SPARK: Learning unseen chemistry to plan experiments

Bing Yan [1,2] Zachary Ulissi [2] Benjamin Kurt Miller [2,*]
Kyunghyun Cho [1,*]

[1]Department of Computer Science, New York University, New York, NY 10011, USA.
[2]Meta FAIR.
[*]Correspondence: bkmi@meta.com, kyunghyun.cho@nyu.edu.

AI models are widely used to predict experimental outcomes. Here we use AI to infer the mechanisms behind these outcomes and plan measurements to distinguish competing explanations. We introduce SPARK, a simulation based inference framework that learns a joint posterior over kinetic parameters. Continuous changes in reaction rates connect related mechanisms within a shared network, allowing parameter uncertainty and pathway activity to be inferred together. For cyclic voltammetry, a neural signal encoder and flow matching model infer 42 parameters in a network of 19 elementary reactions. Posterior mean predictions reconstruct simulated signals with a median error of 4.3% of the current range, without conditioning on a pre-selected mechanism. SPARK's posterior sampling after training is orders of magnitude faster than L-BFGS-B fitting. While multiple mechanisms can explain the same signal, they can lead to different chemical states under new conditions. SPARK uses these differences to plan experiments to identify the correct mechanism. To demonstrate this, we revisit published electrochemical and thermal studies and discover unresolved reaction steps and surface populations, yielding quantitative predictions for new measurements. SPARK moves computation from explanation to planning: inferring the chemistry that measurements leave unseen and proposing experiments that can reveal it.

Introduction

Scientific understanding advances by proposing explanations for what we observe and designing experiments to test them. In chemistry, experiments can track how the amounts of different molecules change over time without directly revealing the pathways that form or consume them. A reaction mechanism describes the molecular events behind these pathways in terms of individual steps, called elementary reactions. Yet different mechanisms can produce nearly identical measurements. The challenge is to identify which explanations remain possible and what to measure next to distinguish them.

Conventional kinetic analysis usually selects a mechanism in advance and then fits its parameters to the measured signal. This approach is widely used to analyze kinetic data, including voltammograms,[1,2] Tafel plots and catalytic turnover frequencies. A close reconstruction can constrain parameters within the chosen mechanism while leaving alternatives unexplored. Many named mechanisms can be represented with a shared reaction network built from a library of elementary reactions.[1,3] Varying the rate constants of these steps continuously connects related mechanisms, allowing their pathways and parameters to be inferred together. This representation also includes mixtures and intermediate cases that do not fit a single named mechanism.

A shared reaction network turns mechanism identification into a search for parameter combinations that explain the measurements. Simulation based inference learns from physical simulations to estimate how much probability to assign to each region of possible parameter combinations after observing a measurement.[4,5] This distribution is called a joint posterior: it describes uncertainty about all parameters together, including which combinations are more plausible.

Here we introduce SPARK (Simulation based Posterior Amortization for Reaction Kinetics) to connect mechanistic inference with experimental planning. SPARK retains competing mechanisms as different rate combinations within a joint parameter posterior. We evaluate parameter recovery and signal reconstruction

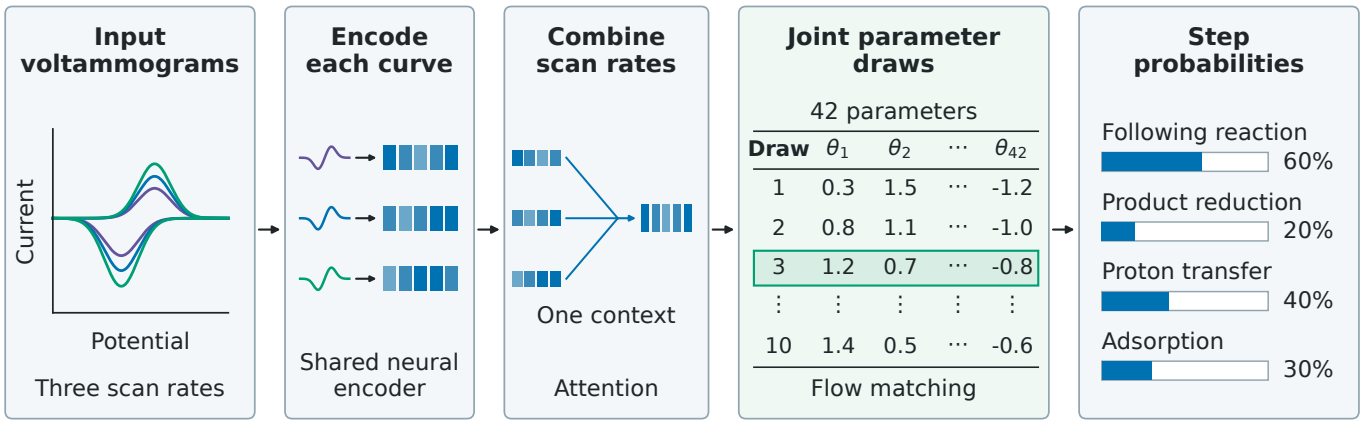


Figure 1: SPARK infers reaction parameters and elementary step probabilities from voltammograms. A shared encoder reads the input curves, an attention module combines their information, and a flow matching estimator generates joint posterior draws. The schematic table shows each draw as one row of 42 parameter values, with one parameter per column. The highlighted row illustrates a complete set whose values stay together during simulation. Only selected rows and columns are displayed. Step probabilities count the fraction of all draws in which the relevant rate or adsorption constant exceeds its activity threshold. Supplementary Table S16 maps combinations of active steps to common named mechanisms.

using simulations with known parameters. We then simulate the inferred parameter combinations under new conditions to identify where competing explanations predict different responses. These differences guide the choice of new measurements to identify the correct mechanism.

We develop SPARK for cyclic voltammetry (Figure 1) and extend the framework to hydrogen evolution and CO oxidation. We demonstrate by literature case studies that SPARK identifies reaction steps and surface states that existing measurements leave unresolved, and then predicts how new experiments could distinguish the alternatives. SPARK thus moves computation from explanation to experimental planning: using existing data to explore unseen chemistry and decide what to measure next.

The SPARK framework

We use cyclic voltammetry to illustrate SPARK (Figure 1). Given one or more voltammograms, SPARK estimates a joint posterior over 42 parameters describing reaction kinetics, thermodynamics, diffusion, surface adsorption and double layer capacitance (Table S15). The prior describes the parameter combinations considered before the curves are observed. The posterior updates their probabilities using the information in the curves, within the assumed reaction model. Several combinations can remain possible when the measurements do not uniquely determine the chemistry.

Physical simulations of a shared network of 19 elementary reactions provide the training data. Each training example pairs one known parameter set with three voltammograms at different scan rates. A shared neural encoder reads each curve, and an attention module combines their information. Flow matching trains a neural network to transform a random vector of 42 numbers into a complete parameter set, guided by the encoded curves.[6] All 42 numbers are transformed together, so the model can learn how the parameters depend on one another.

Repeating this transformation produces posterior draws, shown as rows in Figure 1. Each row contains all 42 parameters for one possible reaction system, and each column corresponds to one parameter. A histogram of one column shows that parameter's distribution; plotting two columns together reveals how their values depend on each other. For example, a faster rate along one pathway may require a slower rate along another to explain the same signal. We preserve these relationships by simulating complete rows under new conditions to identify where their predictions differ. A step's activity probability is the fraction of draws in which its rate or adsorption constant exceeds a defined threshold.

The training prior uses separate ranges for negligible and active steps, with uniform sampling within each range (Supplementary Table S2). This allows the shared network to represent individual pathways and mixtures. Varying the electron transfer rate constant connects reversible and quasireversible electron transfer

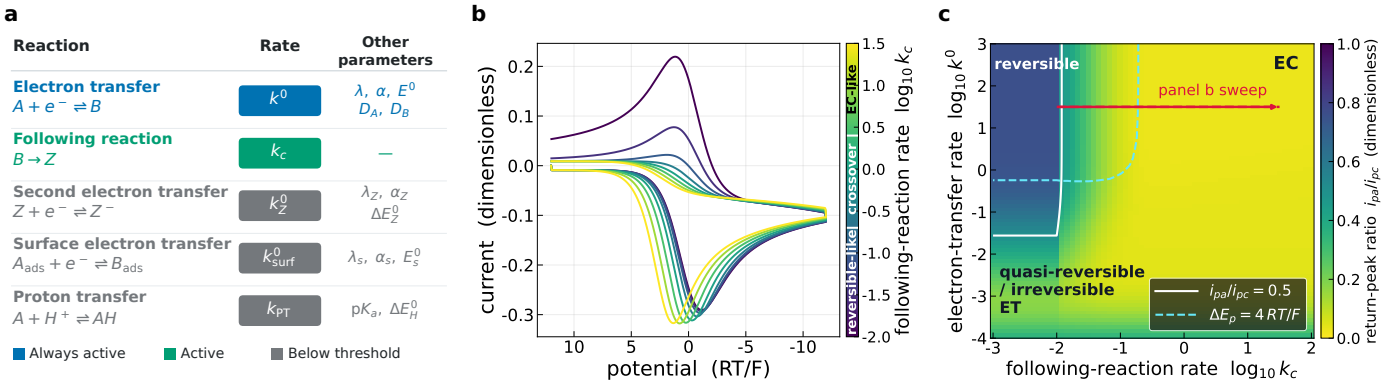


Figure 2: A shared reaction network connects different mechanisms. (a) The primary electron transfer and following chemical reaction are active, while the other steps lie below their activity thresholds. The network includes parameters for reaction kinetics, thermodynamics, diffusion and surface adsorption. (b) Increasing the following reaction rate constant, k_c, continuously suppresses the return peak. (c) A zone diagram shows how the response changes as k_c and the electron transfer rate constant, k^0, vary. Color indicates the return peak ratio. Supplementary Fig. S1 shows additional zone diagrams for other mechanisms and parameter combinations.

(E), while increasing the rate constant of a following chemical step connects E to electron transfer followed by chemistry (EC, Figure 2).[3] Additional zone diagrams illustrate continuous changes between other mechanisms (Figure S1).

Results

Testing the inferred chemistry

Each joint posterior draw describes a complete reaction system that can be simulated under the input conditions. We test these posterior predictions by comparing the mean simulated current at each point with the input curves. Reconstruction error measures the root mean squared difference as a percentage of the current range, with smaller values for closer agreement. Across 120 simulated examples, SPARK achieves a median error of 4.3% without a supplied mechanism (Table 1). Supplementary S2.1 defines this and the following metrics.

Reconstructing a signal does not establish that its underlying chemistry has been recovered. We therefore test step and parameter recovery on 3,000 examples, whose parameters are known. For step detection, we call a step active when it is active in more than half the draws. SPARK recovers the complete set in 56% of examples and achieves micro-$F_1 = 0.78$. This score penalizes both missed and falsely included steps, and 1 means perfect detection.

For each parameter, we compare the median of its draws with the known value using R^2: 1 means exact recovery, while 0 always predicting the average value. When the associated process is present, capacitance, the primary electron transfer rate constant and reactant diffusivity achieve $R^2 = 0.83$, 0.76 and 0.72, respectively. Transfer coefficients, reorganization energies and product diffusivities are less well recovered (Supplementary Table S3). The draws must therefore describe what remains uncertain, as well as provide an estimate.

A single estimate can conceal distinct possibilities. In Figure 3a, a rate constant's posterior has two peaks, with its median between them. The 90% highest density region (HDR) keeps the most likely ranges separate, and together, they contain 90% of the posterior probability. Their combined width relative to the full prior range measures the remaining uncertainty. A wide region leaves many parameter values plausible; its width alone does not measure inference accuracy. Median widths are 15% for capacitance and 86–91% for product diffusivities (Figure 3b). The smaller widths for rate constants partly reflect their separated prior ranges (Supplementary S2.2).

We then test whether the probabilities assigned to these possibilities match observed frequencies. This is calibration: intervals spanning the middle 90% of draws should contain the known value in about 90 of 100 examples (Figure 3a,c). Average coverage is 91% on the 3,000 examples. Even an unchanged prior can have

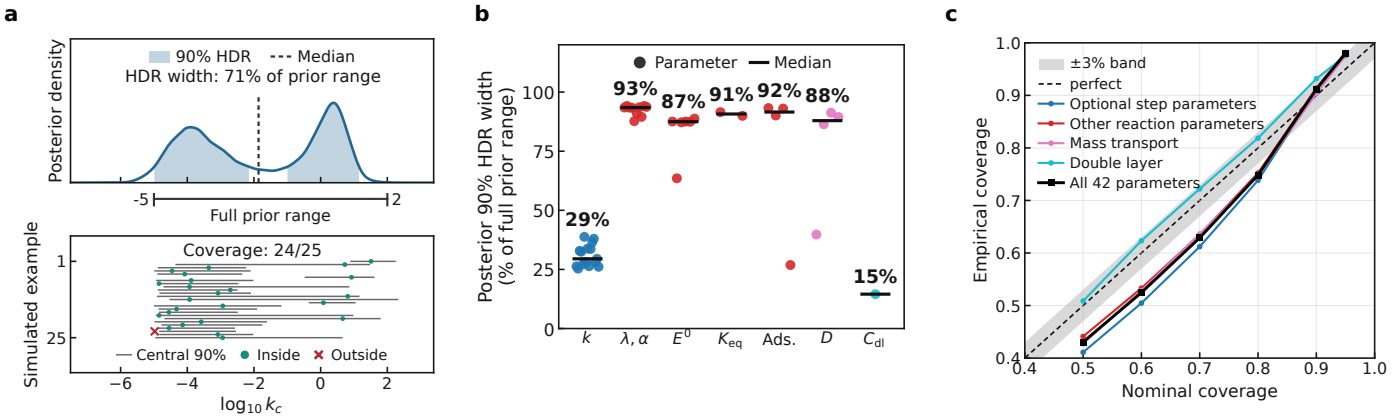


Figure 3: How tightly SPARK constrains parameters and how reliably it reports uncertainty. (a) Top, a posterior with two peaks for the following reaction rate, $\log_{10} k_c$. The shaded 90% highest density region (HDR) keeps two separate intervals whose combined width is 71% of the full prior range. The median lies in the gap. Bottom, central 90% intervals for the first 25 simulated examples of the same parameter. Points mark known values inside (green) or outside (red) each interval; 24 of 25 are inside. This illustrates how coverage is counted; these central intervals are distinct from the HDR above. (b) Median posterior HDR width for each of the 42 parameters, grouped by parameter family and expressed as a percentage of its full prior range. Symbols denote rate constants (k), reorganization energies and transfer coefficients (λ, α), redox potential coordinates (E^0), chemical equilibrium parameters (K_{eq}, including the pK_a offset), adsorption (Ads.), diffusion (D) and capacitance (C_{dl}). Each dot represents one parameter; black horizontal marks and bold percentages show group medians. Horizontal offsets separate overlapping dots. (c) Assigned central interval probability and the fraction of known values inside those intervals across 3,000 model development examples. Curves average over the four calibration groups defined in Supplementary S2.2, and over all 42 parameters. Agreement with the diagonal means that assigned probabilities match observed frequencies. Supplementary S2.1 defines the metrics; S2.2 gives prior distributions, width comparisons and further probability checks.

correct coverage without learning from the curves. HDR width and calibration therefore assess both the range of remaining possibilities and the reliability of their probabilities.

These checks explain why signal reconstruction alone cannot guide model design (Table 1a). Replacing flow matching with a neural spline flow improves reconstruction but lowers the minimum 90% interval coverage to 59%. Replacing the learned curve encoder with predefined summaries reduces step recovery. The learned encoder and flow matching balance signal reconstruction, step recovery and uncertainty.

Direct optimization provides a further comparison (Table 1b). Given the true mechanism, L-BFGS-B reconstructs signals more closely, with 1.3% error; a random mechanism raises this to 6.0%. Optimizing the full network without a supplied mechanism gives 8.9% error on a separate 120-signal benchmark, with the paired SPARK comparison in Supplementary S5. After training, SPARK generates 1,000 joint draws in 0.091 s on a GPU, excluding forward simulation. A single full network optimization start takes 425–450 s on a CPU; most reach the time limit. Table 1 gives all three optimization settings, and Supplementary S5 reports convergence and timing details. SPARK makes alternative parameter combinations available for further simulation, together with their uncertainty.

From competing explanations to new measurements

Posterior draws can also guide a new measurement. We group them by their active steps and ask where the groups predict different responses. Three simulated voltammograms at pH 7, excluded from training, illustrate this approach (Figure 4). SPARK assigns 49% posterior probability to electron transfer alone and 24% to electron transfer with proton transfer (PCET). Each probability is the fraction of joint draws in that group. Mean predictions from both groups reconstruct the input curves, with median errors of 2.1% and 1.9% of the current range. The measured currents leave the role of proton transfer unresolved.

This difference suggests changing the proton concentration. We simulate joint parameter draws from each group across pH at a scan rate ten times slower than the slowest input curve. At this slower rate, the two explanations predict greater separation at pH 0 than at pH 7, relative to the spread within each group (Figure 4c,d). SPARK thus proposes a specific measurement: a slower voltammogram at pH 0. The uncertainty left by the original measurements becomes a concrete proposal for testing the competing explanations.

Table 1: Model components and comparison with parameter optimization. (a) Coverage gives the mean and minimum across parameters for central 90% intervals. (b) The true and random mechanism fits use the same 120-signal test set as a. Errors are median percentages of the current range. Sampling time covers 1,000 GPU draws after training. Optimization times are measured ranges for one signal based first start, restricted to complete simulated signals and including timeouts. Supplementary S2.1 defines the metrics. Methods and Supplementary S5 describe the protocols.

a Model components

Model	Step detection micro-F_1	90% coverage mean / minimum	Reconstruction error (%)	Sampling time 1K joint draws (s)
SPARK: signal + flow matching	0.76	0.90/0.86	4.3	0.091
Signal + neural spline flow	0.71	0.90/0.59	2.6	0.025
Curve summaries + flow matching	0.62	0.91/0.86	7.3	0.091

b Parameter optimization

Method	Mechanism supplied	Error (%)	One-start time range (s)
L-BFGS-B	None (full network)	8.9	425–450
L-BFGS-B	True	1.3	124–506
L-BFGS-B	Random	6.0	124–507

Testing published chemical interpretations

We next ask what SPARK can add to published chemical interpretations. For Co–ClCN, SPARK identifies reaction sequences that guide a more detailed chemical network. We also train posterior estimators for the hydrogen evolution and CO oxidation reactions. Each case connects inferred parameter combinations to a measurement that tests the unresolved chemistry (Figure 5). Supplementary S3 describes the implementation, simulated evaluation and experimental inference for each model.

Co–ClCN: tracing the return current

Sheng et al.[7] studied how a cobalt porphyrin activates the alkyl chloride ClCN, written here as RCl. After the electrode reduces Co^{II} to $[Co^{I}]^{-}$, two routes can form the same cobalt–alkyl product. (1) Direct reaction with RCl gives R–Co^{III} and Cl^{-}. (2) The radical route first releases $R^{\bullet}$, Cl^{-} and Co^{II}, the radical is then captured by Co^{II} to form R–Co^{III}. In the paper, it remains unresolved how the radical's further reductions contribute to the CVs.

We first apply the same trained, 42-parameter flow matching SPARK model used in the preceding sections to each measured CV. At 26–106 equivalents, its posterior strongly favors sequences with further electron transfer. At 53 equivalents, 87% of draws contain electron transfer followed by chemistry and product reduction (ECE), and 72% contain two successive electron transfers (EE). These sequences can coexist with coupling reactions in the same draw (Supplementary Fig. S4). The general model thus points to further reduction after product formation.

Guided by these inferred sequences and the source's formation chemistry, we build an explicit network of eight steps (Figure 5a). The radical and cobalt–alkyl complex can each accept another electron from the electrode or from dissolved $[Co^{I}]^{-}$. A new estimator with 17 parameters learns from 16,384 simulated CVs. On 160 simulated cases evaluated after model selection, its median reconstructed CV error is 0.68%, and nominal 90% parameter intervals have 88.3% mean coverage (Supplementary S3.1).

At 26–106 equivalents of ClCN, the full posterior gives errors of the median reconstructed curves of 2.5–5.2%. Keeping only unchanged draws with small reference rates for all four later reductions removes the return peak and raises errors to 17–19% (Supplementary Fig. S5). Therefore, SPARK extends the literature's reaction network into a quantitative test of the following electron transfers.

At 26 and 53 equivalents, the inferred cobalt–alkyl redox potential is near −1.28 V versus Fc/Fc^{+}. Voltammetry of an independently generated cobalt–alkyl complex would test whether the predicted redox response belongs to that product. The blank control also reveals inference bias in this potential, making the direct product measurement decisive for its chemical assignment (Supplementary S3.1).

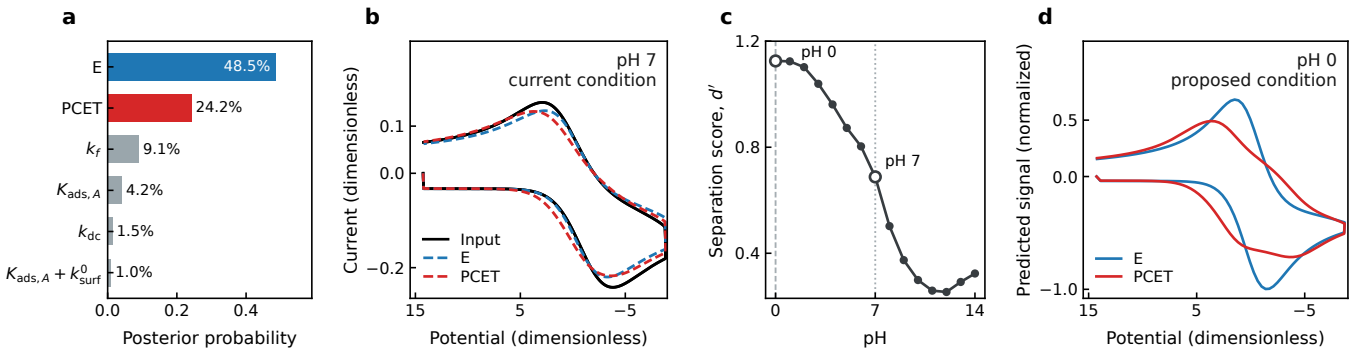


Figure 4: Competing explanations guide the next measurement. (a) Mechanism probabilities inferred from three synthetic voltammograms at pH 7. (b) Mean reconstructions for electron transfer alone (E, blue) and electron transfer with proton transfer (PCET, red), compared with the synthetic input (black) at the current condition, pH 7 and $\sigma = 0.804$. (c) Predicted separation across pH at $\sigma = 0.0804$, ten times slower than the input in b. The score compares the difference between mechanism means with their spread, rising from 0.69 at pH 7 to 1.13 at pH 0. (d) Mean predicted responses at the proposed pH 0 and $\sigma = 0.0804$. In d, curves are peak normalized before averaging within each group. Panels c and d describe proposed measurements; uncertainty enters the separation score defined in Supplementary S2.1.

Hydrogen evolution: testing coverage

Jung et al.[8] measured the potentials needed to drive hydrogen evolution on Pt across pH and current. The Volmer step forms adsorbed hydrogen, H*. Hydrogen can then leave through an electrochemical Heyrovsky step or by Tafel recombination of two H* species. These steps give the VH, VT and VHT pathways.

SPARK uses the paper's rate equations in a model with 11 parameters that solves for potential and H* coverage at each applied current. We train a new estimator on 10,966 simulated examples, each containing potentials at the 120 measured pH and current conditions. Across 1,397 simulated validation examples, SPARK identifies the HER pathway in 94.5% of cases.

For the literature signals, SPARK predicts 82% VH and 18% VHT, with less than 0.1% VT. The mean reconstructed potentials from successful simulations have an error of 15 mV (Supplementary S3.2). Thus, the inference favors hydrogen release through Heyrovsky while retaining a possible Tafel contribution. Both the VH and VHT fits reproduce the potentials with similar errors, 11.6 and 10.7 mV, yet predict different H* coverages (Figure 5b, Supplementary Fig. S6).

The next measurement should therefore probe the surface adsorbed H* directly. At the same applied current of 1.273 mA cm^{-2}, measuring H* coverage across pH would test the more saturated surface predicted by VHT against the lower, more variable coverage predicted by VH. A VT pathway instead predicts coverage that stays constant with pH at fixed current. SPARK turns the mechanism uncertainty acknowledged by Jung et al. into a specific comparison of surface populations (Supplementary S3.2).

CO oxidation: a coverage test for oxygen activation

Allian et al.[9] studied how O_2 reacts on Pt during CO oxidation. O_2 can adsorb and split before reacting with CO (direct activation), or react with adsorbed CO during activation (CO-assisted activation). Both routes explain the measured rates, while DFT calculations and measured activation energies favored CO assistance. Their infrared measurements supported a CO-covered surface but did not establish its coverage at lower controlled CO pressures.

We train a separate, 16-parameter SPARK model that includes both routes and allows CO coverage to vary. Inference uses the published rates, CO desorption measurements and uncertain DFT energies, without using CO coverage. SPARK reconstructs the rates (Supplementary Fig. S7) and favors CO assistance, while retaining parameter draws in which direct activation dominates. Training, evaluation and route probabilities are given in Supplementary S3.3; Supplementary S4 defines the energetic inputs.

The new prediction concerns where these explanations imply different CO coverages. We follow the same direct and CO-assisted parameter groups across pressure and temperature (Figure 5c). At 0.01 kPa CO and

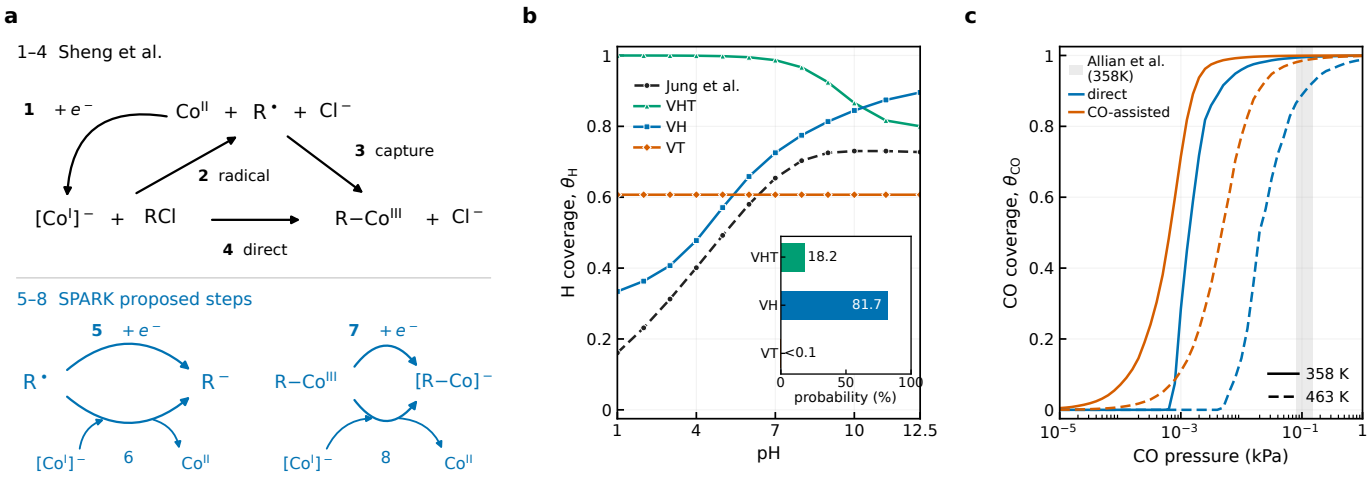


Figure 5: Reaction products and surface coverage provide tests of published mechanisms. (a) Co–ClCN network, with RCl = $Cl(CH_2)_3CN$ and TPP omitted. Black shows the formation routes discussed by Sheng et al.; blue shows the additional reduction network proposed by SPARK to explain the return current. The side arrows in steps 6 and 8 show $[Co^I]^-$ donating an electron and returning as Co^{II}. Electron transfer is reversible in the model; arrows show forward directions. (b) H* coverage predicted by separate VHT, VH and VT fits at 1.273 mA cm^{-2}. V (Volmer) supplies H*; H (Heyrovsky) releases H_2 using one H*; T (Tafel) combines two H*. Gray uses Jung et al.'s rounded parameters. The inset shows SPARK probabilities with priors centered on the fits; the curves use individual fitted parameter sets. (c) Median CO coverage for the direct and CO-assisted parameter groups at 358 K (solid) and 463 K (dashed), with 10 kPa O_2. Groups remain fixed at the 358 K reference. Gray marks the infrared pressure range reported by Allian et al. at 358 K. Supplementary S3 gives reconstructions, parameter checks and uncertainty for all three cases.

10 kPa O_2, their median coverages are similar at 358 K: 96% and 99%, respectively. At 463 K, they fall to 15% and 77%. This difference suggests measuring CO coverage across pressure at both temperatures to test the predicted loss of adsorbed CO. The experiment should avoid complete CO consumption and measure gas composition at the catalyst, so the predictions can be tested at known CO pressures. The prediction intervals overlap; Supplementary S3.3 gives their widths and the measurement protocol.

Discussion

SPARK turns uncertainty about a reaction mechanism into a basis for choosing the next measurement. A shared network retains related explanations, and joint parameter draws preserve the combinations that support them. Simulating these draws reveals consequences beyond the measured signal. The pH example shows how changing an experimental condition can separate predicted responses. The published studies show why measuring a product or surface population can address chemistry that the kinetic signal leaves unresolved.

The reactions and prior ranges supplied to SPARK define which explanations it can consider. Calibration and chemical checks show how inference errors and prior choices affect the results (Supplementary S2.2; S3). New measurements could narrow the remaining alternatives or reveal the need to expand the reaction network.

SPARK connects the explanation of an observation to a test of the chemistry behind it. Computation becomes a way to explore what remains unseen and plan the experiment that can reveal it.

Methods

Simulating voltammograms

The general CV model contains 19 elementary reactions described by 42 parameters. The simulator calculates reactions in solution and at the electrode together with diffusion, adsorption and the electrical response of the cell. Double layer capacitance is inferred with the reaction parameters. Uncompensated resistance is sampled once per example and shared across its scans; it enters the simulated voltage drop but is not one of the 42

inferred parameters. Supplementary S1 gives the equations and numerical methods; Supplementary S6 lists the reactions.

We sample parameter sets from the ranges in Supplementary Table S2. Optional reactions have separate ranges for negligible and active rates, allowing individual pathways and mixtures. Activity is defined by a reference rate or adsorption constant, rather than by its effect on a particular curve. The training library contains 200,000 examples, each with three scan rates. A separate 15,000-example library supports model development and evaluation. Measurement noise and baseline drift are added during training (Supplementary S1).

Learning and sampling the posterior

A shared neural encoder reads each voltammogram, and an attention layer combines the curves without depending on their order.[10] Training uses three curves per example; the encoder can accept one or more curves at inference. Flow matching learns to transform Gaussian random samples into parameter sets consistent with the input curves.[6] Each resulting draw contains all 42 parameters, preserving their relationships. The training objective and network settings are given in Supplementary S1.

A step's probability is the fraction of joint draws in which its reference rate or adsorption constant exceeds the activity threshold. Draws with the same complete set of active steps form a mechanism group; its probability is its fraction of all draws. Group probabilities are counted before forward simulation or any reconstruction check. Supplementary Table S16 maps step combinations to common mechanism names, and Supplementary S2.3 defines the grouping rules.

Checking the inferred chemistry

Simulations with known parameters allow us to check both the inferred chemistry and its uncertainty. Parameter recovery compares posterior medians with the known values using R^2. Parameters belonging to optional reactions are scored when a reaction using them is enabled in the generating example. Step detection uses micro-F_1, which counts both missed and falsely included steps; exact recovery requires the complete set to be correct. The main recovery analysis uses 3,000 development examples.

We assess uncertainty in two ways. The width of the 90% highest density region measures the range of likely parameter values, as a fraction of the full prior range. Interval coverage checks how often the known value falls within the stated probability interval. A separate set of 3,000 newly simulated examples tests whether known values occupy the expected positions among posterior draws. Supplementary S2.1 defines the metrics, and S2.2 gives the probability checks. Supplementary Table S1 lists the examples and draw counts for each analysis.

For CV reconstruction, we simulate complete joint draws, average their currents and calculate the root mean squared error divided by the measured current range. We report the median across scans and then across examples. The generic CV benchmarks supply the generated resistance for reconstruction. Draws are chosen before simulation; failures are recorded, and parameters are not adjusted to improve agreement. The case studies state their own signal units, averaging rules and treatment of failed simulations in Supplementary S3.

Comparing models and optimization

We test the two main model components by replacing flow matching with a neural spline flow,[11] or replacing the learned curve encoder with 21 predefined signal summaries (Table 1a). Step detection and interval coverage use the same 9,005 simulated examples. Signal reconstruction uses 120 examples and up to 16 joint draws per example. Supplementary S5 gives the losses, summary definitions and additional comparisons.

L-BFGS-B provides a comparison with direct parameter optimization.[12] It searches within specified parameter bounds under three settings: the true mechanism, a random mechanism fixed before fitting, or the full network without a supplied mechanism. We report reconstruction error for the complete prediction with the lowest optimization objective found across eight starts. The full network search varies all 42 parameters and is compared with the unchanged SPARK model on the same 120 signals. The true and random mechanism errors use the earlier reconstruction set. Parameter constraints, search limits and convergence results are given in Supplementary S5.

Sampling time covers 1,000 joint draws on a GPU after training and excludes forward simulation. Optimization time is reported separately for one start on a CPU, including starts stopped at the time limit that retain complete predictions. These timings describe different workloads; the optimization errors use eight starts. Supplementary S2.1 and Supplementary Table S14 define the timing comparison.

Choosing the next measurement

We first reconstruct the input signal from each mechanism group's joint draws. Groups that reproduce the input with similar errors provide competing explanations. We then simulate these same explanations under new conditions and compare the difference between their mean signals with the spread within each group. The score d' is larger when the predictions separate more clearly (Supplementary S2.1).

For the simulated example in Figure 4, one group includes proton transfer and the other does not. We therefore vary pH from 0 to 14 at a scan rate ten times slower than the slowest input curve. Each predicted current is divided by its largest absolute value before scoring, so the comparison measures differences in curve shape. The largest score identifies pH 0 as the proposed condition. Supplementary S2.3 gives the reconstruction checks and full protocol. The literature cases instead compare product redox behavior or surface coverage, as described below.

Models for the published experiments

The Co–ClCN study first applies the same general CV model used in the main evaluation. Its inferred sequences guide an explicit eight-step chemical network, for which we train a separate 17-parameter estimator. Each titration then conditions its own posterior. The comparison with weak later reductions retains unchanged posterior draws below the stated rate thresholds, without refitting (Supplementary S3.1).

For hydrogen evolution, an 11-parameter model uses Jung et al.'s rate equations to calculate potential and hydrogen coverage at each imposed current.[8] Training examples use parameter ranges centered on separate fits to the measured potentials. The inferred probabilities therefore depend on ranges chosen using those same measurements. The main estimator samples all 11 parameters together and counts their active steps to obtain pathway probabilities. The coverage curves in Figure 5b come from the separate fits; the inset probabilities come from SPARK (Supplementary S3.2).

For CO oxidation, a 16-parameter model includes direct and CO-assisted oxygen activation and solves for surface populations without assuming CO saturation. Each catalyst's rates, CO desorption measurements and uncertain DFT energies condition a separate posterior; CO coverage is not supplied. We group draws by the route carrying most O_2 activation at a reference condition and follow those same groups across pressure and temperature. Supplementary S3.3 gives the model and prediction protocol; Supplementary S4 specifies the energetic inputs. The three specialized estimators use neural spline flows; Supplementary S3 gives their training and evaluation on simulated examples.

Data availability

The Jung et al. HER tables and microkinetic source are distributed with the source publication,[8] and the Allian et al. kinetic curves are digitized from the author manuscript.[9] Processed raw-point campaigns, source-file hashes, atomistic-anchor manifests, immutable benchmark manifests and numerical source data for reported figures are provided in the project repository. Generated training and test libraries will be deposited in a persistent public archive before publication.

Code availability

SPARK source code, simulator definitions, training and evaluation scripts, figure-generation code and environment documentation will be released through the project repository. Versioned checkpoints and commands used for the reported analyses will accompany the archived data release.

Author contributions

B.Y. developed SPARK, implemented the code, conducted the analyses and drafted the manuscript.
Z.U. proposed the case studies and contributed to writing.
B.K.M. guided the simulation based inference methodology, supervised the project and contributed to writing.
K.C. supervised the project and contributed to writing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.K.M. and K.C. at `bkmi@meta.com` and `kyunghyun.cho@nyu.edu`.

References

[1] Bard, A. J. & Faulkner, L. R. *Electrochemical Methods: Fundamentals and Applications* (John Wiley and Sons, 2001), 2nd edn.

[2] Compton, R. G., Laborda, E. & Ward, K. R. *Understanding Voltammetry: Simulation of Electrode Processes* (Imperial College Press, 2014).

[3] Savéant, J.-M. *Elements of Molecular and Biomolecular Electrochemistry: An Electrochemical Approach to Electron Transfer Chemistry* (John Wiley and Sons, 2006).

[4] Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences* **117**, 30055–30062 (2020).

[5] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22**, 1–64 (2021).

[6] Wildberger, J. *et al.* Flow matching for scalable simulation-based inference. In *Advances in Neural Information Processing Systems*, vol. 36, 16837–16864 (Curran Associates, Inc., 2023).

[7] Sheng, H. *et al.* Autonomous closed-loop mechanistic investigation of molecular electrochemistry via automation. *Nature Communications* **15**, 2781 (2024).

[8] Jung, O., Jackson, M. N., Bisbey, R. P., Kogan, N. E. & Surendranath, Y. Innocent buffers reveal the intrinsic pH- and coverage-dependent kinetics of the hydrogen evolution reaction on noble metals. *Joule* **6**, 476–493 (2022).

[9] Allian, A. D. *et al.* Chemisorption of CO and mechanism of CO oxidation on supported platinum nanoclusters. *Journal of the American Chemical Society* **133**, 4498–4517 (2011).

[10] Lee, J. *et al.* Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 3744–3753 (PMLR, 2019).

[11] Durkan, C., Bekasov, A., Murray, I. & Papamakarios, G. Neural spline flows. In *Advances in Neural Information Processing Systems*, vol. 32, 7511–7522 (Curran Associates, Inc., 2019).

[12] Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208 (1995).

[13] Talts, S., Betancourt, M., Simpson, D., Vehtari, A. & Gelman, A. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788* (2018). URL https://arxiv.org/abs/1804.06788.

[14] Hyndman, R. J. Computing and graphing highest density regions. *The American Statistician* **50**, 120–126 (1996).

[15] Kingma, D. P. & Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, vol. 31, 10215–10224 (Curran Associates, Inc., 2018).

[16] Tejero-Cantero, A. *et al.* sbi: A toolkit for simulation-based inference. *Journal of Open Source Software* **5**, 2505 (2020).

[17] Greenberg, D., Nonnenmacher, M. & Macke, J. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 2404–2414 (PMLR, 2019).

Supplementary Information

Supplementary S1 describes the forward model and additional mechanism transitions. S2 defines the evaluation metrics and reports parameter recovery and calibration checks. S3 documents the three literature case studies, and S4 gives the thermal parameter priors and conditioning model. S5 reports the model comparisons and ablations. S6 lists the outer sphere reactions and their combinations. Figure and table numbers are independent of section numbers.

S1 CV simulation, training and mechanism transitions

The shared network produces continuous changes in voltammogram shape as its rate constants vary. Figure S1 extends Figure 2 with three additional parameter slices.

Outer sphere forward problem. For each diffusing species s with concentration $c_s(x,t)$,

$$\frac{\partial c_s}{\partial t} = D_s \frac{\partial^2 c_s}{\partial x^2} + \sum_r \nu_{sr} r_r(\mathbf{c}, \boldsymbol{\Theta}, E_{\text{elec}}), \tag{1}$$

with initial and far-field concentrations fixed by the sampled campaign. At the electrode, diffusive flux equals the stoichiometric sum of interfacial reaction rates; surface coverages $\boldsymbol{\Theta}$ obey the corresponding site-balance ordinary differential equations. The measured current is

$$i = i_F + C_{\text{dl}} \frac{\mathrm{d}E_{\text{elec}}}{\mathrm{d}t}, \qquad E_{\text{elec}} = E_{\text{programmed}} - iR_u. \tag{2}$$

Here i_F is the signed faradaic current, positive for oxidation. The solver advances concentrations and adsorbed populations on an expanding spatial grid using backward Euler and Newton iteration. An outer secant iteration solves the implicit voltage drop, with capacitive current calculated from the change in electrode potential. Surface conversion from $\mathrm{A_{ads}}$ to $\mathrm{B_{ads}}$ is applied after this update, converting the fraction $1 - \exp(-k_{\text{surfchem}}\Delta t)$ of $\mathrm{A_{ads}}$. All optional $\log_{10}$ rate controls use threshold -2; generated negligible and active rates are separated by an unsampled interval. The reactions and coordinate names are in Table S15; Table S2 gives their generating ranges.

Parameter sampling and scan conversion. Each simulated example contains three scan rates drawn from $\log_{10}\sigma \sim \mathcal{U}(-1, 2)$. One value of uncompensated resistance is sampled per example and shared by all three scans. Resistance affects the signals but is not included in the inferred parameter vector. Before each scan, the generator subtracts $\frac{1}{2}\log_{10}\sigma$ from solution electron-transfer rate coordinates and $\log_{10}\sigma$ from chemical and surface electron-transfer log-rate coordinates and log adsorption constants. Controls assigned negligible behavior are set to -8, and the solver omits steps below its thresholds. Their precise values within the negligible range therefore cannot be recovered from the signal. Scan conversion can also move a reference-active control below the solver threshold. Since the solver time step already depends on scan rate, the physical interpretation of this additional conversion remains unvalidated. The explicit Co–ClCN model instead fixes the simulator constants at the measured scan rate (Supplementary S3.1).

Neural encoder and posterior training. The general CV model uses a shared one-dimensional convolutional encoder for each curve and an attention layer to combine curves independently of their order.[10] The encoder and combined representation both have width 160, with four attention heads. The flow-matching velocity network has five hidden layers of width 256. Parameters are centered and scaled using the midpoint and half-width of the nominal network input ranges, rather than the full generating ranges used for the HDR denominator. Samples are transformed back to the original coordinates without clipping.

For a sampled parameter vector $\boldsymbol{\theta}_1 \sim p_0(\boldsymbol{\theta})$ and simulated observation $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta}_1)$, flow matching connects a Gaussian base sample to the generating parameters along a probability path. At a randomly selected point on that path, the network learns its velocity u_t:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\boldsymbol{\theta}_1,\mathbf{x},\boldsymbol{\theta}_t} \left\| v_\phi(\boldsymbol{\theta}_t, t, e_\psi(\mathbf{x})) - u_t(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_1) \right\|_2^2 . \tag{3}$$

Here e_ψ is the signal encoder. The Gaussian path noise floor is 10^{-3} in the scaled coordinates.[6] Sampling starts from independent standard Gaussian vectors and integrates the learned velocity with 128 Euler steps from the noise endpoint to the parameter endpoint. Each vector remains a complete joint draw.

The saved training configuration uses 200,000 examples, Adam with initial learning rate 3×10^{-4}, cosine learning-rate decay, 60 epochs and batch size 512. Gradients are clipped at norm 5. The separate development library contains 15,000 examples. The flow-matching objective has no classification, interval-coverage or variance penalty. Probability checks use simulations excluded from training. The neural spline baselines use the additional loss described in Supplementary S5; the case-study estimators have the objectives given in Supplementary S3 and S4.

Measurement noise. Fresh noise is applied to the raw currents before their normalization. Polynomial baseline drift, temporally correlated noise and independent Gaussian noise are enabled separately with probabilities 0.5, 0.5 and 0.8 per curve. The peak is the largest absolute current before noise is added. The drift has randomly selected degree 1–3 and Gaussian coefficients scaled by a uniform fraction 0.005–0.05 of the peak; the highest-degree coefficient is halved. It is evaluated on equally spaced points in $[-1, 1]$. These fractions set coefficient scales, rather than the maximum drift. The correlated component filters independent Gaussian values with $y_n = 0.98y_{n-1} + \epsilon_n$, normalizes the resulting standard deviation, and scales it by a uniform fraction 0.005–0.03 of the peak. It is filtered noise, rather than an exact $1/f$ process. The independent Gaussian standard deviation is sampled uniformly from 0.001–0.02 of the peak current. Capacitance and resistance remain physical simulator terms and are not added again as noise.

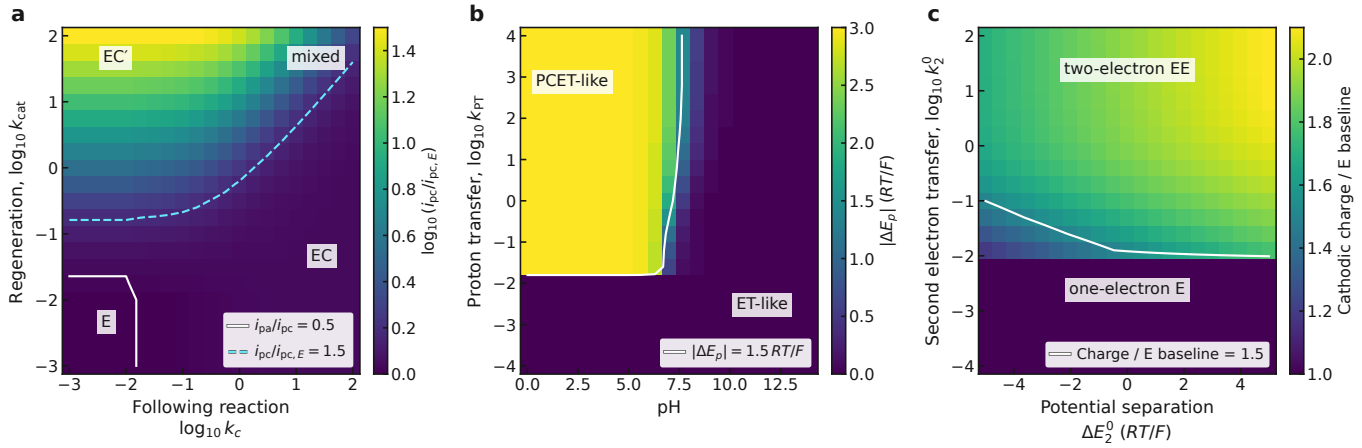


Figure S1: Additional outer sphere mechanism slices. (a) Following chemistry versus catalytic regeneration, colored by peak-current enhancement. (b) Proton-transfer rate versus pH, colored by peak shift. (c) Second electron transfer rate versus formal-potential separation, colored by charge ratio. White and dashed contours mark the response boundaries defined inside each panel.

S2 Evaluation metrics and calibration

The main recovery calculation gives micro-$F_1 = 0.784$ and exact step recovery 0.557. The noise ablation in Supplementary S5 retains the original evaluation draws and therefore reports slightly different values. The recovery and HDR populations share 202 campaigns. The historical ordered 120-case list was not recovered, so its overlap with the new 120-case comparison is not known. The latter uses the first 16 of 512 saved SPARK draws; the same cases also supply the Summary + RQ-NSF reconstruction. For the historical 9,005-case parameter and step comparison, launcher and evaluator defaults differ, and the actual draw-count argument was not retained. The 16-draw reconstruction count is recorded separately.

S2.1 Definitions of the evaluation metrics

The following definitions apply to the main outer sphere evaluation. Let $i = 1, \ldots, N$ index simulated campaigns, $j = 1, \ldots, D$ index parameter coordinates and $m = 1, \ldots, M_i$ index posterior draws. The known

Table S1: Populations used for the outer sphere evaluation. Draw counts are per campaign. The development library contains 15,000 simulations excluded from training; analyses drawn from that library can overlap. Only the locked SBC population below was generated separately for the independent calibration check.

Evaluation	Campaigns	Joint draws	Population and use
Main parameter and step recovery	3,000	512	Development examples; measurement noise and posterior seeds are recorded.
Central coverage curves (Figure 3c)	3,000	512	Original development calculation; historical noise and posterior draws.
HDR widths (Figure 3a,b)	1,000	8,192	Fixed examples from the development library; the first 25 also supply the coverage illustration.
Independent SBC	3,000	255	Separately generated, locked campaigns; independent calibration check.
Model components: steps and coverage (Table 1a)	9,005	Not retained	Clean E and solution chemistry examples used for model comparison.
Earlier reconstruction comparison (Table 1a,b)	120	16	Historical clean E and solution subset; model components and true/random mechanism fits.
Full network comparison (Table 1b)	120	16	Newly fixed clean subset; identical observations for SPARK and optimization, with checkpoints fixed.

generating vector is $\boldsymbol{\theta}_i^\star$, and each $\boldsymbol{\theta}_i^{(m)} \sim q_\phi(\boldsymbol{\theta} \mid \mathbf{x}_i)$ is a complete joint draw. Marginal summaries below use coordinates from these vectors; forward predictions always use the complete vectors. The indicator $\mathbf{1}[A]$ equals 1 when condition A holds and 0 otherwise. The coordinates include the logarithms listed in Supplementary Table S3. Table S1 gives the evaluation populations and draw counts; Supplementary S5 details the model comparisons.

Parameter recovery: R^2. For parameter j, let $\mathcal{E}_j$ contain the examples in which a reaction using that parameter is enabled in the generating recipe. We calculate

$$\widehat{\theta}_{ij} = \operatorname{median}_m \theta_{ij}^{(m)}, \qquad \overline{\theta}_j^\star = \frac{1}{|\mathcal{E}_j|} \sum_{i \in \mathcal{E}_j} \theta_{ij}^\star, \tag{4}$$

$$R_j^2 = 1 - \frac{\sum_{i \in \mathcal{E}_j} (\widehat{\theta}_{ij} - \theta_{ij}^\star)^2}{\sum_{i \in \mathcal{E}_j} (\theta_{ij}^\star - \overline{\theta}_j^\star)^2}. \tag{5}$$

All 512 draws enter each median, including draws that place a truly active step below its threshold. The eligibility set uses the known generating recipe, not inferred activity or reconstruction quality. Shared parameters use the union of the relevant reactions. Surface reaction parameters require adsorption to be enabled as well as their own reaction; capacitance is eligible when its generating $\log_{10}$ value exceeds -2. Historical all-example comparisons instead set $\mathcal{E}_j = \{1, \ldots, N\}$. The calculation uses the listed coordinate: for example, recovery of k^0 is scored on $\log_{10} k^0$. A score of 1 means exact recovery, 0 matches the eligible examples' mean, and a negative score is worse than that constant prediction. The implementation leaves R^2 undefined if the denominator is below 10^{-6}.

The updated recovery calculation uses the original 3,000 examples with saved seeds for new measurement noise and posterior draws; the original noise realizations were not retained. Full-population and eligible-example scores are saved in `active_r2_v1.json`. Historical estimator and noise comparisons retain their original all-example scores.

Active steps: micro-F_1 and exact recovery. For the $K = 15$ optional controls, the known activity label is $z_{ik}^\star = \mathbf{1}[g_{ik}^\star > -2]$, where g_k is its reference rate control. The posterior activity probability p_{ik} is the fraction of draws whose rate control exceeds its reference threshold, as in Equation (29). We predict $\widehat{z}_{ik} = \mathbf{1}[p_{ik} > 0.5]$

and pool counts over all step–example pairs:

$$\mathrm{TP} = \sum_{i,k} z^\star_{ik}\widehat{z}_{ik}, \qquad \mathrm{FP} = \sum_{i,k}(1 - z^\star_{ik})\widehat{z}_{ik}, \qquad \mathrm{FN} = \sum_{i,k} z^\star_{ik}(1 - \widehat{z}_{ik}), \tag{6}$$

$$F_{1,\mathrm{micro}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad A_{\mathrm{exact}} = \frac{1}{N}\sum_i \mathbf{1}[\widehat{\mathbf{z}}_i = \mathbf{z}^\star_i]. \tag{7}$$

Micro-F_1 gives equal weight to each step–example pair, penalizing both missed and falsely included steps. Exact recovery requires all optional controls to be correct in the same example. The always included primary electron transfer and capacitance are excluded from these counts. These scores use one thresholded prediction per example; they are distinct from the probabilities of complete mechanism groups in Equation (50).

Case study probability and parameter scores. For HER, let p_{im} be the inferred probability of family m and $z^\star_{im}$ its one-hot generating label. Accuracy is the fraction of cases in which the family with the largest p_{im} is correct. The reported family and probability scores are

$$F_{1,\mathrm{macro}} = \frac{1}{3}\sum_{m\in\{\mathrm{VH,VT,VHT}\}} \frac{2\mathrm{TP}_m}{2\mathrm{TP}_m + \mathrm{FP}_m + \mathrm{FN}_m}, \qquad B_{\mathrm{HER}} = \frac{1}{N}\sum_i \sum_{m\in\mathcal{M}} (p_{im} - z^\star_{im})^2, \tag{8}$$

where $\mathcal{M}$ also includes Volmer-only for the Brier score. Counts use the most probable family. For the binary thermal event of direct-route dominance, the Brier score is $B_{\mathrm{thermal}} = N^{-1}\sum_i (p_{i,\mathrm{d}} - z^\star_{i,\mathrm{d}})^2$, without the redundant second class. A smaller Brier score means that the assigned probabilities are closer to the known outcomes; the two conventions should not be compared numerically across case studies.

The HER normalized parameter error uses all eligible active-coordinate pairs $\mathcal{A}$ from the generating family:

$$\mathrm{MAE}_{\mathrm{active}} = \frac{1}{|\mathcal{A}|}\sum_{(i,j)\in\mathcal{A}} \frac{|\operatorname{median}_m \theta^{(m)}_{ij} - \theta^\star_{ij}|}{b_j - a_j}, \tag{9}$$

where $[a_j, b_j]$ is the allowed range of that coordinate. It measures absolute parameter error as a fraction of the allowed range. Supplementary S3.2 and S3.3 state the case populations, posterior summaries and evaluation conditions.

Calibration: central interval coverage. Let $\widehat{Q}_{ij}(u)$ be the empirical posterior quantile, using linear interpolation between ordered draws. For assigned probability β, define

$$C_{ij,\beta} = \left[\widehat{Q}_{ij}\left(\frac{1-\beta}{2}\right), \widehat{Q}_{ij}\left(\frac{1+\beta}{2}\right)\right], \tag{10}$$

$$\widehat{c}^{\,Q}_j(\beta) = \frac{1}{N}\sum_i \mathbf{1}[\theta^\star_{ij} \in C_{ij,\beta}], \qquad \bar{c}(\beta) = \frac{1}{D}\sum_j \widehat{c}_j(\beta), \qquad c_{\min}(\beta) = \min_j \widehat{c}_j(\beta). \tag{11}$$

The superscript Q denotes the quantile implementation. It gives the coverage in Table 1a, the coverage columns accompanying the updated R^2 results and the 25-example illustration in Figure 3a. Coverage uses all examples, without the parameter eligibility filter used for R^2. Table 1 reports the mean and minimum across coordinates. Calibration means that empirical coverage follows the assigned probability across repeated simulations, rather than only at $\beta = 0.9$.

The development curves in Figure 3c use the closely related rank implementation. With $r_{ij} = \sum_m \mathbf{1}[\theta^{(m)}_{ij} < \theta^\star_{ij}]$ and $M_i = M$, it is

$$\widehat{c}^{\,R}_j(\beta) = \frac{1}{N}\sum_i \mathbf{1}\left[\frac{1-\beta}{2} \le \frac{r_{ij}}{M} \le \frac{1+\beta}{2}\right]. \tag{12}$$

A group curve averages these coordinate coverages. The gray ±3 percentage-point band is a visual guide, not a confidence interval or a formal test boundary. Finite draws make rank and interpolated-quantile coverage

slightly different. The independent rank audit randomizes exact ties and uses $M = 255$. For assigned probability β, it retains κ_β central integer ranks:

$$\kappa_\beta = \min(M+1, \max(1, \text{round}[\beta(M+1)])), \tag{13}$$

$$\ell_\beta = \left\lfloor \frac{M+1-\kappa_\beta}{2} \right\rfloor, \qquad u_\beta = \ell_\beta + \kappa_\beta - 1, \qquad \beta_M = \frac{\kappa_\beta}{M+1}. \tag{14}$$

Coverage counts ranks from ℓ_β through u_β, inclusive. Under an exact posterior their combined probability is β_M. At nominal $\beta = 0.9$, this is $230/256 = 0.8984$. Supplementary S2.2 gives the full rank test. Central interval coverage and HDR coverage are separate quantities; neither is determined by the region's width.

Remaining uncertainty: HDR width. For a marginal density $q_{ij}(t)$, a β highest density region contains the points with the largest density:

$$\{t : q_{ij}(t) > \lambda_{ij,\beta}\} \subseteq H_{ij,\beta} \subseteq \{t : q_{ij}(t) \geq \lambda_{ij,\beta}\}, \qquad \int_{H_{ij,\beta}} q_{ij}(t)\,\mathrm{d}t = \beta. \tag{15}$$

If the density is flat at the cutoff, a suitable subset of that plateau is included; several equally short choices can exist. In practice, an independent set of posterior draws supplies a Gaussian kernel density estimate $\widehat{q}_{ij}$. Separate cutoff draws give

$$\widehat{\lambda}_{ij,\beta} = \text{quantile}_{1-\beta}\left\{\widehat{q}_{ij}(\theta_{ij}^{(m)}) : m \in \mathcal{B}_{\text{cut}}\right\}, \qquad \widehat{H}_{ij,\beta} = \{t : \widehat{q}_{ij}(t) \geq \widehat{\lambda}_{ij,\beta}\}. \tag{16}$$

Bandwidth choice, cutoff estimation and numerical containment checks use separate draws, as described in Supplementary S2.2. If the estimated region has disjoint intervals $[\ell_{ijr}, u_{ijr}]$, its normalized width and reported parameter summary are

$$W_{ij} = \sum_r (u_{ijr} - \ell_{ijr}), \qquad s_{ij} = 100\frac{W_{ij}}{b_j - a_j}, \qquad S_j = \text{median}_i\, s_{ij}. \tag{17}$$

Here $\beta = 0.9$ and $[a_j, b_j]$ spans the actual generating bounds on the listed coordinate. The denominator includes prior gaps; the numerator excludes gaps between HDR intervals. No clipping to the prior bounds or capping at 100% is applied. Figure 3b shows S_j; each black group summary is the median of S_j over parameters in that group. This percentage describes width, not posterior probability. The region aims to contain 90% of posterior draws regardless of its width. Prior widths use the same denominator. HDRs depend on the coordinate: an HDR calculated in $\log_{10} k$ need not equal the transformed HDR calculated in k.

Signal reconstruction: normalized root mean squared error. Let y_{ivn} be the observed signal at point n of scan v in campaign i, and let $G_{ivn}(\boldsymbol{\theta})$ be its simulated value. A scoring transform z is the identity for signed currents and $\log_{10}$ for positive turnover frequencies. For the valid draws $\mathcal{A}_{iv}$ used in a reconstruction, define

$$\widehat{y}^z_{ivn} = \frac{1}{|\mathcal{A}_{iv}|}\sum_{m\in\mathcal{A}_{iv}} z\left(G_{ivn}(\boldsymbol{\theta}_i^{(m)})\right), \qquad \Delta_{iv} = \max_n z(y_{ivn}) - \min_n z(y_{ivn}). \tag{18}$$

Known conditions and the declared nuisance parameters, such as resistance, are supplied to each simulation. For n_{iv} valid signal points,

$$e_{iv} = \frac{\sqrt{n_{iv}^{-1}\sum_n [\widehat{y}^z_{ivn} - z(y_{ivn})]^2}}{\Delta_{iv}}, \qquad e_i = \text{median}_v\, e_{iv}, \qquad E_{\%} = 100\,\text{median}_i\, e_i. \tag{19}$$

Both blocks of Table 1 use this definition on signed currents. We average simulated curves before scoring; we do not average the errors of individual draws or simulate a vector of parameter medians. The reconstruction denominator uses the observed signal range, rather than the predicted range or peak. Scans with zero signal range have no defined NRMSE; none occurs in the matched full network comparison. Omitting Δ_{iv} gives RMSE in the scoring units. Maximum point error replaces the square root numerator by $\max_n |\widehat{y}^z_{ivn} - z(y_{ivn})|$.

The posterior reconstructions in Table 1 use up to 16 draws that each yield a complete campaign; Supplementary S5 reports simulation failures and the separate check using the optimizer's bounds. All 1,920 unfiltered predictions succeed in the full network comparison. For the E and PCET groups in Figure 4, $\mathcal{A}_{iv}$ contains successful draws from that group for each scan separately. Their 2.1% and 1.9% errors are $100e_i$ for the single example. Mechanism probabilities count all posterior draws before reconstruction, including those whose forward simulation fails. The ClCN overlays and thermal parity figure use pointwise medians as stated in their methods; the HER potential errors are unnormalized RMSE in mV.

For L-BFGS-B, a single candidate vector replaces the posterior average. The full network search selects candidates by the mean squared normalized scan error,

$$\mathcal{J}_i(\boldsymbol{\theta}) = \frac{1}{V_i}\sum_{v=1}^{V_i} e_{iv}(\boldsymbol{\theta})^2, \qquad \widehat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}_i} \mathcal{J}_i(\boldsymbol{\theta}), \tag{20}$$

where $\mathcal{C}_i$ contains all complete candidates evaluated across the starts. The earlier true and random mechanism fits instead select by the sum of unnormalized per-scan current errors,

$$\mathcal{J}_i^{\text{supplied}}(\boldsymbol{\theta}) = \sum_v \frac{1}{n_{iv}} \sum_n \left[G_{ivn}(\boldsymbol{\theta}) - y_{ivn}\right]^2. \tag{21}$$

Each search reports Equation (19) for its selected vector. The optimization objectives and reported median errors are therefore different summaries. Supplementary S5 gives the parameter constraints for each setting.

Inference time. The unit in Table 1a is seconds to generate $M = 1{,}000$ complete joint draws for one input campaign after training:

$$T_{1000} = t_{\text{complete}} - t_{\text{start}}, \tag{22}$$

with timestamps measured in seconds. It is the time for the whole draw batch, not for one parameter or one draw. The historical 91 and 25 ms values describe GPU posterior sampling and exclude training and subsequent forward reconstruction. Their raw timing records were not recovered, so the repeat count, averaging rule, exact GPU and inclusion of encoding cannot be specified from the retained results.

For the full network optimization, the budget per campaign is $B = R\tau = 8 \times 450 = 3{,}600$ s, where R is the number of starts and τ the allowance per start. Recorded worker time is $T_i^{\text{work}} = \sum_{r=1}^{R} \Delta t_{ir}$; total work sums this over campaigns. The median summed time is $\text{median}_i\, T_i^{\text{work}} = 3{,}600$ s (rounded). Table 1b instead reports the minimum and maximum first-start times among cases with complete simulated signals. If $\mathcal{V}$ is that set, the range is $[\min_{i\in\mathcal{V}} \Delta t_{i1}, \max_{i\in\mathcal{V}} \Delta t_{i1}]$. Time-limited starts remain included; starts with failed predictions are counted separately. Parallel execution changes elapsed time, not this sum. Table 1a expresses the historical 91- and 25-ms batch timings as 0.091 and 0.025 s. The CPU optimization and GPU sampling times describe different hardware and workloads, rather than a controlled speed ratio.

Choosing the next measurement: separation score. For the pH search, interpolate each simulated CV onto $J = 128$ equally spaced points along scan progress. For a candidate condition c and mechanism group g, normalize each successful curve separately and calculate

$$x_{gmn}(c) = \frac{I_{gmn}(c)}{\max_n |I_{gmn}(c)| + 10^{-30}}, \tag{23}$$

$$\mu_{gn}(c) = \frac{1}{M_g}\sum_m x_{gmn}(c), \qquad v_{gn}(c) = \frac{1}{M_g}\sum_m [x_{gmn}(c) - \mu_{gn}(c)]^2. \tag{24}$$

Here M_g is the number of successful curves in group g at condition c, which can vary with condition. The variance uses divisor M_g, rather than $M_g - 1$. For two groups A and B,

$$d'(c) = \frac{\sqrt{J^{-1}\sum_n [\mu_{An}(c) - \mu_{Bn}(c)]^2}}{\frac{1}{2}\left[\sqrt{J^{-1}\sum_n v_{An}(c)} + \sqrt{J^{-1}\sum_n v_{Bn}(c)}\right] + 10^{-6}}. \tag{25}$$

The numerator measures the difference between mean shapes; the denominator measures spread within the two groups. Larger values favor conditions where their predictions differ more clearly. This score uses parameter

uncertainty, without added measurement noise or covariance weighting, and is not a probability of successful discrimination. The implementation assigns zero when either group has fewer than two valid curves. For Figure 4, the candidate grid is pH $0, 1, \ldots, 14$ at $\sigma = 0.080433$; pH 0 gives the largest score. The pH 7 and pH 0 scores both use this slower rate. Their reconstruction and design calculations use separate posterior samples.

S2.2 Calibration and recovery checks

The full parameter table reports recovery and uncertainty on their stated evaluation populations (Table S3). Independent simulation based calibration (SBC) tests whether the generating values occupy the expected positions among posterior draws (Figures S2 and S3). These checks detect probability errors that average interval coverage can hide.[13]

Discrete-rank SBC. For each independently simulated campaign and coordinate, SBC uses $L = 255$ posterior draws and the tie-randomized integer rank

$$r = \sum_{\ell=1}^{L} \mathbf{1}[\theta^{(\ell)} < \theta^{\star}] + U\{0, \ldots, T\}, \tag{26}$$

where T is the number of draws tied with the generating value $\theta^{\star}$. Under exact calibration, r is discrete uniform on $\{0, \ldots, 255\}$; physical parameter values are never normalized for this calculation. Posterior draws use independent base samples passed through the deterministic flow, so the MCMC thinning requirement does not apply to this estimator. We control the combined family-wise error rate at $\alpha = 0.05$ without assuming independent parameter coordinates. Half of α is assigned to empirical-CDF diagnostics, using a Dvoretzky–Kiefer–Wolfowitz band with a union bound across all $D = 42$ coordinates. The other half is assigned to 16 equal-null-mass histogram bins, using exact binomial tails with a Bonferroni correction across all $D \times 16$ cells and both tails. The global null is rejected if either diagnostic family crosses its bound. This construction permits arbitrary dependence among coordinates while retaining the standard assumption that simulated SBC campaigns are independent. The checked population is the accepted-simulation law: the locked manifest records 3,000 accepted campaigns and 134 timed-out proposals that were redrawn.

For coordinate j and integer rank k, the empirical and null cumulative probabilities and the simultaneous statistic are

$$\widehat{F}_j(k) = \frac{1}{N} \sum_i \mathbf{1}[r_{ij} \le k], \qquad F_0(k) = \frac{k+1}{L+1}, \qquad \Delta_{\mathrm{CDF}} = \max_{j,k} |\widehat{F}_j(k) - F_0(k)|. \tag{27}$$

The cumulative test rejects if

$$\Delta_{\mathrm{CDF}} > \epsilon_{\mathrm{CDF}}, \qquad \epsilon_{\mathrm{CDF}} = \sqrt{\frac{\log(2D/\alpha_{\mathrm{CDF}})}{2N}}, \qquad \alpha_{\mathrm{CDF}} = 0.025. \tag{28}$$

For histogram bin b, containing 16 of the 256 possible ranks, the count is $H_{jb} = \sum_i \mathbf{1}[r_{ij} \in \mathcal{B}_b]$. Its null distribution is Binomial$(N, 1/16)$. With $B = 16$ bins and $\alpha_{\mathrm{hist}} = 0.025$, each tail is tested at $\alpha_{\mathrm{hist}}/(2DB)$. The two test families together control the probability of any false rejection at $\alpha_{\mathrm{CDF}} + \alpha_{\mathrm{hist}} = 0.05$.

For $N = 3{,}000$, the global rank-CDF bound is 0.0368 and the simultaneous histogram-count interval is 135–244 per bin. The observed maximum CDF deviation is 0.0855, and both diagnostic families reject. Mean central-rank coverage is 0.9095 for an interval whose exact discrete-null mass is $230/256 = 0.8984$; coverage ranges from 0.874 to 0.941 across coordinates. Central coverage is close to its nominal value, although the rank test detects calibration error. Exact ties are negligible. All generating values and posterior draws are retained without clipping. Supplementary S5 separately checks reconstruction using draws within the actual generating bounds.

The near 90% average coverage does not imply agreement at every probability level. In the same independent calibration audit, central 50% intervals contain the true reactant diffusivity D_A in 65% of examples and the second reduction product diffusivity D_C in 37%. SPARK assigns too little probability to the first central range and too much to the second, on average. These errors explain why checking several interval levels and the full rank distribution adds information beyond one coverage number.

Table S2: Generating ranges for the 42 outer sphere parameters. Ranges use dimensionless simulator coordinates, including the indicated logarithms. Sampling is uniform within each selected range. Optional controls use separate negligible and active ranges; the intervals between them are not sampled. The primary electron transfer range depends on the selected reaction family. Failed simulations can change the parameter distribution among accepted examples, which supplies the prior reference in Supplementary S2.

Coordinate(s)	n	Generating range(s)
$\log_{10} k_1^0$	1	$[-4, -1.5]$ for mediated examples; $[-1, 4]$ otherwise
$\log_{10}(k_2^0, k_c, k_Z^0, k_{\text{cat}}, k_f, k_{\text{med}}^0)$	6	$[-5, -2.5]$ negligible; $[-0.5, 2]$ active
$\log_{10}(k_{\text{disp}}, k_{\text{dim}}, k_{\text{rad-sub}}, k_{\text{dc}}, k_{\text{med}}, k_{\text{surf}}^0)$	6	$[-5, -2.5]$ negligible; $[-0.5, 4]$ active
$\log_{10} k_{\text{surfchem}}$	1	$[-5, -2.5]$ negligible; $[-0.5, 3]$ active
$\log_{10} k_{\text{PT}}$	1	$[-5, -2.5]$ negligible; $[1, 6]$ active
$\log_{10} K_{\text{ads},A}$	1	$[-5, -2.5]$ negligible; $[0, 4]$ active
$\log_{10} K_{\text{ads},B}$	1	$[-5, 4]$
$\log_{10} C_{\text{dl}}$	1	$[-3, -2.5]$ or $[-1.7, -0.3]$, each with probability 1/2
$\log_{10} \lambda_1$	1	$[0.5, 3]$
$\log_{10}(\lambda_2, \lambda_Z, \lambda_{\text{med}}, \lambda_{\text{surf}})$	4	$[0.3, 3]$
α_1	1	$[0.4, 0.6]$
$\alpha_2, \alpha_Z, \alpha_{\text{med}}, \alpha_{\text{surf}}$	4	$[0.3, 0.7]$
$\log_{10}(D_A, D_B, D_C, D_{\text{med}})$	4	$[-0.4, 0.4]$
E^0	1	$[-1, 1]$
$\Delta E_2^0, \Delta E_Z^0, \Delta E_H^0$	3	$[-5, 5]$
E_{med}^0	1	$[-3, 3]$ for mediated examples; $[-8, 8]$ otherwise
E_{surf}^0	1	$[-2, 2]$
$\log_{10} K_{\text{eq}}$	1	$[-3, 3]$
$\log_{10} \Gamma_{\text{sat}}$	1	$[-1, 1]$
g_{Frumkin}	1	$[-3, 3]$
pK_a offset	1	$[-6, 6]$ for proton coupled examples; $[-7, 7]$ otherwise

Parameter recovery when reactions are enabled. The eligible counts range from 129 to 3,000. Thermodynamic and diffusion parameters shared by several reactions use the union of their enabled examples. For instance, ΔE_2^0 and D_C are evaluated in 465 examples with either the second electron transfer or disproportionation enabled, whereas the second electron transfer rate uses 269 examples. The full posterior is retained for every estimate. For the following reaction, 147 of 695 enabled examples have a posterior median below the activity threshold; these account for 87% of its squared error. The smaller spread of true rates within the active range also makes these errors large relative to the R^2 baseline. Its active-example R^2 is -6.07, compared with 0.49 over all examples in the same run. The all-example score partly reflects recognition of the widely separated active and inactive ranges. Posterior probabilities and uncertainty are evaluated separately.

Generating prior and posterior region widths. The outer sphere generator first selects a reaction family and then samples its parameters. Fourteen optional kinetic rate constants have separate uniform ranges for negligible and active behavior. The primary electron transfer rate also has two ranges, selected according to whether mediation is present. The adsorption constant $K_{\text{ads},A}$ and capacitance have two ranges; many other parameters have one uniform range. Supplementary Table S2 lists all 42 coordinates. The prior therefore already has separated bands before any signal is observed. Their presence in a posterior does not by itself show that the measurement reveals distinct mechanisms.

The full prior range is the span between the outermost generating bounds, not the nominal range used to scale neural network inputs. The prior reference contains 14,000 accepted simulated examples, separate from the 1,000 cases used for the HDR calculation. This reference retains changes in the generating distribution caused by failed simulations. Within each allowed prior interval, we estimate density with a Gaussian kernel reflected at its boundaries and weight the interval by its observed probability. No prior density is spread across gaps or beyond its bounds. The prior 90% HDR selects the highest density portions until their combined probability is 0.9.

For each of the original 1,000 cases, we draw 8,192 complete parameter vectors from the fixed checkpoint. The saved noisy inputs and their encoded representations are reused exactly. We retain all draws, including values in prior gaps or outside its bounds. For each parameter, 4,096 draws estimate a Gaussian kernel density;

1,024 separate draws choose its bandwidth by predictive log density. The candidates range from 1/64 to 8 times the sample based Silverman bandwidth. None of the selected bandwidths lies at a search boundary.

A further 2,048 draws set the HDR cutoff to the tenth percentile of their estimated density values. The region contains points above this cutoff, with boundaries located by interpolation. This density quantile construction targets probability in the sampled posterior, rather than probability in a smoothed approximation. [14] The remaining 1,024 draws test containment independently. The mean is 89.94%, with parameter means from 89.85% to 90.01%. This checks the numerical construction of each posterior region; the separate tests using known generating values assess the accuracy of the posterior itself.

We check numerical stability on 64 cases across all 42 parameters. Doubling grid resolution changes widths by less than 0.009 percentage points for 95% of these marginals. Halving and doubling the selected bandwidth gives corresponding differences below 0.72 and 0.56 points. Repeating the calculation with 4,096 total draws gives a median absolute difference of 1.19 points and a 95th percentile of 4.87 points. For the prior, five separate fit and check splits give mean containment of 89.61%; halving or doubling bandwidth changes its normalized widths by at most 0.45 points. Flat prior bands can admit several HDRs of the same width, so small differences in their endpoints are not interpreted as chemical alternatives.

The example in Figure 3a is the ninth saved case ($\log_{10} k_c$), chosen to illustrate two separated posterior peaks. It has two HDR intervals, each containing at least 15% of the draws, a gap wider than 0.8 log units and fewer than 3% of draws outside the full prior bounds. Selection does not use the true value or its coverage. The two intervals are $[-4.990, -2.150]$ and $[-0.996, 1.153]$, compared with a full prior range of $[-5, 2]$. Their total width is 71.29% of this range; the median lies at -1.862, in the gap. This example illustrates the distribution's shape, rather than the typical width: the median HDR width for this parameter remains 37.97% across all 1,000 cases.

The lower illustration uses the first 25 saved cases in their original order, with no selection by coverage. Each central interval runs from the fifth to the 95th percentile of all 8,192 draws. It contains the known value in 24 of 25 cases, compared with 908 of 1,000 cases (90.8%) over the full HDR cohort. For the same parameter and cohort, the 90% HDR contains the known value in 94.8% of cases. These are distinct checks: central intervals and HDRs select different regions of the distribution, so their coverage need not agree for an approximate posterior.

Parameter groups in Figure 3. Figure 3b groups all 42 coordinates into seven parameter families. Reorganization energies (λ) and transfer coefficients (α) share one group. Rate constants include the primary electron transfer and all optional reaction rates. The six redox coordinates include reference potentials and potential differences. Equilibrium constants comprise K_{eq} and the pKa offset. Adsorption parameters comprise $K_{\mathrm{ads},A}$, $K_{\mathrm{ads},B}$, Γ_{sat} and the Frumkin interaction parameter. Each coordinate belongs to exactly one group; Table S3 gives every coordinate and its individual HDR width.

The summaries below are medians across parameters within each group. Each posterior parameter summary is the median across 1,000 cases. Both columns use the same full prior range for each parameter. The individual widths are unchanged by regrouping.

Parameter family	n	Prior HDR width (%)	Posterior HDR width (%)
Rate constants	15	35.4	29.5
Reorganization energies and transfer coefficients	10	89.7	93.4
Redox potentials	6	89.7	87.5
Equilibrium constants	2	89.1	90.7
Adsorption parameters	4	89.6	91.6
Diffusion coefficients	4	89.6	87.9
Capacitance	1	59.4	14.5

The calibration curves in Figure 3c retain a separate grouping used in their original calculation. "Optional step parameters" denotes 15 activity controls: 14 rate constants and $K_{\mathrm{ads},A}$. "Other reaction parameters" contains the primary electron transfer rate, five reorganization energies, five transfer coefficients, six redox coordinates, two equilibrium coordinates and the three remaining adsorption parameters. The other curves summarize four diffusion coefficients, capacitance, and all 42 parameters. These calibration groups differ from the parameter families in panel b.

Parameter summaries and checks are retained in `hdr_full_prior_v3.json`. These widths describe marginal uncertainty on the specified coordinates. They do not measure changes in parameter dependence or replace checks of posterior probabilities.

Table S3: All 42 outer sphere coordinates. R^2 compares posterior medians with known values in the n examples where a reaction using the parameter is enabled. Negative values mean that the median performs worse than predicting the average true value within that subset. All 512 draws enter each estimate. Central 90% coverage uses all 3,000 examples. HDR width is the median posterior 90% width as a percentage of the full prior range, using 1,000 cases with 8,192 draws each. Values use the listed coordinates, including logarithms.

Parameter	n	R^2	cov	HDR (%)
Electron transfer ($k^0, \lambda, \alpha, E^0$)				
$\log_{10} k_1^0$	3000	+0.76	0.91	28
$\log_{10} \lambda_1$	3000	+0.13	0.91	90
α_1	3000	+0.04	0.91	91
E^0	3000	+0.42	0.89	64
$\log_{10} k_2^0$	269	-5.57	0.91	34
$\log_{10} \lambda_2$	269	-0.01	0.93	93
α_2	269	-0.01	0.94	94
ΔE_2^0	465	+0.59	0.88	88
$\log_{10} k_Z^0$	270	-9.71	0.91	34
$\log_{10} \lambda_Z$	270	-0.02	0.88	88
α_Z	270	-0.02	0.93	94
ΔE_Z^0	270	+0.63	0.89	89
$\log_{10} k_{\mathrm{med}}^0$	213	-3.10	0.91	33
$\log_{10} \lambda_{\mathrm{med}}$	213	+0.16	0.93	93
α_{med}	213	+0.03	0.93	94
E_{med}^0	213	+0.88	0.88	87
$\log_{10} k_{\mathrm{surf}}^0$	458	-1.15	0.91	26
$\log_{10} \lambda_{\mathrm{surf}}$	458	+0.09	0.93	93
α_{surf}	458	+0.01	0.93	94
E_{surf}^0	458	+0.53	0.89	88
Chemical steps (homogeneous)				
$\log_{10} k_c$ (following)	695	-6.07	0.90	38
$\log_{10} k_{\mathrm{disp}}$ (DISP)	129	-7.47	0.91	26
$\log_{10} k_{\mathrm{cat}}$ (EC′)	390	-8.58	0.91	36
$\log_{10} k_f$ (CE)	302	-20.86	0.89	39
$\log_{10} K_{\mathrm{eq}}$ (CE)	302	+0.44	0.90	90
$\log_{10} k_{\mathrm{dim}}$ (dimer.)	449	-5.78	0.91	33
$\log_{10} k_{\mathrm{rad-sub}}$ (radical)	289	-1.43	0.92	27
$\log_{10} k_{\mathrm{dc}}$ (disp./comp.)	218	-11.77	0.90	28
$\log_{10} k_{\mathrm{med}}$ (cross-ET)	213	-0.79	0.91	25
$\log_{10} k_{\mathrm{surfchem}}$ (surface)	211	-1.76	0.92	29
Diffusion / transport				
$\log_{10} D_A$	3000	+0.72	0.91	40
$\log_{10} D_B$	3000	+0.09	0.90	86
$\log_{10} D_C$	465	+0.02	0.90	91
$\log_{10} D_{\mathrm{med}}$	213	+0.70	0.89	89
Adsorption / surface				
$\log_{10} K_{\mathrm{ads},A}$	529	-3.00	0.91	27
$\log_{10} \Gamma_{\mathrm{sat}}$	529	+0.56	0.90	90
g_{Frumkin}	529	+0.15	0.92	93
$\log_{10} K_{\mathrm{ads},B}$	529	+0.16	0.93	93
PCET (proton)				
$\log_{10} k_{\mathrm{PT}}$	468	-13.95	0.89	26
pK_a offset	468	+0.17	0.92	92
ΔE_H^0	468	+0.50	0.88	87
Double layer (C_{dl})				
$\log_{10} C_{\mathrm{dl}}$	1473	+0.83	0.92	15

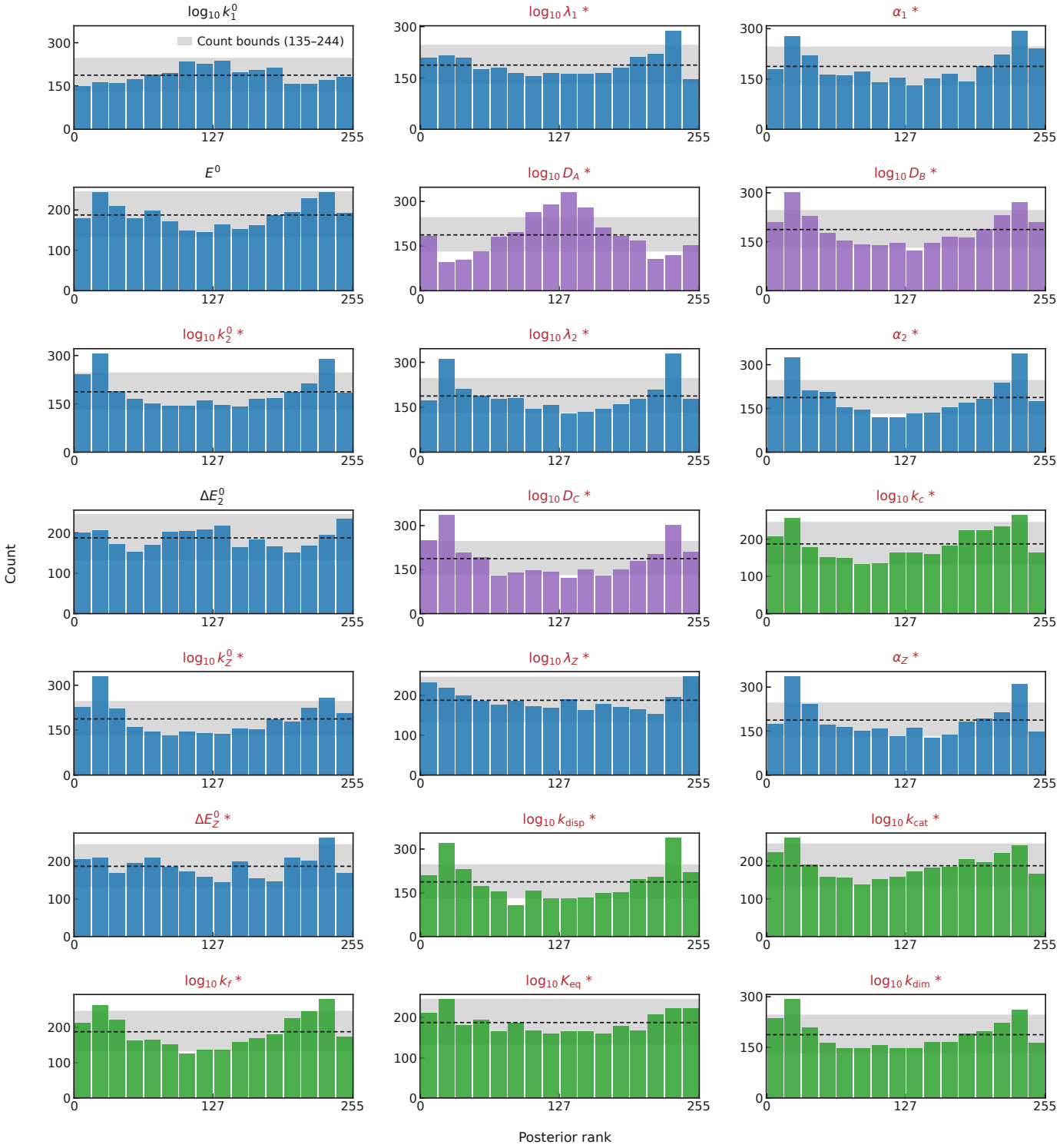


Figure S2: Posterior rank checks for the first 21 parameters. Each panel uses 255 posterior draws and 16 equal probability bins for integer ranks 0–255 on a fixed cohort of 3,000 simulations. The gray count interval, 135–244, is a simultaneous reference for all 42 histograms at $\alpha = 0.025$. Red labels and stars mark histograms that cross this reference. A separate bound on cumulative ranks, combined by a union bound, gives joint error control at $\alpha = 0.05$ without assuming independent parameters. Both tests reject exact calibration; interval coverage is reported separately. Figure S3 shows the remaining parameters.

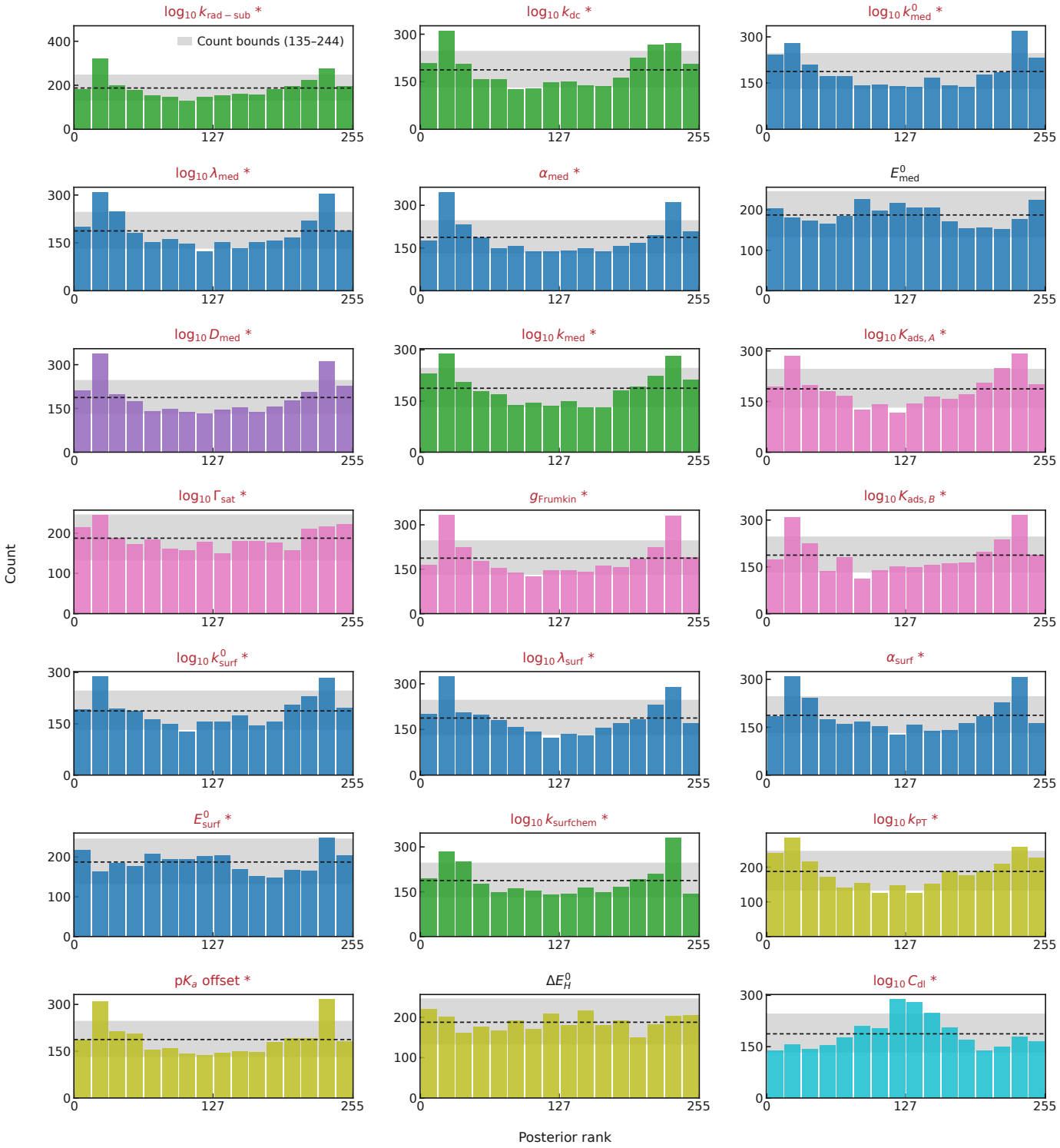


Figure S3: Posterior rank checks for the remaining 21 parameters. The cohort, posterior draws, rank bins and simultaneous count reference are the same as in Figure S2. Together, the two figures report all 42 coordinates. Red labels and stars mark histograms that cross the simultaneous count reference.

S2.3 Comparing pathways and choosing conditions

For optional step k, let $g_k^{(m)}$ be the reference $\log_{10}$ rate or adsorption control in posterior draw m, with activity threshold $\tau_k = -2$. Its activity and posterior probability are

$$z_k^{(m)} = \mathbf{1}\left[g_k^{(m)} > \tau_k\right], \qquad p_k = \frac{1}{M}\sum_{m=1}^{M} z_k^{(m)} \approx P(g_k > \tau_k \mid \mathbf{x}). \tag{29}$$

Linked reactions share the controls listed in Supplementary Table S15. The decision $p_k > 0.5$ selects a single step set only for the detection metrics; it is distinct from the rate threshold. Mechanism groups retain whole draws with the same complete activity pattern. Their probabilities use all draws before forward simulation, as defined in Equation (50).

Each displayed group is checked by simulating its parameter draws and comparing the mean current with the input curves. The recorded analysis labels a reconstruction as having similar error when

$$\mathrm{NRMSE} \le \max(0.05,\ 1.5\,\mathrm{NRMSE}_{\mathrm{best}}),$$

where the best error is the smallest finite group error for that example. The 0.05 value is a fixed tolerance, not a cutoff calibrated from independent residuals. No additional calibrated absolute-error filter was applied in this analysis. This rule compares reconstructed signals; it does not assign a probability that an experimental signal is adequately described by the model. Case-study reconstruction checks and chemical constraints are given separately in Supplementary S3.

Figure 4 uses simulated example 4498, excluded from training, at pH 7 and dimensionless scan rates $\sigma = 0.80433$, 4.20266 and 97.87286. Panel b shows the slowest input scan. The E and PCET errors, 0.021 and 0.019, are medians across scans, calculated from each group's mean simulated current and divided by the input current range. All posterior draws contribute to group probabilities, including those whose subsequent forward calculation fails; only successful curves enter reconstruction means.

The proposed-measurement search evaluates 15 pH values from 0 to 14 at $\sigma = 0.080433$, one tenth of the slowest input scan rate. The pH 7 reference in panel c is therefore a prediction at the slower rate, distinct from the input in panel b. Each successful predicted curve is interpolated onto 128 scan-progress points and divided by its largest absolute current before calculating group means and the separation score (Equations (24) and (25)). The search identifies pH 0 as the largest-score condition; panel d shows this prediction. Reconstruction and design use separate posterior samples. The score describes predicted shape separation under parameter uncertainty, without added measurement noise. The literature cases use the product and coverage comparisons in Supplementary S3 rather than this pH score.

S3 Case study details

These sections describe the measurements, reaction models and inference checks behind Figure 5. The supplementary figures compare reconstructed and measured signals; the parameter and uncertainty checks explain how to interpret the proposed experiments.

S3.1 Co–ClCN: implementation, evaluation and inference

Measurements. The publisher's workbook `Fig. 4.xlsx`, sheet `Fig. 4c`, supplies CVs at 0, 26, 53, 106 and 264 equivalents of $Cl(CH_2)_3CN$.[7] We use the first recorded cycle, with 1,000 current points at 0.01 V s^{-1}. The other two cycles repeat that scan rate. Numerical traces from the source's separate experiments at other scan rates are not included in this archive.

Potentials retain the reported Fc/Fc^+ reference and iR correction, so the forward model uses $R_u = 0$. The source $Co^{II/I}$ potential is -1.275 V. The current values match amperes when checked against the published figure, despite the workbook's microampere header; conversion to microamperes is applied only for display. We use the reported 3 mm disk and catalyst concentrations of 1.0000, 0.9975, 0.9950, 0.9901 and 0.9756 mol m^{-3}, calculated from the titration volumes.

Inference with the general CV model. Before constructing the explicit Co–ClCN network, we apply the same trained 42-parameter flow matching model used in the preceding Results sections. Its weights remain unchanged, and the experimental inputs use the normalization saved from the original training library; no Co–ClCN retraining is performed at this stage. Each titration supplies one measured CV and receives an independent posterior. The base calculation uses a reference diffusion coefficient of 4.66×10^{-10} m^2 s^{-1} to convert the experimental signals to the model coordinates. We pool 1,024 complete draws from each of three sampling seeds, giving $M = 3{,}072$ draws per titration. Every draw enters the probabilities in Figure S4; none is filtered by reconstruction error or assigned a new weight.

All sequences include the primary step $A + e^- \rightleftharpoons B$. EE adds $B + e^- \rightleftharpoons C$. ECE contains $B \to Z$ and $Z + e^- \rightleftharpoons Z_r$. EC′ contains regeneration, $B \to A$. DISP contains $B \to Z$ and solution electron transfer, $Z + B \to A + Z_r$. Using the activity indicators in Equation (29), the probability that a draw contains a sequence with required controls $\mathcal{S}$ is

$$P(\mathcal{S} \text{ present} \mid \mathbf{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \prod_{j \in \mathcal{S}} z_j^{(m)}, \qquad z_j^{(m)} = \mathbf{1}[g_j^{(m)} > -2]. \tag{30}$$

Here g_j is the stored $\log_{10}$ reference rate control. The required controls are k_2^0 for EE, $\{k_c, k_Z^0\}$ for ECE, k_{cat} for EC′ and $\{k_c, k_{\text{dispX}}\}$ for DISP. Other steps may be active in the same draw, so these sequence probabilities overlap and need not sum to 100%. They measure posterior support for active sequences, rather than the fraction of molecules following a route.

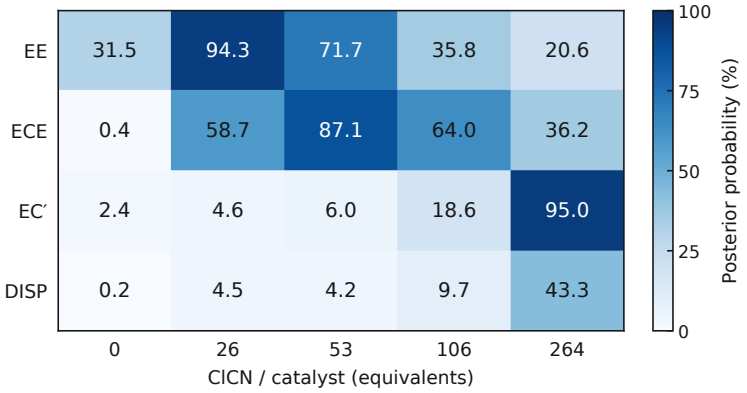


Figure S4: Reaction sequences inferred by the general SPARK model for Co–ClCN. Color and cell labels give the percentage of all 3,072 complete posterior draws containing each sequence under the base diffusion assumption. Each CV is analyzed independently using the same trained 42-parameter model as in the earlier Results sections, without retraining. Sequences may coexist, so probabilities need not sum to 100%. The 264-equivalent prediction has poor CV reconstruction, as discussed in the text.

At 53 equivalents, the largest complete active combination is EE plus ECE and coupling $A + B \to P$, with 18.6% probability. EE plus ECE has 13.7%, while ECE alone has 4.2%. The 87.1% ECE entry therefore combines many possible networks that share this sequence. Halving or doubling the reference diffusion coefficient gives ECE probabilities of 28.5–58.7%, 76.9–88.7% and 59.0–67.0% at 26, 53 and 106 equivalents, respectively. These ranges compare normalization assumptions, not credible intervals.

We also simulate 24 unchanged draws per titration. Under the base diffusion assumption, errors of the mean reconstructed CV are 4.2%, 15.5%, 7.5%, 8.4% and 420.3% of the observed current range at 0, 26, 53, 106 and 264 equivalents. All calculations are finite. At 264 equivalents, the mean is strongly affected by a few poor draws; the median individual-draw error is 25.0%. Across the three diffusion assumptions, the error of the mean remains 120–420%, so the apparent preference for EC′ at this concentration is not assigned a chemical interpretation. These checks assess prediction agreement; they do not establish calibration for this single-CV deployment.

From inferred sequences to an explicit chemical network. The general inference points to chemistry followed by further reduction, but its generic species do not identify the molecular product. We combine

this finding with the source's proposed formation routes to build the eight-step network below. This network supplies a chemical interpretation to test; its eight reactions are not individually discovered by the general estimator.

Sheng et al. identify ClCN as an exception to their simple EC analysis. Direct activation forms R–Co^{III} and Cl^-, whereas radical activation first forms Co^{II}, $R^\bullet$ and Cl^-, followed by radical capture. The source considered radical electrode reduction and acknowledged cobalt–alkyl reduction; the present network examines these together with both solution electron transfer reductions.[7]

The eight reactions in Figure 5a conserve cobalt, the TPP ligand, the alkyl group, chlorine and charge. Steps 5 and 7 reduce the radical and cobalt–alkyl complex at the electrode; steps 6 and 8 reduce them through electron transfer from $[Co^I]^-$. For each solution electron transfer, the forward to reverse rate ratio is $\exp[F(E^0_{\text{acceptor}} - E^0_{\text{Co}})/(RT)]$. The reduced complex is written $[\text{R–Co}]^-$ without assigning the electron to the metal or ligand. Subsequent Co–C cleavage and the lifetimes of the reduced products are outside this model.

Table S4: Explicit Co–ClCN reaction network. Numbers match Figure 5a. The TPP ligand is omitted. Steps 2 and 3 form the radical route to the same neutral cobalt–alkyl product as step 4. Steps 5–8 reduce the free radical or bound product, at the electrode or through dissolved $[Co^I]^-$. These candidate reactions are supplied to the simulator; SPARK infers their parameters.

Step	Reaction
1	$Co^{II} + e^- \rightleftharpoons [Co^I]^-$
2	$[Co^I]^- + RCl \longrightarrow Co^{II} + R^\bullet + Cl^-$
3	$Co^{II} + R^\bullet \longrightarrow R{-}Co^{III}$
4	$[Co^I]^- + RCl \longrightarrow R{-}Co^{III} + Cl^-$
5	$R^\bullet + e^- \rightleftharpoons R^-$
6	$[Co^I]^- + R^\bullet \rightleftharpoons Co^{II} + R^-$
7	$R{-}Co^{III} + e^- \rightleftharpoons [R{-}Co]^-$
8	$[Co^I]^- + R{-}Co^{III} \rightleftharpoons Co^{II} + [R{-}Co]^-$

Reaction and transport calculation. Each solution reaction follows mass action: its rate is its rate constant times the concentrations of its reactants, minus the reverse rate for steps 6 and 8. For species i at distance x from the electrode,

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2} + \sum_{j \in \{2,3,4,6,8\}} \nu_{ij} r_j, \tag{31}$$

where ν_{ij} is the stoichiometric coefficient from Table S4. For each electrode step $j \in \{1, 5, 7\}$, the net reduction flux is

$$q_j = k_j^0 \left[c_{\text{ox},j}(0,t) e^{-\alpha_j \eta_j} - c_{\text{red},j}(0,t) e^{(1-\alpha_j)\eta_j} \right], \qquad \eta_j = \frac{F(E - E_j^0)}{RT}. \tag{32}$$

The species fluxes at the electrode follow these stoichiometries. The current is

$$i(t) = -FA \sum_{j \in \{1,5,7\}} q_j + AC_{\text{dl}} \frac{\mathrm{d}E}{\mathrm{d}t} + i_{\text{offset}} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \tag{33}$$

with reduction negative and A the disk area. Simulated reconstructions omit ϵ_i while retaining charging and offset. Concentrations are in mol m^{-3} internally; rate constants listed in M^{-1} s^{-1} are divided by 1,000 before use with these concentrations.

The initial solution contains only Co^{II} and RCl. The far boundary holds their bulk concentrations fixed. Transport is approximated by one-dimensional planar diffusion; the disk area sets the current scale. We use 96 finite-volume cells with edges at $L(n/96)^2$, where $L = 6\sqrt{D_{\max} t_{\text{scan}}}$. A backward differentiation solver advances the coupled concentrations, with relative tolerance 2×10^{-6} and absolute tolerance 10^{-10}. The recorded potential waveform determines time through cumulative absolute potential change divided by scan rate. The state is carried through the scan reversal. This model and its 17-coordinate estimator are separate from the generic outer sphere simulator and its 42-coordinate estimator.

Joint parameter inference and training. The simulation library contains 20,480 finite CVs: 16,384 for training, 2,048 for validation and 2,048 for testing, generated with separately scrambled Sobol sequences. It reads all 1,000 ordered signed currents, their normalized shape and amplitude, and the known concentrations. A multilayer perceptron with three residual blocks of width 256 supplies a 128-coordinate representation to an autoregressive neural spline flow with eight transforms and eight bins. The competing direct and radical activation rates are represented by their total and ratio during training, then mapped back to the original physical coordinates. All 17 parameters, including diffusion, charging, offset and noise, are inferred jointly. Independent Gaussian noise is added during training at the standard deviation specified by each simulated parameter set.

Training minimizes negative log likelihood with Adam, a batch size of 256 and an initial learning rate of 3×10^{-4}. The schedule reduces the learning rate when validation likelihood stops improving; training allows up to 400 epochs with early stopping. Validation likelihood selects epoch 152 within the chosen training run. Selection among estimator variants also used 40 development cases and the measured CV reconstructions. The checkpoint was then fixed before the independent tests below. Each titration conditions this same estimator independently. The parameter priors are broad search ranges, rather than independently measured chemical limits.

Co–ClCN parameter priors. The 17 coordinates in Table S5 are sampled independently and uniformly within the listed intervals, using a scrambled Sobol sequence with a separate scramble for each split. Thus a uniform prior on a logarithmic coordinate gives a log-uniform prior on the corresponding physical quantity. These intervals define the search space. The transformation to the sum and ratio of the two substrate-activation rates preserves this original joint prior.

Table S5: Priors for the explicit Co–ClCN model. Each listed coordinate has a uniform prior on its interval. Potentials are in V versus Fc/Fc^+. Rate and diffusion coordinates use the units shown inside each logarithm. k_{ex} is a solution electron-transfer prefactor; its relation to the forward and reverse rates is given below.

Quantity	Sampled coordinate	Interval
$Co^{II}/[Co^{I}]^-$ potential	E^0_{Co}	$[-1.300, -1.250]$
Co electron transfer	$\log_{10}[k^0_{Co}/(m\,s^{-1})]$	$[-6, -2]$
Cobalt-species diffusion	$\log_{10}[D_{Co}/(m^2\,s^{-1})]$	$[-10, -8.8]$
Direct alkyl-complex formation	$\log_{10}[k_{direct}/(M^{-1}\,s^{-1})]$	$[-4, 5]$
Radical formation from RCl	$\log_{10}[k_{radical}/(M^{-1}\,s^{-1})]$	$[-4, 5]$
Radical capture by Co^{II}	$\log_{10}[k_{capture}/(M^{-1}\,s^{-1})]$	$[0, 8]$
$R^\bullet/R^-$ potential	$E^0_{radical}$	$[-1.90, -0.95]$
Radical electron transfer	$\log_{10}[k^0_{radical}/(m\,s^{-1})]$	$[-10, -3]$
$R–Co^{III}/[R–Co]^-$ potential	E^0_{bound}	$[-2.00, -1.20]$
Bound-complex electron transfer	$\log_{10}[k^0_{bound}/(m\,s^{-1})]$	$[-10, -3]$
Radical solution reduction	$\log_{10}[k_{ex,radical}/(M^{-1}\,s^{-1})]$	$[-3, 7]$
Bound-complex solution reduction	$\log_{10}[k_{ex,bound}/(M^{-1}\,s^{-1})]$	$[-3, 7]$
Small-species diffusion ratio	$\log_{10}(D_{small}/D_{Co})$	$[-0.4, 0.5]$
Double layer capacitance	C_{dl} (F m^{-2})	$[0, 0.5]$
Current offset	i_{offset} (A)	$[-10^{-7}, 10^{-7}]$
Co transfer coefficient	α_{Co}	$[0.35, 0.65]$
Gaussian current-noise scale	$\log_{10}(\sigma_i/A)$	$[-8, -6.5]$

For each solution reduction, define $K = \exp[F(E^0_{acceptor} - E^0_{Co})/(RT)]$. The sampled prefactor gives $k_f = k_{ex}\min(1, K)$ and $k_r = k_{ex}\min(1, K^{-1})$. The ratio $k_f/k_r = K$ therefore obeys detailed balance, while neither direction exceeds k_{ex}. All four cobalt-containing species share D_{Co}; RCl, $R^\bullet$, R^- and Cl^- share D_{small}. The radical and bound-complex transfer coefficients are fixed at 0.5, and $T = 298.15$ K. Current offset and Gaussian noise are added to the simulated current.

Evaluation on independent simulations. After fixing the checkpoint, we tested 160 fresh synthetic cases, with 32 per concentration, 512 posterior draws and 64 uniformly selected forward draws per case. All 10,240 forward calculations were finite, and all 160 generating parameter sets exactly reproduced their saved clean currents. Errors against clean currents use the full generating CV range, including errors calculated

separately for each scan branch. Clean-truth and noisy-observation errors are reported separately; scores using the observed current range are also archived. Median error of the median CV was 0.68% of the clean current range. Individual errors had median 1.23% and 95th percentile 19.4%; the 95th percentile relative to the known noise standard deviation was 114. Using an interior return maximum above 10% of the maximum absolute current as a peak criterion, median curves recovered 68/70 true peaks and gave false peaks in 2/90 cases without a peak. Mean marginal 90% parameter coverage was 88.3%.

Experimental inference and posterior filtering. We sample 786,432 joint parameter sets per titration, split equally between three fixed seeds. Every draw is finite and remains within the original prior. To examine small additional rates, we define dimensionless scores for steps 5–8:

$$s_5 = k^0_{\mathrm{radical}}\sqrt{t_{\mathrm{scan}}/D_{\mathrm{radical}}}, \quad s_6 = k_{\mathrm{ex,radical}} C_{\mathrm{Co}} t_{\mathrm{scan}},$$
$$s_7 = k^0_{\mathrm{bound}}\sqrt{t_{\mathrm{scan}}/D_{\mathrm{bound}}}, \quad s_8 = k_{\mathrm{ex,bound}} C_{\mathrm{Co}} t_{\mathrm{scan}}.$$

Here $t_{\mathrm{scan}} = \sum_i |E_{i+1} - E_i|/v = 99.9$ s is the recorded scan time, C_{Co} is in mol L^{-1}, electrode rates are in m s^{-1} and exchange rates are in L mol^{-1} s^{-1}. The radical diffusion coefficient is the cobalt diffusion coefficient times the inferred small species ratio; bound species use the cobalt value. Each solution exchange prefactor bounds both its forward and reverse rates.

A draw is retained when all four scores are strictly below 0.01. Its entire joint parameter vector and all eight reactions remain unchanged. No optimization or selection by current error is used. We choose 128 draws uniformly from the full posterior and 128 from the filtered subset before simulation. The same procedure uses cutoffs 0.00316 and 0.0316 for sensitivity checks. At the tighter cutoff, all 105 qualifying draws at 53 equivalents and all 115 at 106 equivalents are used. Overlapping selections share the same forward calculation. All 2,234 unique calculations are finite.

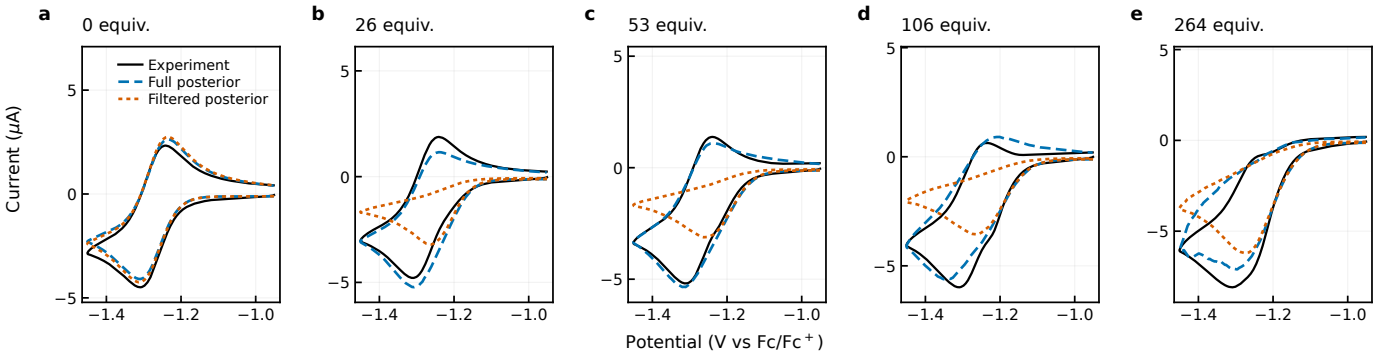


Figure S5: CV reconstruction from the full and filtered SPARK posteriors. Panels a–e show the five ClCN titrations at 10 mV s^{-1}. Black is experiment, blue is the full posterior and orange is the subset with small reference rates for all four additional steps. Each reconstruction is the median current at each potential from 128 uniformly selected parameter draws. Both groups use the full network and unchanged joint parameter vectors. The filtered median loses the return oxidation peak at 26–106 equivalents. Charging and offset are included; no new measurement noise is added.

Table S6: Posterior filtering and CV error. Error is RMS current error divided by the measured full-CV current range. Values describe median curves, which need not arise from one parameter vector. Every fraction uses all 786,432 posterior draws for that titration.

ClCN equivalents	Retained draws	Retained (%)	Full error (%)	Filtered error (%)
0	1,640	0.2085	3.94	3.46
26	385	0.0490	5.20	17.16
53	272	0.0346	2.53	18.26
106	290	0.0369	5.15	19.26
264	773	0.0983	7.31	15.78

The tighter and looser cutoffs preserve the missing return peak at 26–106 equivalents. Across those two cutoffs, filtered median errors are 17.07–17.25%, 18.15–18.44%, 18.95–19.59% and 14.33–16.54% at 26, 53, 106

and 264 equivalents. Individual draws also retain broad errors. At 264 equivalents, their median errors are 15.28% for the full posterior and 15.65% for the filtered subset, despite the larger difference between the two median curves.

The retained fractions describe this region of the learned posterior. They are not probabilities that all four reactions are absent or required: the same joint event already occupies only about 0.44% of the prior. Electrode coefficients can also grow away from their formal potentials, so small reference rates do not ensure negligible reaction throughout the scan. A separate coefficient bound that includes the largest forward and reverse electrode rates across the waveform retains one blank draw and no nonzero draws at the main cutoff. That stricter event also has nearly zero prior mass and is not used to infer chemical impossibility.

Chemical consistency of the parameters. The cobalt diffusion estimates differ across the titration in both full and filtered SPARK (Table S7). Their marginal 90% intervals have no common overlap across all nonzero concentrations. This is a descriptive check across independent posteriors: composition or speciation can change transport, but the present model does not explain those changes. The nonzero full-posterior cobalt potential medians range from −1.266 to −1.260 V, with intervals that include the source value of −1.275 V.

Table S7: Transport estimates across the titration. Cobalt diffusion is in 10^{-10} m^2 s^{-1}. Values are marginal medians from unchanged joint posterior draws.

ClCN equivalents	SPARK	Filtered SPARK
0	5.11	5.44
26	2.57	2.94
53	2.97	3.09
106	4.02	4.10
264	10.37	12.51

The maximum charging current allowed by the prior is 0.0353 μA under these conditions, too small to explain a microampere return peak. Full SPARK accumulates near the offset lower limit, −0.1 μA. These checks assess whether the inferred parameters give a consistent physical description in addition to reconstructing the current.

At 26 and 53 equivalents, SPARK places the bound-product formal potential near −1.281 and −1.283 V, with marginal 90% intervals $[-1.290, -1.270]$ and $[-1.295, -1.270]$ V. The corresponding electrode rate medians are 5.57×10^{-5} and 1.78×10^{-5} m s^{-1}. Radical redox parameters remain much broader. The source does not report formal potentials for these specific products and chose its ordinary scan window to avoid cobalt–alkyl reduction.[7] The inferred accessible redox response therefore needs a product measurement before it can be assigned to R–Co^{III}.

The blank supplies an exact control for inactive chemistry. With no RCl, the model never forms the radical or cobalt–alkyl product, so their kinetic parameters should retain their independent prior marginals. The learned posterior does not do so. For example, 50.63% of blank draws place the bound-product potential between −1.300 and −1.250 V, compared with 6.25% under the uniform prior. At 26 and 53 equivalents the fractions are 96.98% and 94.95%. The blank enrichment exposes inference bias; it cannot be chemical information from the current. This check, together with the reconstruction tails, prevents reading the present sample fractions as calibrated molecular pathway probabilities.

Proposed product measurement. The full and filtered predictions identify later electron transfer as a target for measurement, but the four added steps are not individually established by this comparison. In particular, the blank bias limits the interpretation of the concentrated bound-product potential. We propose generating R–Co^{III} independently and recording its CV under the same solvent, electrolyte and reference conditions, including the predicted response near −1.28 V versus Fc/Fc$^+$. A product response in this range would support assigning part of the return current to the bound-complex redox couple. Its absence would challenge that assignment and require revisiting the inferred parameters or reaction network. This experiment tests the chemical identity behind the reconstructed current; it has not been performed here.

S3.2 Jung Pt HER: implementation, evaluation and coverage predictions

Experimental question and measured data. Jung et al. used modeled hydrogen coverage to interpret Pt HER kinetics and compared alternative descriptions (their Supplementary Figs. S31–S35 and S60). Their Supplementary Fig. S61 shows correlated parameter changes and motivates independent measurements or calculations.[8] The question here is whether pathways that reproduce the same potential measurements imply different surface populations. We compare Volmer–Heyrovsky (VH), Volmer–Tafel (VT) and Volmer–Heyrovsky–Tafel (VHT) descriptions, then use their coverage predictions to specify an additional measurement.

The publisher Pt/0.5 M formate/borate table contains 120 potentials versus NHE at 12 pH values and ten imposed currents. Potential was recorded after 10 s at each imposed current; the simulator treats these endpoints as steady state. Extracted MAT data are verified against the publisher's supplemental ZIP. Jung's equations are retained for source checks. Their rounded Table-S24 parameters do not reproduce the reported SSR; fitting the source equations locally gives a median log-current RMSE across pH curves of 0.085 decades. This current-space source check is separate from the potential-endpoint errors reported below.

Physical model, coordinates and priors. The HER implementation uses the source elementary rate expressions for Volmer and Heyrovsky reactions supplied by hydronium or water, together with Tafel recombination. It has 11 continuous parameters and predicts steady H* coverage and potential at imposed current. The Volmer step supplies H*, Heyrovsky forms H_2 by consuming one H*, and Tafel combines two H*. A shared rate scale and site density are combined into one current-scale parameter. Both the original and continuous parameter representations predict H* coverage, which the experiment did not measure.

Let $k_{\mathrm{V}}^{\max}$ and $k_{\mathrm{H}}^{\max}$ be the larger of the hydronium and water rate constants for the Volmer and Heyrovsky steps. We infer the dimensionless ratios

$$d_{\mathrm{V}} = \log_{10} \frac{k_{\mathrm{V,H_3O}}}{k_{\mathrm{V,H_2O}}}, \qquad d_{\mathrm{H}} = \log_{10} \frac{k_{\mathrm{H,H_3O}}}{k_{\mathrm{H,H_2O}}}, \qquad r_{\mathrm{H}} = \log_{10} \frac{k_{\mathrm{H}}^{\max}}{k_{\mathrm{V}}^{\max}}, \qquad r_{\mathrm{T}} = \log_{10} \frac{k_{\mathrm{T}}}{k_{\mathrm{V}}^{\max}}.$$

The common rate scale and site density Γ cannot be inferred separately from these steady-state measurements. We combine them as $S = 1000F\Gamma k_{\mathrm{V}}^{\max}$ in $\mathrm{mA\,cm^{-2}}$, with Γ in $\mathrm{mol\,cm^{-2}}$, and set $k_{\mathrm{V}}^{\max} = 1$ when evaluating relative rates. The cathodic current is $j = -2S\widetilde{r}_{\mathrm{H_2}}$, where $\widetilde{r}_{\mathrm{H_2}}$ is the hydrogen formation rate in this relative time scale. No separate site-density parameter is inferred. The Frumkin coordinate is $g = f_{\mathrm{Jung}}/2$; the source interaction factors are fixed at 0.5.

The prior proposals are centered on independent VH, VT and VHT fits to all 120 potentials in Jung et al.'s 0.5 M formate/borate Pt dataset. These fits determine the centers in Table S8; they are not Jung et al.'s rounded published parameters. Each family receives one third of the proposals. For 90% of each family's proposals, coordinates are drawn independently from normal distributions with the listed centers and scales, truncated to the listed bounds. The two activity coordinates have an additional restriction: active r_{H} or r_{T} lies above -5.75, and inactive values lie below -6.25. For the remaining 10%, these two coordinates are uniform on $(-6, -5.5)$ when active and $(-6.5, -6]$ when inactive; the other nine coordinates retain the same truncated normal distributions.

For coordinate bounds $[a_j, b_j]$, normalize each fitted center as $u_{mj} = (\mu_{mj} - a_j)/(b_j - a_j)$, where m denotes the three families. The common scale is

$$s_j = (b_j - a_j) \min\left[0.12,\ \max\left(f_j,\ 0.35\sqrt{\frac{1}{3}\sum_m (u_{mj} - \bar{u}_j)^2}\right)\right],$$

with normalized floors $\mathbf{f} = (0.04, 0.05, 0.04, 0.04, 0.05, 0.05, 0.035, 0.04, 0.05, 0.05, 0.035)$ in table order. Exact centers and samples within standardized distance 0.15 of any center are excluded. Failed proposals and failed simulations are recorded without replacement. The accepted training library therefore conditions this proposal distribution on successful generation and need not retain equal family weights. All finite reaction rates remain in the simulator; the threshold $r = -6$ labels activity without removing a step. Because the measurements set the prior centers, the experimental posterior probabilities depend on these fitted priors.

Table S8: Coordinates and proposal distributions for the 11-parameter HER model. Bounds are limits of the allowed coordinates, not uniform prior intervals. The three columns give normal-distribution centers, and s_j gives the common standard deviation before truncation. Centers and scales are rounded to six significant figures. The text specifies the separate sampling rule near the activity thresholds. All coordinates are dimensionless except ΔG_{H} (eV) and the current scale, expressed as $\log_{10}[S/(\mathrm{mA\,cm^{-2}})]$.

Coordinate	Bounds	VH center	VT center	VHT center	s_j
ΔG_{H}	$[-0.60, 0.60]$	-0.0251456	0.0370754	-0.179567	0.048
g	$[-6, 6]$	0.643606	2.43602	0.599102	0.6
$\log_{10} S$	$[-14, 16]$	0.827822	4.45278	2.71753	1.2
d_{V}	$[-20, 20]$	1.52441	-8	7.56538	2.24258
$\alpha_{\mathrm{V,H_3O}}$	$[0.02, 0.98]$	0.575762	0.08	0.561800	0.0806694
$\alpha_{\mathrm{V,H_2O}}$	$[0.02, 0.98]$	0.08	0.08	0.911340	0.1152
r_{H}	$[-20, 20]$	-0.784651	-13	-0.701376	2.02233
d_{H}	$[-20, 20]$	1.91093	0	2.70844	1.6
$\alpha_{\mathrm{H,H_3O}}$	$[0.02, 0.98]$	0.897253	0.5	0.667020	0.0570012
$\alpha_{\mathrm{H,H_2O}}$	$[0.02, 0.98]$	0.532144	0.5	0.0800002	0.0720946
r_{T}	$[-20, 20]$	-13	-5.49988	-5.03581	1.4

HER rate equations and imposed-current calculation. The 11-coordinate model retains the elementary rate expressions from Jung et al.'s source model.[8] Let θ be H* coverage and $v = 1 - \theta$. The source concentration factors are

$$h = 10^{-\mathrm{pH}}, \qquad w = 55.5, \qquad o = 10^{\mathrm{pH}-14}, \qquad a = 0.000793650794,$$

for hydronium, water, hydroxide and H_2, respectively. The solution concentrations are expressed relative to 1 M; in particular, the water factor is 55.5, not a pure-water activity of one. The source H_2 factor a is fixed.

Write G_{H} for the molar free energy obtained from the coordinate ΔG_{H} in eV, with 1 eV per adsorbate equal to $96.48533212\ \mathrm{kJ\,mol^{-1}}$. The equilibrium constants are

$$K_{\mathrm{V}} = e^{-G_{\mathrm{H}}/(RT_{\mathrm{eq}})}, \qquad K_{\mathrm{H}} = e^{G_{\mathrm{H}}/(RT_{\mathrm{eq}})}, \qquad K_{\mathrm{T}} = e^{2G_{\mathrm{H}}/(RT_{\mathrm{eq}})}, \qquad T_{\mathrm{eq}} = 293.15\ \mathrm{K}. \tag{34}$$

The potential-dependent factors retain the source's separate temperature $T_{\mathrm{rate}} = 298.15$ K. Define

$$E_X^0 = \frac{RT_{\mathrm{rate}}}{F} \ln K_X, \qquad z_{X,a} = \frac{F(E - E_X^0)}{RT_{\mathrm{rate}}}, \qquad z_{X,w} = \frac{F(E - E_X^0 + \delta)}{RT_{\mathrm{rate}}},$$

where $X \in \{\mathrm{V}, \mathrm{H}\}$, E is in V versus NHE, and $\delta = 0.82818766347$ V. The indices a and w denote hydronium and water. The Frumkin coordinate is $g = f_{\mathrm{Jung}}/2$, and all source interaction symmetry factors are fixed at $1/2$. With net rates defined as forward minus reverse,

$$\begin{aligned}
\mathcal{R}_{\mathrm{V},a} &= k_{\mathrm{V},a} \left[hv\, e^{-g\theta - \alpha_{\mathrm{V},a} z_{\mathrm{V},a}} - \frac{w\theta}{K_{\mathrm{V}}}\, e^{g\theta + (1-\alpha_{\mathrm{V},a}) z_{\mathrm{V},a}} \right], \\
\mathcal{R}_{\mathrm{V},w} &= k_{\mathrm{V},w} \left[wv\, e^{-g\theta - \alpha_{\mathrm{V},w} z_{\mathrm{V},w}} - \frac{o\theta}{K_{\mathrm{V}}}\, e^{g\theta + (1-\alpha_{\mathrm{V},w}) z_{\mathrm{V},w}} \right], \\
\mathcal{R}_{\mathrm{H},a} &= k_{\mathrm{H},a} e^{g\theta} \left[h\theta\, e^{-\alpha_{\mathrm{H},a} z_{\mathrm{H},a}} - \frac{awv}{K_{\mathrm{H}}}\, e^{(1-\alpha_{\mathrm{H},a}) z_{\mathrm{H},a}} \right], \\
\mathcal{R}_{\mathrm{H},w} &= k_{\mathrm{H},w} e^{g\theta} \left[w\theta\, e^{-\alpha_{\mathrm{H},w} z_{\mathrm{H},w}} - \frac{aov}{K_{\mathrm{H}}}\, e^{(1-\alpha_{\mathrm{H},w}) z_{\mathrm{H},w}} \right], \\
\mathcal{R}_{\mathrm{T}} &= k_{\mathrm{T}} e^{2g\theta} \left[\theta^2 - \frac{av^2}{K_{\mathrm{T}}} \right].
\end{aligned} \tag{35}$$

These expressions preserve the source factors in both directions. In particular, the Heyrovsky and Tafel interaction factors multiply their respective forward and reverse terms equally. The identities $K_{\mathrm{V}} K_{\mathrm{H}} = 1$ and $K_{\mathrm{V}}^2 K_{\mathrm{T}} = 1$ follow from Equation (34); they do not by themselves establish thermodynamic closure of all concentration and interaction factors.

The coordinates in Table S8 set the rate constants, in the relative time scale with $k_{\mathrm{V}}^{\max} = 1$, as

$$k_{\mathrm{V},a} = 10^{\min(d_{\mathrm{V}},0)}, \qquad k_{\mathrm{V},w} = 10^{\min(-d_{\mathrm{V}},0)},$$
$$k_{\mathrm{H},a} = 10^{r_{\mathrm{H}}+\min(d_{\mathrm{H}},0)}, \quad k_{\mathrm{H},w} = 10^{r_{\mathrm{H}}+\min(-d_{\mathrm{H}},0)}, \quad k_{\mathrm{T}} = 10^{r_{\mathrm{T}}}.$$

At each pH and positive imposed cathodic-current magnitude J, we solve

$$\mathcal{R}_{\mathrm{V}} - \mathcal{R}_{\mathrm{H}} - 2\mathcal{R}_{\mathrm{T}} = 0, \qquad 2S(\mathcal{R}_{\mathrm{H}} + \mathcal{R}_{\mathrm{T}}) = J, \tag{36}$$

where $\mathcal{R}_{\mathrm{V}} = \mathcal{R}_{\mathrm{V},a} + \mathcal{R}_{\mathrm{V},w}$ and $\mathcal{R}_{\mathrm{H}} = \mathcal{R}_{\mathrm{H},a} + \mathcal{R}_{\mathrm{H},w}$. The signed current is $j = -J$, and S is in $\mathrm{mA\,cm^{-2}}$. We retain a unique stable coverage root, for which the derivative of the coverage balance with respect to θ is negative, and solve directly for E on a monotone current branch within $[-1.5, 0.25]$ V versus NHE. Failed, ambiguous or out-of-range solutions are recorded. RHE potentials, when used, follow $E_{\mathrm{RHE}} = E_{\mathrm{NHE}} + (\delta/14)\mathrm{pH}$. All finite rates enter both balances; activity thresholds only label the sampled parameter sets.

Specialized estimator and synthetic training. For this campaign, SPARK uses a separately trained HER posterior estimator. Its inputs are the ordered pairs of imposed $\log_{10}$ current and observed NHE potential within each pH curve, together with pH and point masks. Three one-dimensional convolutional layers encode each curve; local center/span normalization is accompanied by the unnormalized centers and spans so that absolute potential and current information is retained. Attention pools the pH curves into a context vector. The encoder preserves the order of current conditions within a curve, while pooling over pH views. The network is initialized separately for each training seed and learns the 11-coordinate HER posterior from HER simulations; the 42-coordinate outer sphere flow matching checkpoint is not used for this case.

Two posterior constructions share this encoder and the same 11 coordinates. The single-flow model uses one bounded rational-quadratic neural-spline flow (RQ-NSF), with no family classifier or family label as input. For each joint draw, $r_{\mathrm{H}} > -6$ and $r_{\mathrm{T}} > -6$ determine whether Heyrovsky and Tafel are active. The four combinations give VH, VT, VHT and Volmer-only probabilities by counting joint parameter draws; marginal gate probabilities are not multiplied. The comparator uses a three-family classifier and one full 11-coordinate RQ-NSF per family,

$$q_\phi(m, \boldsymbol{\theta} \mid \mathbf{x}) = q_\phi(m \mid \mathbf{x}) q_{\phi,m}(\boldsymbol{\theta} \mid \mathbf{x}), \qquad m \in \{\mathrm{VH}, \mathrm{VT}, \mathrm{VHT}\}.$$

Its reported family probabilities are classifier outputs. Thresholding its parameter draws is a separate diagnostic because the full flows retain every coordinate, including those below the activity threshold. Neither construction receives the true family as an encoder input. Both use encoder and context widths of 128, four attention heads, and flows with ten coupling layers, hidden width 192 and eight spline bins.

The pilot proposes 24,000 training and 3,000 validation campaigns on the measured pH/current grid, with independent seeds for parameter sampling and measurement noise. Each proposal supplies a known parameter vector; the simulator calculates a fresh set of potentials and adds independent Gaussian noise with a 5-mV standard deviation. Measured potentials are used to set the proposal centers but are not copied into these synthetic signals. Using the same observations to set the prior centers and condition the posterior makes this an empirical-Bayes analysis. Complete imposed-current solutions are obtained for 10,966 training and 1,397 validation campaigns, or 45.7% and 46.6% of proposals. The remaining 13,034 and 1,603 proposals have incomplete current inversion and are retained in the failure records without replacement. The accepted validation set contains 538 VH, 661 VT and 198 VHT campaigns; 204 accepted cases come from the proposals near the activity thresholds. Thus the evaluated distribution includes the effects of both the fitted priors and unequal solver acceptance.

All three seeds (0, 1, 2) are trained for 80 epochs with AdamW, batch size 128, initial learning rate 5×10^{-4}, cosine learning-rate decay, weight decay 10^{-6} and gradient clipping at 20. The single-flow loss combines the negative log likelihood of all 11 generating coordinates with a binary cross-entropy term for gates read from posterior samples,

$$\mathcal{L}_{\mathrm{single}} = -\left\langle \log q_\phi(\boldsymbol{\theta} \mid \mathbf{x}) \right\rangle + 8\ \mathrm{BCE}(\widehat{\mathbf{p}}, \mathbf{y}), \qquad \widehat{p}_X = \frac{1}{128} \sum_{s=1}^{128} \mathrm{sigmoid}\left(\frac{r_X^{(s)} + 6}{0.15} \right), \quad y_X = \mathbf{1}[r_X^{\mathrm{true}} > -6], \tag{37}$$

where $X \in \{\mathrm{H}, \mathrm{T}\}$ and BCE is averaged over cases and gates. The sampled gate term uses no warmup and a one-epoch ramp to weight 8. It provides gate supervision through the parameter samples without a separate classifier; its use does not establish calibrated gate probabilities. The comparator minimizes the full conditional negative log likelihood plus five times the family classification cross-entropy. No active-coordinate mask is applied to either likelihood.

The single-flow checkpoint minimizes validation negative log likelihood subject to macro-$F_1 \geq 0.65$ and multiclass Brier score ≤ 0.25, evaluated with 128 draws on a fixed prefix of at most 512 validation cases. Its likelihood uses the full accepted validation set. The comparator selects the checkpoint with the lowest weighted validation objective among five-epoch checkpoints. Neither checkpoint selection uses an untouched test set; the validation scores below therefore describe model development.

Synthetic evaluation and its limits. Both frozen estimators are evaluated on the same 1,397 accepted validation campaigns using 512 posterior draws per case and training seed. The generating parameters supply the family labels and the values used to check parameter intervals. Table S9 reports equal means over the three seeds. The paired difference in accuracy, separate-family minus single-flow, is 0.010 (case-bootstrap 95% interval 0.0048–0.0160). This bootstrap resamples case IDs while retaining both models and all seeds for each case; it measures variation across cases conditional on the three fitted seeds.

Table S9: HER development-set evaluation. Scores use the same 1,397 accepted synthetic validation campaigns and are averaged over three training seeds. This set also selected checkpoints. Family accuracy uses the most probable family; macro-F_1 averages the three generating families. The Brier score is the mean sum of squared probability errors over VH, VT, VHT and Volmer-only, with the comparator assigning zero to Volmer-only. Parameter absolute error uses coordinate-wise posterior medians, divides each error by its allowed coordinate range, and averages over coordinates active in the generating family. Interval coverage is the fraction of those generating coordinate values inside equal-tail 90% posterior intervals.

Metric	Single flow	Separate-family model
Family accuracy	0.945	0.956
Macro-F_1	0.921	0.936
Multiclass Brier score	0.0820	0.0677
Mean normalized active-coordinate absolute error	0.0422	0.0397
Central 90% active-coordinate interval coverage	0.875	0.880

The interval check pools active parameter coordinates and does not constitute a rank-based simulation calibration test or a calibration test of experimental mechanism probabilities. More directly, a separate matched intervention audit asks whether the inferred family changes when only the two activity coordinates change. Each VH/VT/VHT triplet has identical values for the other nine parameters, conditions, masks and noise. Each triplet is attempted once, without replacing failed solutions. Seven of 24 proposed triplets yield valid solutions for all three members (21 members), all near the activity thresholds. None of the 12 proposals at the source rate scale or four proposals centered on VHT yields a complete triplet. The fixed checkpoints are evaluated in the recorded order with 512 posterior draws per member. Single-flow accuracies across seeds are 0.333, 0.333 and 0.381; the separate-family model gives 0.333 for every seed. Neither ensemble recovers all three members of any triplet. The models therefore do not reliably distinguish these gate changes in the matched cases that the solver accepts. Their high accuracy on the fitted-prior validation library does not establish reliable discrimination after these parameter correlations are removed. The small accepted sample also does not establish that the mechanisms are indistinguishable under all conditions.

Inference from the measured campaign. Each frozen checkpoint receives the same 120 measured potentials and generates 4,096 joint posterior draws. For the single flow, the VH probability ranges from 0.669 to 0.895 across seeds, with mean 0.817. Mean probabilities are 0.182 for VHT, 0.00049 for VT and 0.000081 for Volmer-only. The separate-family classifier gives means of 0.753, 0.245 and 0.0019 for VH, VHT and VT. The main inset uses the single-flow probabilities averaged equally over the three seeds and omits the small Volmer-only mass. These probabilities summarize the learned posterior under the fitted priors and activity thresholds. They are not calculated from the deterministic fit errors or established as calibrated probabilities for the measured mechanism.

For reconstruction, 64 parameter draws per model and seed are passed through the simulator without added measurement noise. The mean fractions yielding complete campaigns are 79.7% for the single flow and 78.1% for the separate-family model. The mean potential curve over complete draws has RMSE 15.0 and 13.0 mV, respectively, averaged over seeds. These reconstruction summaries condition on successful simulation; the family probabilities above count the posterior before this filtering. The accompanying potential intervals describe parameter-induced noiseless curve spread and do not include measurement noise. Because this single observed campaign has no known mechanism label and also supplies the prior centers, its reconstruction and posterior masses do not provide an independent mechanism accuracy score. Under these assumptions, the inference favors a pathway containing Heyrovsky while retaining a possible Tafel contribution.

Coverage predictions and the proposed experiment. Independent point fits of VHT and VH reproduce the 120 measured potentials with RMSE 10.7 and 11.6 mV; VT gives 25.5 mV (Figure S6). These fits also set the prior centers, but the coverage curves in Figure 5b each come from one fitted parameter vector. They are not posterior mean coverage curves or uncertainty intervals. Coverage is solved directly at the recorded 1.2732395 mA cm^{-2} current, without interpolation. The predicted H* coverage ranges across pH are 0.800–1.000 for VHT, 0.334–0.896 for VH and a constant 0.607 to the displayed precision for VT. The plotted current is representative and was not selected to maximize separation. In the VT limit, the fixed imposed current fixes Tafel turnover and therefore coverage through the pH-independent Tafel law. VT coverage remains constant with pH at each current to the displayed precision, but rises from 0.491 at 0.471 mA cm^{-2} to 0.988 at 21.65 mA cm^{-2}.

The rounded Table-S24 parameters predict coverages of 0.161–0.730, with potential RMSE 118.9 mV in this network. That curve documents the printed parameters and is not an equally fitted fourth model; the full precision fit from Jung et al. was not available. The present comparison therefore extends the source's recognized parameter ambiguity into a concrete prediction: measure hydrogen coverage across pH at the same imposed current, particularly where the fitted VH and VHT curves separate. Such a measurement would test the surface populations behind the two similar potential fits. The proposed coverage measurement has not been performed here, and separation of point-fit predictions alone does not assign a mechanism.

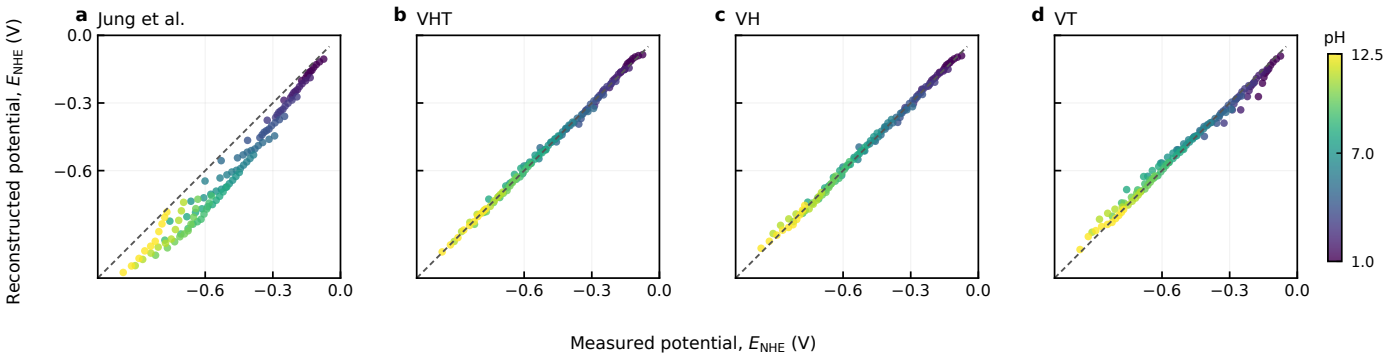


Figure S6: Potential-endpoint reconstruction of the Jung Pt campaign. Predicted versus measured NHE-referenced 10-s potentials for all 120 pH–current conditions. The dashed line marks parity and color denotes pH. VHT, VH and VT use separate point fits, with RMSE 10.7, 11.6 and 25.5 mV. The Jung et al. panel evaluates the rounded values printed in Table S24 and has RMSE 118.9 mV under the source-rate evaluation shown here; the full precision fit from Jung et al. was not available for this comparison.

S3.3 Allian CO oxidation: routes and coverage predictions

Allian et al.'s rates and spectra supported both direct and CO-assisted oxygen activation, whereas their DFT calculations and apparent activation energies favored the assisted route.[9] Direct activation splits adsorbed O_2; assisted activation reacts gas phase O_2 with adsorbed CO, producing CO_2 and adsorbed O. Both routes remove adsorbed O through reaction with CO.

This case asks whether the measured rates and energetic evidence distinguish the two oxygen-activation routes, and which additional observable could test the remaining alternatives. The implementation and

synthetic evaluation below precede inference from the literature measurements. The complete parameter prior, evidence likelihood and training protocol are given in Supplementary S4.

Implementation of the specialized SPARK model. The thermal case uses a separately trained, 16-coordinate rational quadratic neural spline flow (RQ-NSF). Its forward model resolves CO, molecular oxygen, atomic oxygen and vacant sites without imposing CO saturation. The coordinates comprise five reference rate constants and their activation energies, two adsorption equilibrium constants and their enthalpies, a site-normalization factor and a log-rate noise scale (Table S12). Thus the application uses the SPARK strategy of conditioning a joint parameter posterior on simulated signals, with a thermal simulator and posterior estimator specific to this case. It does not apply the generic 42-coordinate outer sphere flow-matching estimator to thermal data.

Let c, m, o and v denote the fractions of sites occupied by CO, O_2, O and vacancies, respectively. Pressures are in bar. The net adsorption fluxes and reaction fluxes are

$$J_{\mathrm{CO}} = k_{\mathrm{CO,ads}} p_{\mathrm{CO}} v - k_{\mathrm{CO,des}} c, \qquad J_{\mathrm{O_2}} = k_{\mathrm{O_2,ads}} p_{\mathrm{O_2}} v - k_{\mathrm{O_2,des}} m, \tag{38}$$

$$r_{\mathrm{d}} = k_{\mathrm{d}} m v, \qquad r_{\mathrm{a}} = k_{\mathrm{a}} p_{\mathrm{O_2}} c v, \qquad r_{\mathrm{LH}} = k_{\mathrm{LH}} o c. \tag{39}$$

The balances are

$$\dot{c} = J_{\mathrm{CO}} - r_{\mathrm{a}} - r_{\mathrm{LH}}, \qquad \dot{m} = J_{\mathrm{O_2}} - r_{\mathrm{d}}, \tag{40}$$

$$\dot{o} = 2r_{\mathrm{d}} + r_{\mathrm{a}} - r_{\mathrm{LH}}, \qquad \dot{v} = -J_{\mathrm{CO}} - J_{\mathrm{O_2}} - r_{\mathrm{d}} + 2r_{\mathrm{LH}}, \tag{41}$$

with $c + m + o + v = 1$. The predicted TOF is $s(r_{\mathrm{a}} + r_{\mathrm{LH}})$, where s converts the model site count to the reported normalization. We check the stability of each steady state. If several stable solutions exist, dynamics starting from a surface predosed with CO select the state. Molecular O_2 occupies one site, and the model approximates neighboring occupations by mean coverages.

Each rate constant follows $k(T) = k_{\mathrm{ref}} \exp[-E(1/T - 1/T_{\mathrm{ref}})/k_{\mathrm{B}}]$. Adsorption equilibrium constants follow the same relation with ΔH in place of E. For both adsorbates, $k_{\mathrm{des}}(T) = k_{\mathrm{ads}}(T)/K(T)$ and $E_{\mathrm{des}} = E_{\mathrm{ads}} - \Delta H$, enforcing detailed balance for adsorption. O_2 splitting and CO_2 formation are irreversible, with CO_2 removed as a gas product. The model therefore imposes adsorption equilibrium relations without enforcing thermodynamic closure across the full network. It has no explicit OOCO intermediate.

Training and route summaries. The finite simulation library contains 65,536 training campaigns, 4,096 validation campaigns and a separate 4,096-campaign test partition. Training generates fresh rate noise and optional desorption and DFT evidence, using the three evidence modes and uncertainty ranges in Supplementary S4. A 62-input encoder retains every measured log rate through an invertible within-series contrast representation, together with masks, catalyst design and energetic evidence. Its 128-coordinate context conditions an eight-layer RQ-NSF over all 16 parameters. Negative log density supplies the training loss; validation loss selects the checkpoint. Each catalyst's measurements subsequently condition a separate posterior without retraining. The held-out results below use specified subsets of the test partition, rather than all 4,096 test campaigns.

For a parameter draw at temperature T and pressures $\mathbf{p}$, define

$$f_{\mathrm{a}}(\boldsymbol{\theta}; T, \mathbf{p}) = \frac{r_{\mathrm{a}}}{r_{\mathrm{a}} + r_{\mathrm{d}}}, \qquad \widehat{P}_{\mathrm{d}}(T, \mathbf{p}) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\left[f_{\mathrm{a}}(\boldsymbol{\theta}^{(j)}; T, \mathbf{p}) < \tfrac{1}{2} \right], \tag{42}$$

where M is the number of posterior draws with resolved fluxes. The first quantity is the fraction of O_2 activation events using the assisted route; direct splitting of one O_2 counts as one event even though it forms two O atoms. The second estimates the posterior probability of direct dominance at the stated condition. Assisted dominance uses $f_{\mathrm{a}} > 1/2$. These are probabilities of carrying the larger activation flux, rather than probabilities that a route exists or measures of rate control. Both routes remain in each parameter draw. Ties and unresolved assignments are reported separately.

Evaluation on synthetic campaigns. The focused route check uses 64 held-out parameter truths for each catalyst design, with 256 unfiltered posterior draws per truth and evidence mode. The same 128 truths and noisy rate observations are used across the three evidence modes; the 32 truths used in the earlier development audit are excluded. Ground truth is the dominant route obtained by simulating the generating parameters. Majority assignment thresholds the inferred dominance probability at one half. The Brier score is the mean squared error of that probability against the binary generating event. Route-interval coverage instead asks whether a central 90% interval for the continuous f_a contains its generating value.

The reference conditions are 423 K, 0.1 kPa CO and 3.7559 kPa O_2 for D087, and 443 K, 0.1 kPa CO and 5.0179 kPa O_2 for D009. All truth and posterior route evaluations are finite, with no exact ties or excluded draws. Table S10 shows useful route discrimination under these conditions. Prior-only Brier scores are 0.2298 for D087 and 0.2497 for D009, using 32,768 unconditioned prior simulations per design at the same reference conditions. Their majority-assignment accuracies are 42/64 and 34/64, respectively. The posterior scores therefore improve on these prior baselines, although adding DFT does not improve the scores relative to desorption evidence alone in this cohort.

Table S10: Synthetic evaluation of the specialized CO oxidation posterior. Each row uses the same 64 held-out truths for its catalyst design and 256 posterior draws per truth. Correct assignments compare the inferred majority route with the route of the generating parameter vector. Brier scores measure probability error, with lower values better. Coverage refers to central 90% intervals for the continuous assisted activation fraction. These test-subset results evaluate inference under the simulator and its sampled evidence model; they do not supply known mechanism labels for the literature measurements.

Design	Evidence	Correct	Brier	90% coverage
D087	Rates	58/64	0.0814	96.9%
D087	Rates + desorption	59/64	0.0816	95.3%
D087	Rates + desorption + DFT	59/64	0.0833	93.8%
D009	Rates	51/64	0.1399	81.2%
D009	Rates + desorption	53/64	0.1193	89.1%
D009	Rates + desorption + DFT	49/64	0.1299	87.5%

The interval widths remain broad: with DFT, the mean 90% widths for f_a are 0.711 and 0.685 for D087 and D009. The audit varies DFT errors and their correlation over the training evidence model; only four truths per catalyst have all three DFT standard deviations at or below 0.10 eV. It therefore does not establish local calibration at the literature energy values or an exact DFT uncertainty at which the preferred route changes.

Full-curve replay uses 16 truths per catalyst and evidence mode from this final cohort, with eight preselected posterior draws per truth. All 96 clean generating-curve replays match the stored library within 10^{-8} decades, and all 768 posterior curves are finite. Joint posterior accuracy nevertheless remains limited. With DFT evidence, the median individual-draw log-rate RMSE divided by the known generating noise is 4.02 for D087 and 2.52 for D009; the corresponding 95th percentiles are 81.6 and 13.1. An exact conditional-noise check under the stated Gaussian likelihood also finds noise values too small for their paired kinetic residuals. These errors limit the experimental route probabilities and coverage predictions below. Finite simulations and useful majority-route discrimination do not establish calibration of the full joint posterior.

Inference from the literature measurements. The main coverage predictions use D087, the catalyst with exposed-Pt fraction 0.87. Its inputs are a CO desorption energy of 63 kJ mol^{-1} with an assumed Gaussian standard deviation of 6 kJ mol^{-1}, and the reported desorption constant 0.0028 ± 0.0006 s^{-1} at 303 K.[9] For conditioning, we treat the quoted 0.0006 s^{-1} spread as a Gaussian standard deviation and convert it to the log-rate scale using the delta method. DFT inputs are -79, 35 and 15 kJ mol^{-1} for CO adsorption, O_2 uptake and CO-assisted activation, with independent standard deviations of 0.10 eV. These uncertainties are analysis assumptions, and electronic energies approximate effective kinetic enthalpies. The source gives conflicting O_2 uptake barriers: 35 kJ mol^{-1} in the Figure 9 discussion and 48 kJ mol^{-1} in Table 3. We use 35 kJ mol^{-1} for the main predictions and 48 kJ mol^{-1} in separate sensitivity runs. No numerical CO coverage is supplied. The SI parity plot also includes an independently inferred D009 campaign, with exposed-Pt fraction 0.09. It uses CO desorption energy 54 kJ mol^{-1} with assumed standard deviation 6 kJ mol^{-1} and no desorption-rate

constraint at 303 K. Its DFT values and uncertainties match D087. Both campaigns include the same seven measurements from the aggregate temperature series.

The observed campaigns have no known route labels. Their posterior route weights are inference results, while replaying their conditioning measurements checks reconstruction. Across 50 evidence settings, all 6,400 preselected source curves and 51,200 source route evaluations are finite, with no rejection based on reconstruction error.

Supplementary Fig. S7 compares reconstructions with all 20 D087 and 22 D009 conditioning measurements. Each catalyst independently conditions its own joint posterior. All 128 draws per campaign are selected before simulation and retained regardless of route or reconstruction error. Points show median $\log_{10}$ TOF without measurement noise; bars show central 90% intervals. Median-curve RMSE is 0.06105 for D087 and 0.05979 for D009 in $\log_{10}$ TOF. Median individual-draw RMSE is 0.07236 and 0.08400, respectively. Typical residuals exceed the inferred measurement noise. The digitized data use saved marker coordinates that still require checking against the original figure crops.

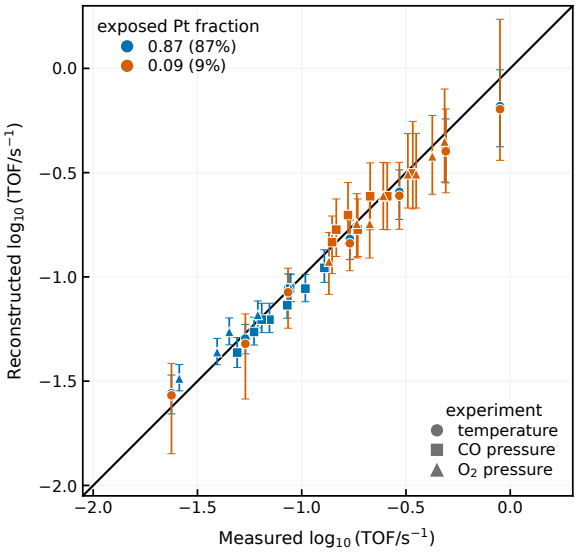


Figure S7: SPARK reconstruction of Allian CO oxidation rates. Measured versus reconstructed $\log_{10}$ TOF for exposed-Pt fractions 0.87 (blue, 20 conditions) and 0.09 (orange, 22 conditions). Circles, squares and triangles mark temperature, CO-pressure and O_2-pressure series. Both campaigns include the same seven temperature measurements. The solid line marks parity. Points and bars show pointwise medians and central 90% intervals from 128 draws selected before simulation for each catalyst posterior. Inference uses the 16-parameter model, DFT inputs and desorption constraints stated in the text. All draws and both oxygen activation routes are retained, without added measurement noise. The intervals describe uncertainty in the model rates, not experimental error bars.

At 423 K, 0.1 kPa CO and 3.7559 kPa O_2, the D087 posterior with desorption and DFT evidence assigns direct dominance to 199 of 1,024 draws (19.4%) and assisted dominance to 825 (80.6%). This preference agrees in direction with Allian et al.'s proposed assisted route, while retaining a competing direct route under the specified priors and evidence errors. The percentages describe this approximate posterior at the stated condition; the synthetic checks do not validate them as exact probabilities for the experimental catalyst.

Sensitivity to kinetic assumptions. Kinetic assumptions also affect these route weights. With the same likelihood and DFT inputs, a Gaussian preference for similar assisted/uptake Arrhenius prefactors lowers direct dominance from 19.4% to 10.5% or 5.7%, when the standard deviation of their $\log_{10}$ ratio is two decades or one decade, respectively. These are optional prior assumptions, not measured constraints or assumptions made by Allian. Importance weighting of the saved posterior gives effective sample sizes of 559 and 303. The route probabilities therefore depend on kinetic priors as well as DFT accuracy.

Coverage predictions at the infrared conditions. We simulate all 1,024 D087 posterior draws at 358 K and 10 kPa O_2, changing only the simulation conditions. The CO-pressure grid contains 55 values from 10^{-5}

to 1 kPa, including 0.001, 0.004, 0.01, 0.08, 0.1 and 0.15 kPa. Since 358 K lies below the 403–463 K rate series used for inference, this prediction extrapolates the inferred temperature dependence.

At the coverage reference of 358 K, 0.1 kPa CO and 10 kPa O_2, the direct route dominates in 151 draws (14.7%) and the assisted route in 873 (85.3%). Every draw has a resolved assignment. These groups differ from those at the 423-K route-check reference because the conditions differ. Membership stays fixed across all displayed temperatures and CO pressures, so each curve follows the same parameter draws. Both routes remain in every simulation.

The main curves show pointwise median coverage without added measurement noise. Each median combines many parameter vectors and need not correspond to a single one. Table S11 reports the corresponding uncertainty at the pressure highlighted in the main text.

Allian et al.'s infrared data provide qualitative evidence for CO coverage.[9] At 358 K and 10 kPa O_2, nearly constant CO bands from 0.08 to 0.15 kPa support a CO-covered surface, without giving absolute coverage or numerical uncertainty. Below 0.08 kPa, the bands disappear as CO conversion becomes complete. This does not define zero coverage at a known nonzero CO pressure at the catalyst. Our simulator holds gas pressures fixed and omits conversion, contact time, heat transfer and the infrared response. The low-pressure curves therefore predict coverage for a controlled pressure measurement; they do not simulate the reported reactor transition.

Temperature dependence of the coverage predictions. The same 1,024 draws and fixed 151/873 groups are simulated at 358, 383, 403, 423, 443 and 463 K, using 55 CO pressures and 10 kPa O_2. The DFT inputs and posterior remain unchanged. Predictions at 358 and 383 K extrapolate below the 403–463 K rate series. Negative CO adsorption enthalpy makes K_{CO} decrease as temperature rises. The oxygen reactions have their own inferred barriers, so the coverage response depends on both CO binding and reaction rates.

Table S11 gives the group medians, their uncertainty and the median gap at a fixed pressure. At 0.01 kPa CO, the gap rises across all six temperatures. The largest gap on each pressure grid ranges from 0.522 to 0.627 without rising steadily, while its location moves from 0.000794 to 0.01 kPa. Heating therefore shifts the region of strongest separation toward higher CO pressure. This comparison uses predicted coverages, without accounting for measurement precision or reactor behavior.

Table S11: CO coverage and uncertainty across temperature. Group medians and central 90% intervals at 0.01 kPa CO and 10 kPa O_2. The gap is the CO-assisted median minus the direct median, in percentage points. The same 151 direct-group and 873 CO-assisted-group draws are retained at every listed condition; groups are assigned at 358 K, 0.1 kPa CO and 10 kPa O_2. Intervals describe parameter uncertainty without added measurement noise.

T (K)	Direct: median [5%, 95%]	CO-assisted: median [5%, 95%]	Gap (points)
358	0.960 [0.0117, 0.9996]	0.994 [0.7135, 0.9999]	3.4
383	0.888 [1.32×10^{-4}, 0.9992]	0.986 [0.6776, 0.9998]	9.8
403	0.758 [6.43×10^{-6}, 0.9985]	0.970 [0.5639, 0.9995]	21.2
423	0.564 [3.67×10^{-7}, 0.9968]	0.937 [0.3047, 0.9989]	37.3
443	0.340 [7.14×10^{-9}, 0.9931]	0.875 [0.0890, 0.9975]	53.5
463	0.145 [1.92×10^{-12}, 0.9874]	0.772 [0.0015, 0.9948]	62.7

Central 90% intervals overlap at every evaluated condition across all six temperatures. Main Figure 5c shows the 358- and 463-K medians as solid and dashed curves, without added measurement noise.

Of 337,920 steady-state calculations, 337,915 yield valid solutions. Five failures affect one direct-group draw, away from the 0.01-kPa comparisons and the largest-gap conditions. Only those five summaries use 150 rather than 151 direct-group draws. No draw is excluded elsewhere because it fails at another condition. Successful solutions have a maximum site-balance error of 2.1×10^{-13}, and reacting CO coverage remains below the CO-only equilibrium value. A numerical solver failure is not evidence that a physical steady state is absent.

The primary calculations start from CO adsorption equilibrium before admitting O_2. For cases with multiple stable solutions, we also start from the state selected at the preceding pressure. The solutions differ at 26 states: eight at 443 K and 18 at 463 K, all between 0.1259 and 0.5012 kPa CO. Some draws switch between CO-rich and O-rich surfaces, but the largest change in a group median is below 0.0005. None of the 0.01-kPa comparisons changes. Both solutions and their starting conditions are archived.

A fixed parameter group can change its leading route as conditions change. At 0.01 kPa CO, 23, 58, 79, 95, 114 and 136 draws from the original assisted group become direct across the six temperatures. All 151 original direct draws remain direct through 423 K, and 150 do so at 443 and 463 K. There are no ties or unresolved assignments at this pressure. The curves thus follow the same parameter groups across conditions, with both routes present.

Proposed coverage measurement. The predicted difference suggests measuring absolute CO coverage as a function of CO pressure at 358 and 463 K, holding O_2 at 10 kPa and including 0.01 kPa CO. At that pressure the direct/assisted group medians change from 0.960/0.994 at 358 K to 0.145/0.772 at 463 K, increasing their gap from 3.4 to 62.7 percentage points. The experiment should maintain partial conversion and measure gas composition at the catalyst, so that nominal inlet pressure can be related to the pressure entering the model. Absolute coverage and its measurement uncertainty are needed to compare with these predictions; disappearance of the infrared signal under complete conversion is insufficient. The overlapping prediction intervals and joint inference errors make this a candidate test of the inferred surface states, without establishing an optimal design or a guaranteed mechanism assignment.

S4 Combining kinetic and energetic evidence

The current CO oxidation model conditions one joint posterior on measured rates, CO desorption measurements and uncertain DFT energies. The hydrogen evolution case uses kinetic measurements alone. This section defines the thermal prior and the evidence model used for the predictions in Figure 5c; the surface balances, source values and experimental conditions are given in Supplementary S3.3.

Parameters and prior. The 16 coordinates have independent uniform priors over the ranges in Table S12. All reference constants are defined at $T_{\mathrm{ref}} = 358$ K. Positive constants are sampled uniformly on their listed $\log_{10}$ coordinates. The same prior is used for both catalysts; their measured rates and evidence condition separate posteriors.

Table S12: Prior for the explicit CO oxidation model. Bounds apply to the listed coordinates. Rate constants use seconds and pressures in bar: $k_{\mathrm{CO,ads}}$, $k_{\mathrm{O_2,ads}}$ and k_{a} have units $\mathrm{bar}^{-1}\,\mathrm{s}^{-1}$; k_{d} and k_{LH} have units s^{-1}; adsorption equilibrium constants have units bar^{-1}. Energies are in eV. The site scale s is dimensionless, and σ_x is the standard deviation of $\log_{10}$ TOF noise, in decades.

Coordinate	Lower	Upper
$\log_{10} k_{\mathrm{CO,ads,ref}}$	3	8
$\log_{10} K_{\mathrm{CO,ref}}$	2	10
$E_{\mathrm{CO,ads}}$	0	0.3
ΔH_{CO}	−2	−0.4
$\log_{10} k_{\mathrm{O_2,ads,ref}}$	−4	8
$\log_{10} K_{\mathrm{O_2,ref}}$	−3	6
$E_{\mathrm{O_2,ads}}$	0	1.5
$\Delta H_{\mathrm{O_2}}$	−1	−0.01
$\log_{10} k_{\mathrm{d,ref}}$	−4	7
E_{d}	0	1.5
$\log_{10} k_{\mathrm{a,ref}}$	−4	8
E_{a}	0	1.5
$\log_{10} k_{\mathrm{LH,ref}}$	−2	8
E_{LH}	0	1.5
$\log_{10} s$	−0.3	0.3
$\log_{10} \sigma_x$	−2	−0.5

Evidence and conditional distribution. Each parameter draw predicts five additional observables,

$$h(\boldsymbol{\theta}) = (E_{\mathrm{CO,des}},\ \log_{10} k_{\mathrm{CO,des}}(303\ \mathrm{K}),\ \Delta H_{\mathrm{CO}},\ E_{\mathrm{O_2,ads}},\ E_{\mathrm{a}})\,, \tag{43}$$

where $E_{\mathrm{CO,des}} = E_{\mathrm{CO,ads}} - \Delta H_{\mathrm{CO}}$ and

$$\log_{10} k_{\mathrm{CO,des}}(303\ \mathrm{K}) = \log_{10} k_{\mathrm{CO,ads,ref}} - \log_{10} K_{\mathrm{CO,ref}} - \frac{E_{\mathrm{CO,des}}}{k_{\mathrm{B}} \ln 10} \left(\frac{1}{303\ \mathrm{K}} - \frac{1}{358\ \mathrm{K}} \right). \tag{44}$$

The first two coordinates describe desorption measurements; the final three are compared with DFT energies. The assisted barrier is referenced to gas-phase O_2 and adsorbed CO, as in the source comparison. Electronic energies approximate effective kinetic enthalpies in this analysis.

Let $\mathbf{x}$ contain measured $\log_{10}$ rates and let $\mathbf{y}_{\mathrm{exp}}$ and $\mathbf{y}_{\mathrm{DFT}}$ contain the available desorption and DFT values. Conditional on $\boldsymbol{\theta}$, rate errors are independent Gaussian with standard deviation σ_x. Desorption errors are independent Gaussian with their supplied standard deviations. DFT errors have likelihood

$$L_{\mathrm{DFT}}(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}_{\mathrm{DFT}}; h_{\mathrm{DFT}}(\boldsymbol{\theta}), \Sigma_{\mathrm{DFT}}), \qquad \Sigma_{\mathrm{DFT}} = D_\sigma \left[(1-\rho) I + \rho \mathbf{1}\mathbf{1}^{\mathsf{T}} \right] D_\sigma, \tag{45}$$

where D_σ contains the three supplied DFT standard deviations. The target is

$$p(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}_{\mathrm{exp}}, \mathbf{y}_{\mathrm{DFT}}, \boldsymbol{\sigma}, \rho) \propto p_0(\boldsymbol{\theta})\, p_x(\mathbf{x} \mid \boldsymbol{\theta}) L_{\mathrm{exp}}(\boldsymbol{\theta})\, L_{\mathrm{DFT}}(\boldsymbol{\theta}). \tag{46}$$

Missing evidence contributes no likelihood term. The reported main predictions use $\rho = 0$ and DFT standard deviations of 0.10 eV. D087 includes both desorption measurements; D009 includes only desorption energy. No numerical surface coverage is supplied.

Training and input representation. Each training batch generates fresh Gaussian log-rate noise and fresh energetic evidence. Three input modes are chosen with equal probability: rates alone, rates with desorption evidence, and rates with desorption and DFT evidence. Evidence standard deviations are sampled uniformly in their logarithms: 0.03–0.15 eV for desorption energy, 0.03–0.30 decades for the desorption rate, and 0.03–0.50 eV for each DFT energy. The DFT error correlation is sampled uniformly from $[0, 0.8]$.

Within each rate series, an invertible transformation retains the first log rate and subtracts it from every other log rate. Noise is added before this transformation. The 62 encoder inputs contain the 22 padded rate coordinates, 22 observation masks, two catalyst indicators, five evidence values, five evidence masks, five log standard deviations and ρ. Rate and evidence values are normalized using training statistics. Missing rate and evidence entries are set to zero; missing evidence also has zero log-standard-deviation input, and ρ is zero when DFT evidence is absent.

The encoder supplies a 128-coordinate context to an eight-layer rational quadratic spline flow with eight bins. Parameter coordinates are mapped from their bounded ranges to logit coordinates. Training minimizes negative log density of the simulated parameter vectors, without interval-containment or variance penalties. The library contains 65,536 training, 4,096 validation and 4,096 test campaigns. Adam uses learning rate 3×10^{-4}, cosine decay, batch size 1,024 and 500 epochs; validation loss selects the checkpoint. The retained inference checkpoint preserves the trained ActNorm parameters by marking their initialization complete. Posterior draws are transformed back to the original coordinates before simulation, with no selection by reconstruction error.

S5 Baselines and ablations

Comparing model components. Table 1a compares flow matching with a rational quadratic neural spline flow using the same signal encoder and training set.[6,11] Both are evaluated on the same 9,005 clean E and solution chemistry campaigns. The spline model has 14 coupling layers, eight bins per transform and ActNorm.[15] It uses the containment penalty below; flow matching does not. Reconstruction uses the same separate 120-campaign subset and averages 16 complete simulated currents per case before scoring. The 91- and 25-ms sampling times use 1,000 draws on a GPU and exclude training and forward simulation. The retained timing record does not specify the GPU model, repeat count or whether encoding is included, so it does not support a hardware-matched speed ratio.

Replacing curves with summaries. To test the signal encoder, we keep each posterior estimator fixed and replace the CNN and Set-Transformer with a multilayer perceptron. Its inputs are 21 summary statistics: peak potentials and currents, peak separation, half-peak width and charge integral for each scan rate. Such summaries are commonly used in simulation based inference.[16,17] The resulting Summary-FMPE and Summary-RQNSF models use the same losses as their signal counterparts: no containment penalty for flow matching and the interval containment penalty for the spline flow. The learned signal encoder improves step recovery (Table 1a). Summary-RQNSF intervals contain the true values almost universally (1.00 mean and 0.99 minimum coverage at the nominal 0.90 level). This indicates intervals that are too broad, rather than accurate uncertainty estimates. Overcoverage persists across the tested values of λ_{cal}. Summary-RQNSF reconstruction uses a new fixed subset of 120 campaigns from the same test population, because the historical ordered subset was unavailable. Its median error is 6.6%, calculated from the 1,318 successful original posterior simulations; 602 of 1,920 draws failed or timed out. A failed scan excludes the complete draw. No replacement draws, parameter adjustment or rejection by reconstruction error are used. All 120 cases retain at least one successful draw, but none retain all 16. Large finite errors remain included: 13 of 120 case errors exceed 1,000%. The median therefore describes reconstruction conditional on successful simulation and does not summarize its poor upper tail.

Noise during training. Training with measurement noise improves step detection and primary-rate recovery even when evaluation signals are clean (Table S13). Capacitance and resistance remain in the physical simulator in every comparison.

Table S13: Effect of measurement noise during training. "Clean" and "noisy" denote measurement noise augmentation off and on. Capacitance and resistance remain physical forward-model effects in every cell. The bold row is the deployment setting. Scores use the original evaluation draws. The primary rate is present in every example; its R^2 therefore uses all examples.

Train	Test	cov. (90%)	micro-F_1	exact set	k^0 R^2
clean	clean	0.886	0.681	0.408	0.606
clean	noisy	0.879	0.636	0.365	0.520
noise-aug	clean	0.909	0.792	0.563	0.786
noise-aug	**noisy**	**0.908**	**0.786**	**0.554**	**0.762**

Containment penalty for the spline baseline. The spline baselines use central intervals at nominal masses $\{0.5, 0.7, 0.8, 0.9, 0.95\}$, estimated from 128 joint draws. The generic flow matching model and the three case-specific estimators do not use this penalty.

For the outer sphere spline-flow baselines, b indexes batch items, d parameter coordinates, $\ell \in \mathcal{L}$ nominal interval levels and k posterior draws. Normalized truth θ_{bd} and central interval $[L_{bd\ell}, U_{bd\ell}]$ define

$$I^{\text{soft}}_{bd\ell} = \text{sigm}(a_{\text{cal}}(\theta_{bd} - L_{bd\ell}))\ \text{sigm}(a_{\text{cal}}(U_{bd\ell} - \theta_{bd}))\,, \tag{47}$$

$$I^{\text{ST}}_{bd\ell} = I^{\text{hard}}_{bd\ell} + I^{\text{soft}}_{bd\ell} - \text{stopgrad}(I^{\text{soft}}_{bd\ell}), \tag{48}$$

$$\mathcal{L}_{\text{cal}} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \text{mean}_{b,d}\left[w_{bd}(I^{\text{ST}}_{bd\ell} - \ell)^2\right], \qquad w_{bd} \propto \max\Big(\text{sd}_k(\theta^{(k)}_{bd}), 10^{-4}\Big)^{-1}. \tag{49}$$

The weights are normalized to mean one and held fixed during gradient calculation. Here $I^{\text{hard}}_{bd\ell} = \mathbf{1}[L_{bd\ell} \le \theta_{bd} \le U_{bd\ell}]$. Adjacent sigmoid factors in the first line are multiplied, and the boundary sharpness a_{cal} increases from 5 to 200 after warm-up. The forward pass squares individual containment errors before averaging; it is not the batch-level objective $(\text{mean}\, I - \ell)^2$ and therefore does not directly target nominal coverage. The equation specifies the implemented penalty, which encourages wider intervals. Calibration is assessed from empirical coverage.

Optimization with a supplied mechanism. The earlier comparison in Table 1b reports median errors of 1.3% with the true active step set and 6.0% with a seeded random set constrained to differ from truth. Its 120 campaigns are the historical reconstruction cohort in Table 1a, where SPARK gives 4.3%. The original

ordered case list, per-case fits and measured optimization times were not recovered; these entries retain the previously reported results.

The supplied step set remains fixed during each fit. Parameters used by the primary electron transfer, the selected steps and double layer capacitance are optimized within their bounds; transfer coefficients are fixed at 0.5. Omitted rates stay at the simulator floor, and selected rates are bounded above their activity thresholds. Both settings use eight starts, a 500-s allowance per start and a total 3,600-s limit per campaign. The first start uses a signal based guess and later starts are random. The objective sums mean squared current errors across scans; reported errors are normalized by each scan's current range. These settings differ from the full network comparison below, which frees all 42 coordinates. The older 23–632-s timing range came from a separate nine-mechanism benchmark with 15 starts; it is not a timing measurement for these eight-start fits.

Optimization without a supplied mechanism. The full network comparison in Table 1b uses the same 120 campaigns for both methods, with 36 E and 84 solution chemistry examples. A fixed random permutation of the archived 15,000-campaign library (seed 12345) selects this cohort. It also supplies the separate Summary + RQ-NSF reconstruction above. It is a new selection; its overlap with the historical subset in Table 1a is unknown. The simulator, parameter ordering and checkpoint are fixed before the comparison. Known parameters reproduce all 120 stored campaigns, with maximum NRMSE 1.80×10^{-8}.

For optimization, all 42 coordinates are scaled to $[0, 1]$ using the actual generating bounds, with a finite difference step of 0.001 in these scaled coordinates. The continuous bounds include the gaps between generating bands. The simulator's reference activity thresholds and scan conversion remain unchanged, including flat regions for negligible steps. A start can try up to eight prior draws to obtain a complete simulation; failure checks count toward its 450-s budget and do not use reconstruction error. Each search allows at most 150 iterations and 15,000 objective evaluations, with a 60-s limit per forward scan. The true parameters, family and active steps are excluded from initialization and fitting.

All 960 starts produce a complete reconstruction, but only 25 meet the optimizer's stopping criteria; 935 reach the time limit. The selected best fit comes from a converged search in 3 of 120 cases. Median reconstruction error is 8.89% for L-BFGS-B and 3.99% for SPARK (Figure S8a,b). L-BFGS-B gives the smaller error in 45 cases. The respective 95th percentiles are 21.7% and 14.4%; all finite errors, including SPARK's largest error of 426.5%, are retained. Median summed search time is 3,600 seconds of CPU worker time per case, with 118.98 worker hours across all 120 cases; parallel execution reduces elapsed time. This is not a controlled hardware speed ratio.

Saved search trajectories show continued improvement near the time limit (Figure S8c). Increasing the allowance from 300 to 450 s per start lowers median error from 9.83% to 8.89%. The objective improves in 109 of 120 cases, by more than 1% in 101. These are results within the stated budget, not estimates of the best achievable error after convergence. Search frequencies are not posterior probabilities.

Time for one optimization start. Table 1b uses one signal based first start for each of the same 120 timing cases in all three settings. For the full network, we use the first of the eight archived starts. For true and random mechanisms, we run a separate timing benchmark with the original parameter constraints and optimizer settings, one start and a 500-s limit. The frozen simulator and input curves match the full network comparison. Each setting uses one CPU thread per search; paired true/random cases run on the same CPU node. Timers include objective evaluations but exclude input loading, program startup and final scoring. The full network timer also includes its initial feasible-candidate check. The reported errors still select the best of eight starts and are not the errors achieved within the one-start times.

Timing ranges retain starts that produce complete finite predictions for all three scans, including starts stopped at the time limit. The true/random checks also require an objective below the simulator failure penalty. No filtering uses reconstruction accuracy. A constant failure penalty can satisfy a numerical stopping criterion without producing a valid signal, so we count such outcomes as failed predictions. Table S14 reports both the complete-prediction subset and all attempts. The 500-s limit is checked between objective evaluations, so recorded times can slightly exceed it. Four true-mechanism starts and four random-mechanism starts ended with abnormal optimizer termination but retained complete predictions. The two failed random-mechanism predictions returned the fixed failure penalty and a numerical convergence flag; they are excluded from the complete-prediction timing range. The new timing runs use Intel Xeon Platinum 8592+ CPUs, NumPy 1.26.4

and SciPy 1.17.1. Each optimization setup retains its own solver settings as stated above; these are practical execution times, not a controlled speed ratio.

Table S14: Outcomes of the one-start timing benchmark. Each setting uses the same 120 signals. A complete prediction contains finite currents for all three scans. The stopping-criteria count also requires a complete prediction; the time-limit count covers all attempts. Ranges are measured elapsed seconds, rounded to the nearest second. Main Table 1 uses the complete-prediction range.

Mechanism supplied	Complete predictions	Met stopping criteria	Reached time limit	Complete-prediction range (s)	All-attempt range (s)
True	120/120	22	94	124–506	124–506
Random	118/120	15	99	124–507	44–507
None (full network)	120/120	1	119	425–450	425–450

Applying the same bounds to SPARK. The approximate SPARK posterior can extend outside the optimizer's generating bounds: 1,757 of the first 1,920 joint draws have at least one such coordinate, affecting 5.36% of coordinate values. These training ranges are not universal limits of chemical validity. As a separate check, we retain the first 16 of each case's saved 512 draws that lie within all 42 optimizer bounds. We do not clip values, generate replacement draws or select by reconstruction error. There are 114 cases with 16 eligible draws and six with 7–15, giving 1,896 successful complete predictions and no excluded cases. Median error is 3.87%. This check enforces the same outer bounds, not the gaps or the full joint structure of the generating prior.

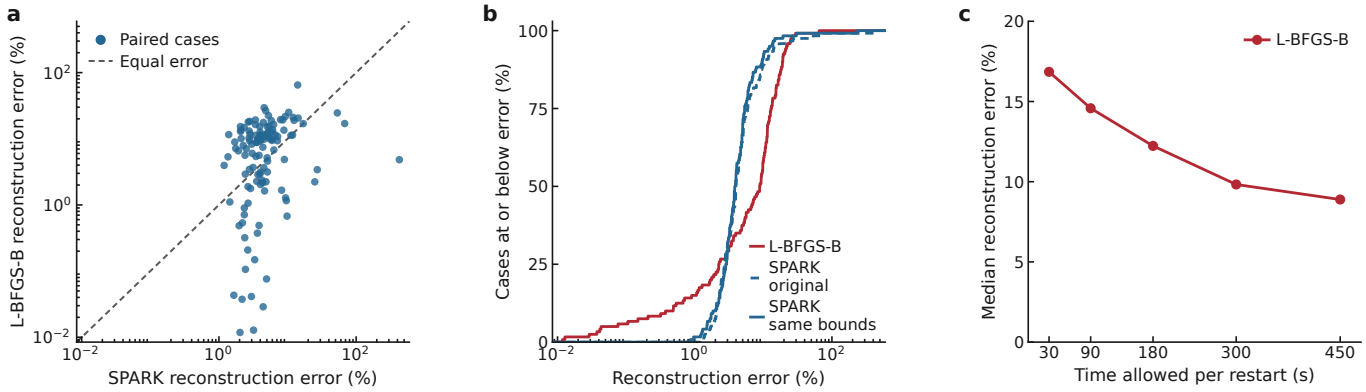


Figure S8: Full network optimization and SPARK on the same signals. All comparisons retain the same 120 examples. (a) Reconstruction errors for the best L-BFGS-B parameter set and SPARK's mean prediction from 16 joint draws. Points below the diagonal favor optimization. (b) Cumulative error distributions, including a separate SPARK calculation using draws within the optimizer's bounds. All finite errors are shown. (c) Median optimization error against the allowed time per start, reconstructed from saved searches. Each case uses eight starts; the objective selects the best candidate at every time cutoff. All 120 cases have valid predictions at each cutoff.

S6 Outer sphere reaction-network dictionary

The step probabilities in Figure 1 use the shared controls defined here. Supplementary Table S15 lists the elementary reactions and their parameters. Supplementary Table S16 maps combinations of active steps to common mechanism names.

For each joint posterior draw, we apply the activity thresholds to all 15 optional controls and record the complete combination of active steps. The primary electron transfer is always included, and linked reactions share the controls listed in Supplementary Table S15. The thresholds are $\tau_k = -2$ for the $\log_{10}$ controls at the model's reference scale. Draws with the same complete combination form one mechanism group. If $\mathbf{z}_{\mathcal{M}}$

denotes that combination, its posterior probability is

$$P(\mathcal{M} \mid \mathbf{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left[\mathbf{z}^{(m)} = \mathbf{z}_{\mathcal{M}}\right]. \tag{50}$$

The denominator includes all posterior draws, including when only the most probable groups are displayed. We count whole draws to preserve dependencies among parameters; multiplying the separate step probabilities would lose these dependencies. The naming table gives familiar labels to some combinations. Other combinations retain their full set of active steps, including mixtures of familiar mechanisms.

Table S15: Outer sphere reaction-network dictionary. A step is labeled present when its rate coordinate exceeds the activity threshold; linked steps share the indicated gate. The listed coordinates enumerate all 42 inferred parameters. R_u is an unlabeled nuisance and is not included.

Symbol	elementary process and named region(s)	presence coordinate	other coordinates
			$\log_{10} k_1^0$, $\log_{10} \lambda_1$, α_1;
k_1^0	$A + e^- \rightleftharpoons B$ *E; reference couple for all regions*	always on	E_1^0, $\log_{10} D_A$, $\log_{10} D_B$
k_2^0	$B + e^- \rightleftharpoons C$ *EE; potential-inversion chemistry*	$\log_{10} k_2^0$	$\log_{10} \lambda_2$, α_2, ΔE_2^0, $\log_{10} D_C$
k_c	$B \rightarrow Z$ *EC; with k_Z^0: ECE; with k_{disp}: DISP*	$\log_{10} k_c$	—
k_Z^0	$Z + e^- \rightleftharpoons Z_r$ *ECE*	$\log_{10} k_Z^0$	$\log_{10} \lambda_Z$, α_Z, ΔE_Z^0
k_{disp}	$Z + B \rightarrow A + Z_r$ *DISP*	$\log_{10} k_{\mathrm{disp}}$	—
k_{cat}	$B \rightarrow A$ *catalytic EC′*	$\log_{10} k_{\mathrm{cat}}$	—
k_f	$W \rightleftharpoons A$ *CE*	$\log_{10} k_f$	$\log_{10} K_{\mathrm{eq}}$
k_{dim}	$2B \rightarrow P$ *radical–radical coupling; DIM*	$\log_{10} k_{\mathrm{dim}}$	—
$k_{\mathrm{rad-sub}}$	$A + B \rightarrow P$ *radical–substrate coupling*	$\log_{10} k_{\mathrm{rad-sub}}$	—
k_{dc}	$A + C \rightleftharpoons 2B$ *disproportionation/ comproportionation*	$\log_{10} k_{\mathrm{dc}}$	equilibrium tied to ΔE_2^0
			$\log_{10} \lambda_{\mathrm{med}}$, α_{med};
k_{med}^0	$M_o + e^- \rightleftharpoons M_r$ *mediated redox catalysis*	$\log_{10} k_{\mathrm{med}}^0$	E_{med}^0, $\log_{10} D_{\mathrm{med}}$
k_{med}	$M_r + A \rightarrow M_o + B$ *mediated redox catalysis*	$\log_{10} k_{\mathrm{med}}$	—
$K_{\mathrm{ads},A}$	$A \rightleftharpoons A_{\mathrm{ads}}$ *Langmuir/Frumkin adsorption*	$\log_{10} K_{\mathrm{ads},A}$	$\log_{10} \Gamma_{\mathrm{sat}}$, g_{Frumkin}
$K_{\mathrm{ads},B}$	$B \rightleftharpoons B_{\mathrm{ads}}$ *adsorption-coupled surface chemistry*	slaved to $K_{\mathrm{ads},A}$	$\log_{10} K_{\mathrm{ads},B}$
k_{surf}^0	$A_{\mathrm{ads}} + e^- \rightleftharpoons B_{\mathrm{ads}}$ *surface-confined ET; Laviron limit*	$\log_{10} k_{\mathrm{surf}}^0$	$\log_{10} \lambda_{\mathrm{surf}}$, α_{surf}, E_{surf}^0
k_{surfchem}	$A_{\mathrm{ads}} \rightarrow B_{\mathrm{ads}}$ *surface EC*	$\log_{10} k_{\mathrm{surfchem}}$	—
k_{PT}	$A + H^+ \rightleftharpoons AH$; $B + H^+ \rightleftharpoons BH$; $AH + e^- \rightleftharpoons BH$ *PCET; thermodynamically closed square scheme*	$\log_{10} k_{\mathrm{PT}}$	pK_a offset, ΔE_H^0
C_{dl}	double layer current *cell physics; not a chemical mechanism*	activity readout from $\log_{10} C_{\mathrm{dl}}$	—

Table S16: Elementary step combinations and common mechanism names. Every row includes the primary electron transfer $A + e^- \rightleftharpoons B$. The listed optional controls are above their activity thresholds; every unlisted optional control is below its threshold. These rows define complete step combinations. Additional active steps give a different combination, which can contain more than one familiar pathway. E denotes electron transfer and C denotes a chemical reaction.

Name	Optional elementary reactions	Active controls ($\log_{10}$)
E	None	None
EC	$B \rightarrow Z$	k_c
ECE	$B \rightarrow Z$; $Z + e^- \rightleftharpoons Z_r$	$k_c,\ k_Z^0$
DISP	$B \rightarrow Z$; $Z + B \rightarrow A + Z_r$	$k_c,\ k_{\mathrm{disp}}$
EC′	$B \rightarrow A$ (catalytic regeneration)	k_{cat}
CE	$W \rightleftharpoons A$	k_f
EE	$B + e^- \rightleftharpoons C$	k_2^0
PCET square scheme	$A + H^+ \rightleftharpoons AH$; $B + H^+ \rightleftharpoons BH$; $AH + e^- \rightleftharpoons BH$	k_{PT} (shared control)