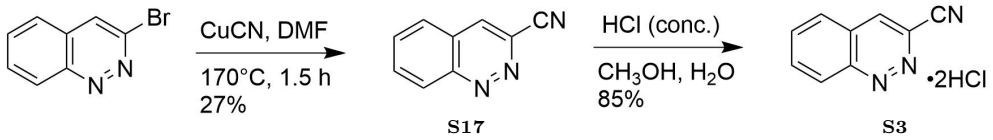


Fig. S43 Synthetic route to 3-cinnolinecarbonitrile hydrochloride (**S3**).

S10.19 3-Cinnolinecarbonitrile (S17)

CuCN (430 mg, 4.798 mmol) was added to a solution of 3-bromocinnoline (500 mg, 2.399 mmol) in DMF (8 mL) in a sealed microwave vial. The reaction mixture was heated at 170 °C under microwave irradiation for 1.5 h. After cooling to room temperature, 2 N HCl was added, and the mixture was stirred for 10 min at room temperature. The solution was then neutralized with 1 N NaOH and extracted with EtOAc. The organic layer was separated, dried over Na_2SO_4, and concentrated *in vacuo*. The crude mixture was purified by flash column chromatography (10–40% EtOAc/hexanes) to

afford the product as a brown solid (27%). ^{1}H NMR (300 MHz, CD_3OD) δ 8.72 (d, J = 8.7 Hz, 1H), 8.34 (s, 1H), 8.10 (m, 1H), 7.95 (m, 2H).

S10.20 3-Cinnolinecarbonitrile hydrochloride (S3)

3-Cinnolinecarbonitrile (**S17**, 100 mg, 0.64 mmol) was dissolved in MeOH (3 mL) and cooled to 0 °C. Concentrated HCl (0.5 mL) was added dropwise at 0 °C, and the reaction mixture was stirred for 2 min at this temperature, then allowed to warm to room temperature and stirred for an additional 5 min. All volatiles were removed under reduced pressure. The resulting solid was washed with CH_2Cl_2 (3×5 mL) and dried under high vacuum overnight to afford the hydrochloride salt as a brown solid (85%). ^{1}H NMR (300 MHz, CD_3OD) δ 8.85 (s, 1H), 8.65 (d, J = 8.7 Hz, 1H), 8.20 (m, 2H), 8.10 (m, 1H). ESI-MS: 156.5 $[M+H]^+$. HPLC purity: > 99.9%.