BATTERYOCV: A PAIRED DATASET AND BENCHMARK FOR HIGH-TO-LOW RATE PROFILE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Battery measurements involve a trade-off between speed and diagnostic value. Higher currents shorten measurement time, but resistance, reaction kinetics, and ion-transport limits shift the measured voltage away from the low-rate reference used to study degradation. Lower currents reduce these effects and provide a more consistent diagnostic reference, but a complete low-rate charge–discharge measurement can take more than 40 hours. We ask how well this reference can be reconstructed from faster, lower-fidelity measurements of the same cell. BatteryOCV contains 182,519 high-rate cycle records and 1,894 repeated low-rate references from 404 industrial NCM/graphite cells, paired on the same physical cells as they degrade. The benchmark task predicts a measured low-rate voltage profile and its charge and discharge capacities from neighboring high-rate voltage cycles, using cell-disjoint splits. We compare signal-processing and classical controls with autoencoder and generative baselines. On unseen cells of the training design, copying, averaging, or smoothing the high-rate curves leaves the discrepancy from the reference largely unchanged, whereas learned mappings reduce it substantially, and a linear ridge model is competitive with the neural networks. Capacity and charge-aligned errors capture what profile error misses, and differential-voltage peak counts provide a coarse check of local features. On held-out product designs, errors increase markedly for both linear and neural models. BatteryOCV provides paired data and a fixed protocol for measuring how much of a slow diagnostic reference can be recovered from fast measurements.

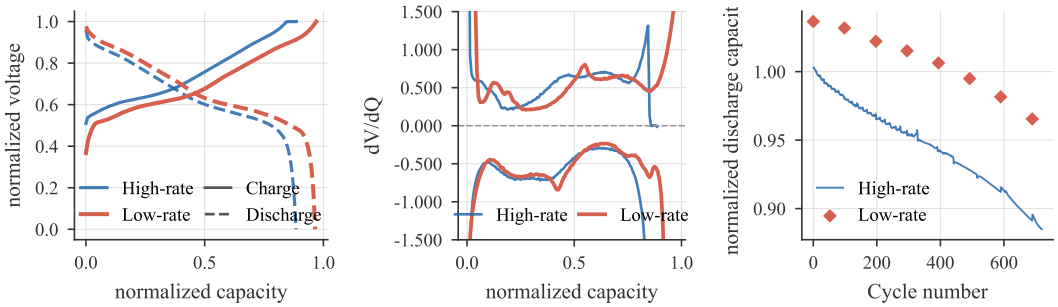
1 INTRODUCTION

Tracking how a lithium-ion cell degrades requires a measurement that can be compared across its life. Low-rate charge–discharge curves serve this purpose: at low current, polarization is small, and the voltage–capacity curve and its derivative used in differential voltage analysis (DVA) reflect changes in lithium inventory and active material (Dubarry et al., 2012; Birkl et al., 2017; Weng et al., 2023). Routine cycling runs at higher currents, where resistance, reaction kinetics, and ion transport distort the curve (Birkl et al., 2017; Barai et al., 2019). As a result, low-rate references are measured only occasionally, since a complete low-rate charge–discharge measurement can take more than 40 hours, whereas the cycling itself runs at high rate, so every cycle yields a high-rate curve. This paper asks how well these faster, lower-fidelity measurements can reconstruct the slow reference.

Fidelity here refers to diagnostic value relative to the low-rate reference, not to instrument accuracy. A high-rate curve records the cell under load and differs systematically from the reference. BatteryOCV pairs routine high-rate cycling with low-rate references measured under a standardized protocol at a common temperature, also for cells cycled at other temperatures (Section 3.3). Following prior DVA studies (Dubarry et al., 2012; Birkl et al., 2017; Weng et al., 2023), we treat the 0.05C reference curve as OCV-like. Figure 1 shows how the measurement rate changes the voltage curve, its differential-voltage features, and the observed capacity trajectory.

Why paired measurements are needed. Training and evaluating a reconstruction model requires a measured reference for every prediction. High-rate cycling histories support capacity and lifetime prediction but contain no low-rate profile. Pairing curves from different cells would confound the rate difference with cell-to-cell variation and degradation. BatteryOCV links each low-rate reference

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Voltage–capacity profile (b) $\mathrm{d}V/\mathrm{d}Q$ (DVA) (c) Capacity trajectory

Figure 1: **(a)** Profiles of the same cell at nearby life stages, measured at 0.33C and 0.05C. Polarization shifts and distorts the high-rate profile relative to the low-rate reference. **(b)** The distortion suppresses diagnostic structure in $\mathrm{d}V/\mathrm{d}Q$. **(c)** High-rate capacity varies across individual cycles, while periodic low-rate measurements provide a consistent view of the degradation trajectory.

to neighboring high-rate cycles of the same cell at the same life stage, and repeats these references along each cell's cycling history, so the mapping is supervised as the cell degrades.

We treat the problem as low-to-high-fidelity reconstruction by conditional regression. A model receives a window of high-rate voltage curves and predicts the low-rate charge–discharge profile together with its two capacities. The capacities are part of the target because profiles are stored on a branch-normalized capacity grid, and shape alone does not give the capacity range on which voltage–capacity and DVA analyses are read. The default window is centered on the reference cycle, so the task reconstructs the reference at an observed life stage.

Evaluation. The primary test holds out cells of Model E, the design with most of the paired data. Controls that predict a constant reference, copy, average, or smooth the high-rate curve, or fit simple corrections show how much error a model removes beyond these simple explanations. Linear, retrieval, and neural baselines are evaluated on the same task. Profile, capacity, charge-aligned, and peak-count errors each measure a different aspect of agreement with the reference. A separate test on held-out designs measures how far the mapping learned on Model E transfers. This split mirrors a common industrial situation. A cell design that reaches production has accumulated long cycling and validation records, whereas a new design starts with few cells and little or no degradation testing. A reconstruction model would in practice be trained on the established design and applied to the new one, where low-rate references are scarcest.

Existing battery datasets and paired-rate studies provide cycling and diagnostic measurements (Table 1). BatteryOCV adds repeated same-cell pairing across hundreds of industrial cells, measured low-rate profile targets, and a common reconstruction protocol. Its cycling histories can also be used for temporal tasks under their own observation and prediction rules.

Contributions. **(1) Paired measurement data.** BatteryOCV pairs the routine high-rate cycling of hundreds of industrial NCM/graphite cells with repeated low-rate references from the same cells, which allow direct supervision and evaluation of the rate-to-reference mapping across degradation.

(2) Reconstruction benchmark. The task predicts the low-rate voltage profile and its charge and discharge capacities from high-rate voltage observations. Fixed cell-disjoint splits evaluate unseen cells within a design and, separately, held-out product designs, using complementary reconstruction metrics.

(3) Baseline results. On unseen Model-E cells, learned mappings reduce profile and capacity errors relative to copying the high-rate curve and to a constant reference. Voltage-only ridge regression is competitive with the autoencoders and has lower capacity error than any of them. On held-out designs, errors increase substantially for both linear and neural models.

Table 1: Comparison with public battery datasets and paired-rate studies. *Systematic pair*: repeated high-/low-rate profiles from the same physical cells across degradation.

Dataset / study	# Cells	Chemistry	Temp.	Fast rate	Slow rate / pair
CALCE (Pecht, 2013)	13	LCO	25 °C	0.5–1C	No
Severson et al. (Severson et al., 2019)	124	LFP	30 °C	3.6–6C	Limited[†]
Oxford (Birkl & Howey, 2017)	8	LCO/NCO	40 °C	Drive cycle	Yes
Che et al. (Che et al., 2026)	94	NCA/SiO_x–Gr	25 °C	C/5	C/40, study-specific
BatteryML (Zhang et al., 2024)	383[*]	Mixed	Mixed	Mixed	Source-dependent
BatteryLife (Tan et al., 2025)	990[*]	Mixed	Mixed	Mixed	Source-dependent
BatteryOCV	**404**	**NCM/Gr**	**20–50 °C**[‡]	**0.33, 0.5C**	**0.05C, systematic**

[*] Aggregation platforms: diagnostic coverage is inherited from constituent datasets. [†] Some cells include C/10 diagnostics, without systematic pairing throughout degradation.
[‡] BatteryOCV reports the measured temperature range; nominal protocol groups are audited in Appendix A.4.

2 Related Work

2.1 Low-rate Diagnostics and Physical Modeling

Differential voltage analysis (DVA) uses the local slope of a voltage–capacity curve to study degradation. Under near-equilibrium measurement conditions, changes in peak locations and spacing can inform analyses of lithium inventory and active-material loss (Dubarry et al., 2012; Birkl et al., 2017; Weng et al., 2023). Higher currents introduce polarization that changes the measured curve and its derivatives (Barai et al., 2019). Porous-electrode models (Newman & Tiedemann, 1975; Doyle et al., 1993) and reduced-order variants (Moura et al., 2017) relate measured voltage to internal states, but require parameter identification and modeling assumptions. BatteryOCV instead learns from data: the supervised target is the measured low-rate full-cell curve, and no electrode state is inferred.

2.2 OCV and Paired-rate Reconstruction

Tian et al. (Tian et al., 2021) reconstruct a complete OCV–capacity curve by estimating electrode-aging parameters from a partial 1C charging segment, with validation on eight Oxford cells. Zhou et al. (Zhou et al., 2024) learn an autoencoder representation of OCV–capacity curves and map partial-charge features into that representation. Both address a missing part of the voltage range. In BatteryOCV the full high-rate charge–discharge range is observed, and the difficulty lies in the rate difference. Our partial-observation experiment varies observation coverage within this task; it does not reproduce either method.

Che et al. (Che et al., 2026) study rate-adaptive diagnosis on 94 cells measured at C/5 and C/40. Their pipeline uses pristine half-cell open-circuit-potential measurements to fit mechanistic states, maps high-rate states to low-rate equivalents, and applies a residual correction. BatteryOCV also connects faster measurements to slower references, but provides full-cell voltage windows and measured profile targets and requires neither half-cell measurements nor pre-fitted mechanistic states. Mechanistic, representation-mapping, and residual-correction approaches can be adapted to the benchmark once their input requirements are stated.

2.3 Battery Datasets and Learning Benchmarks

CALCE (Pecht, 2013), Severson et al. (Severson et al., 2019), and Oxford (Birkl & Howey, 2017) have supported battery degradation research, with different cycling protocols and diagnostic coverage. BatteryML (Zhang et al., 2024) and BatteryLife (Tan et al., 2025) aggregate resources and standardize degradation or lifetime evaluation. BatteryGPT (Hu et al., 2026) predicts future operating trajectories and downstream state of health from early-life time series. BatteryOCV also contains cycling histories suited to such temporal tasks, but the benchmark here targets the slow reference at an observed life stage. Measured low-rate full-cell references supervise predictions from faster measurements of the same cells, and the evaluation uses cell-disjoint splits, simple learned and non-learned controls, and separate profile and capacity metrics.

Table 2: BatteryOCV statistics by product model. *Low-rate targets* are the matched diagnostic profiles; *high-rate cycles* are the operational records.

Model	# Cells	Low-rate targets	High-rate cycles
A	8	40	3,578
B	7	15	1,837
C	4	17	3,062
D	5	30	2,913
E	380	1,792	171,129
Total	**404**	**1,894**	**182,519**

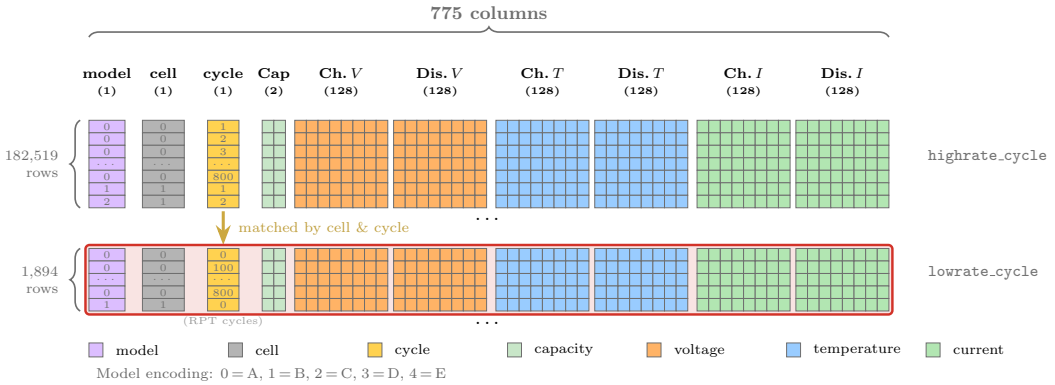


Figure 2: BatteryOCV data structure. Each raw row has 775 columns: seven metadata values and charge/discharge voltage, temperature, and current profiles sampled at 128 points per branch. The benchmark downsamples each voltage branch to 64 points and concatenates them into a 128-point model input and target. Low-rate rows are matched to high-rate rows by cell and cycle index. The schematic omits the two scalar charge/discharge temperature metadata fields; these are included in the seven metadata values and the 775-column total.

3 Dataset Description

3.1 Resource and Normalization

BatteryOCV contains 404 industrial NCM/graphite cells from five anonymized product designs, with 1,894 measured low-rate references and 182,519 high-rate cycles (Table 2). All designs share a broad chemistry family and a common voltage window. Disclosed capacity classes are B/E Small and A/C/D Large; detailed design parameters are not provided as model inputs.

Current and capacity are normalized by nominal capacity. Current is therefore expressed as C-rate, whereas normalized capacity is dimensionless. Voltage is min–max normalized over this common window, whose limits are not disclosed, so profile RMSE is dimensionless. The released rows contain voltage, current, and temperature profiles, but the benchmark uses only high-rate voltage as input. The measured low-rate profile and its capacities serve only as targets.

3.2 Pairing and Observation Availability

Low-rate characterization is inserted periodically during long-term cycling. Each reference is indexed by physical cell and cycle, and matched to a centered window of $W{=}64$ neighboring high-rate cycles. The surrounding input cycles are nearby observations of the same cell. Programmed temperature and rate are generally fixed within a window, so the window gives repeated observations under one operating condition. Loader rules for the few targets without a matching high-rate cycle are given in Appendix A.

3.3 Reference Protocol and Domain Structure

Low-rate references use a standardized reference performance test at approximately 25 °C, including for cells cycled at other temperatures. A low-current discharge to the lower voltage limit precedes the full 0.05C charge–discharge measurement. This reduces polarization and gives a common reference for cells with different operating histories. A trained operator verifies the reference profiles and repeats defective measurements.

High-rate measurements retain transient disturbances, with temperature and rate varying mainly across physical cells and protocol groups. We keep the imbalance between designs on purpose. Model E supplies 380 cells and plays the role of an established design with long test histories; A–D together supply 24 cells and play the role of new designs with little paired data and limited degradation testing. Model E supports the primary cell-disjoint evaluation, and A–D the held-out-design test. The cohorts also differ in degradation coverage: about 20% of Model-E cells lose more than 10% of their initial reference discharge capacity, up to 22.9%, whereas the maximum observed loss on A–D is 7.8%. The held-out groups thus lack the heavily degraded tail of Model E, although their cycling histories are not uniformly shorter and their median capacity loss is not uniformly smaller. Section 4.2 defines the evaluation split; Appendix A.4 reports operating-condition mixtures and degradation coverage.

4 Benchmark Task Definition

4.1 Low-to-High-Fidelity Reconstruction Task

Given neighboring high-rate voltage profiles $\mathbf{X} \in \mathbb{R}^{W \times d}$ from one cell, the model predicts the measured low-rate profile $\mathbf{y} \in \mathbb{R}^d$ and its normalized charge and discharge capacities $\mathbf{c} \in \mathbb{R}^2$:

$$(\hat{\mathbf{y}}, \hat{\mathbf{c}}) = f_\theta(\mathbf{X}). \tag{1}$$

The default is W=64 and d=128: 64 charge points followed by 64 discharge points. Inputs and target are linked by cell and cycle index. The profile lies on a branch-normalized capacity grid, so its voltage shape does not determine its capacity scale; the two capacity outputs place each branch on the nominal-capacity-normalized axis. Both capacities are measured at the reference cycle.

Because the centered window includes cycles after the reference, the task is retrospective: it reconstructs the reference at an observed life stage. Forecasting or strictly causal reconstruction would require a different observation rule.

4.2 Cell-disjoint Reconstruction Evaluation

Both evaluations keep the cells used for fitting and model selection separate from the test cells. Except in the supplementary adaptation study, only Model-E training and validation cells are used for fitting and selection.

Primary unseen-cell test. Model E is divided into 265/39/76 train/validation/test cells using split seed 42, with 1,231/191/370 low-rate targets. This test measures reconstruction on new physical cells of the design seen in training. The split is fixed across training seeds.

Held-out-design test. This test reproduces the setting in which a model trained on an established design must serve a new design with little data of its own. Designs A–D contain 24 cells and 102 references. No A–D example is used for parameter fitting, checkpoint selection, or hyperparameter selection. All designs are NCM/graphite, but product design, capacity class, and protocol mixture change together, so the test measures transfer to related designs without isolating a single factor. A–D also lack the heavily degraded cells present in Model E (Section 3.3).

The two evaluations are reported separately. Pooled errors weight designs and cells by their number of references, and per-design results show the resulting heterogeneity. Appendix B.13 reports a supplementary study in which one or two cells per held-out design supply low-rate references and the remaining cells are used for testing; it is kept separate from the zero-shot results.

4.3 Complementary Reconstruction Metrics

Voltage shape and capacity scale. Profile RMSE is computed over the normalized voltage points, and capacity RMSE over both normalized capacity outputs. Charge-aligned RMSE interpolates the reference and prediction onto their shared charge-capacity interval and measures voltage disagreement there. It reflects shape and scale together over that interval but ignores discharge alignment and the non-overlapping capacity tail, so we report it together with capacity RMSE. Appendix B.2 gives the aggregation rules.

Local-feature count. DVA examines $\mathrm{d}V/\mathrm{d}Q$ structure. We smooth each voltage branch with a Savitzky–Golay filter (window 7, polynomial order 2), take finite differences on its capacity axis, and detect peaks with prominence 0.02. Peak-count error is

$$\mathcal{M}_{\mathrm{cnt}} = \frac{1}{2N} \sum_{j=1}^{N} \sum_{b \in \{\mathrm{ch},\mathrm{dis}\}} \left| N_{\mathrm{peak}}(\hat{\mathbf{y}}_{j,b}) - N_{\mathrm{peak}}(\mathbf{y}_{j,b}) \right|, \tag{2}$$

where peaks are detected on the predicted or reference capacity scale, respectively. Equal counts do not imply correct peak positions or amplitudes, so we use peak count as a coarse check of local features alongside the other metrics. Figure 3 shows examples.

Selection and reporting. Autoencoder, VAE, and GAN checkpoints are selected by Model-E validation profile RMSE; the diffusion model uses its validation loss. The Optuna configurations were chosen with the original multi-metric objective, which included a matched-peak position error. We do not report that position error because it does not penalize unmatched peaks (Appendix B.1).

5 Experiments

5.1 Experimental Setup

Fitting and selection use only Model-E training and validation cells. The primary test has 370 references from 76 unseen Model-E cells; the A–D test has 102 references from 24 cells. Supplementary experiments with target-design support cells are in Appendix B.13.

Seven autoencoders and five generative models each reuse one Optuna-selected configuration and are retrained with seeds 0, 1, and 42 (Akiba et al., 2019). Means and sample standard deviations describe retraining variation on fixed test sets; Optuna is not rerun per seed. Classical controls use validation-selected settings, and the partial-observation experiment uses a separate fixed Conv configuration with seed 42. Appendix B.1 lists the selection objectives.

The reference configuration is a Conv AE with voltage-only input, $W{=}64$, segment bias, and differential loss (Figure 5, Appendix B.10). It is the setting of the input ablation and of the architecture search; the window ablation disables segment bias and differential loss, and the component ablation toggles them. Ablations and architecture comparisons use different optimization settings, so we compare results only within each group.

5.2 Reconstruction from High-rate Measurements

5.2.1 Constant, signal-processing, and learned controls

The controls separate three sources of a low profile error: predicting a typical low-rate reference, processing the high-rate observations directly, and learning an input-dependent mapping from paired examples. The training-target mean predicts the same low-rate profile and capacities for every input. Identity copies the center high-rate cycle; temporal mean and median summarize neighboring cycles; Savitzky–Golay smoothing filters the center profile. Pointwise affine learns a correction at each voltage coordinate. Ridge combines center, window-mean, standard-deviation, and change features, and 1-NN retrieves a paired training target from voltage features. Appendix B.3 reports all twelve controls.

On Model E, profile RMSE is 0.02746 for identity, 0.02743 for temporal mean, and 0.02747 for smoothing: averaging or smoothing the high-rate curves leaves the discrepancy from the reference

Table 3: Selected controls on the 370-target Model-E test. Lower is better. V: voltage summaries; charge-only alignment. All twelve methods: Appendix Table 8.

Method	Profile RMSE	Cap RMSE	Cap-aligned RMSE (ch.)	Peak count
Training-target mean	0.00998	0.04779	0.01941	0.04865
High-rate identity	0.02746	0.04253	0.03770	0.51216
Pointwise affine	0.00972	0.03548	0.01652	0.04865
Ridge (V)	0.00429	0.01647	0.00670	0.05405
1-NN retrieval	0.00433	0.01924	0.00661	0.04865

Table 4: Seven autoencoders on the primary 370-target Model-E test; voltage-only input. Alignment is charge-only. Entries are mean $\pm$ sample standard deviation over three retraining seeds after fixing each architecture's Optuna-selected configuration. **Bold**: lowest mean per column.

Model	Params	Profile RMSE ↓	Cap RMSE ↓	Cap-aligned RMSE ↓	Peak count ↓
Conv+LSTM AE	331K	$\mathbf{0.00362 \pm 0.00014}$	0.01891 ± 0.00046	$\mathbf{0.00615 \pm 0.00011}$	0.062 ± 0.001
LSTM AE	538K	0.00418 ± 0.00019	0.02019 ± 0.00095	0.00673 ± 0.00020	0.098 ± 0.033
Conv AE	160K	0.00397 ± 0.00022	$\mathbf{0.01787 \pm 0.00094}$	0.00646 ± 0.00033	0.063 ± 0.014
MLP AE	4.6M	0.00440 ± 0.00017	0.02320 ± 0.00130	0.00804 ± 0.00019	$\mathbf{0.052 \pm 0.003}$
UNet AE	4.9M	0.00378 ± 0.00021	0.01924 ± 0.00059	0.00637 ± 0.00023	0.084 ± 0.027
BiLSTM AE	967K	0.00755 ± 0.00244	0.03224 ± 0.01124	0.01252 ± 0.00459	0.064 ± 0.018
Transformer AE	431K	0.00553 ± 0.00086	0.02772 ± 0.00886	0.01007 ± 0.00273	0.060 ± 0.014

essentially unchanged. The training-target mean reaches 0.00998, so a typical low-rate curve is already a better profile prediction than the high-rate input itself. Pointwise affine reaches 0.00972, and voltage-only ridge and retrieval reach 0.00429 and 0.00433 (Table 3). The reduction below the constant reference comes from input-dependent corrections.

Capacity inputs differ across control families. Identity and the signal-processing controls use observed high-rate capacities, and the affine corrections are fitted on them; Ridge (V), retrieval, and the reference neural models receive no capacity input.

5.2.2 Neural reconstruction baselines

Table 4 compares seven autoencoders on the same voltage-only task. All of them have lower mean profile and capacity RMSE than the training-target mean. Conv+LSTM AE reaches profile RMSE 0.00362 ± 0.00014 and capacity RMSE 0.01891 ± 0.00046; Conv AE reaches 0.00397 ± 0.00022 and 0.01787 ± 0.00094. With the classical results, this shows that high-rate voltage carries information about the low-rate profile and capacity of an unseen cell of the training design.

Voltage-only ridge is competitive with these models. Its profile RMSE (0.00429) lies between those of LSTM AE and MLP AE, and its capacity RMSE (0.01647) is lower than that of any autoencoder, so reconstruction on Model E does not require a neural model. The rankings reflect the tested configurations and training budgets.

5.2.3 Generative comparison

Three VAEs, a GAN, and a conditional-diffusion model extend the comparison beyond deterministic autoencoders. None of them improves on the best autoencoder profile error on Model E (Appendix B.11).

5.3 Reconstruction Fidelity beyond Voltage-shape Error

Among the autoencoders, Conv+LSTM has the lowest mean profile and charge-aligned errors, Conv the lowest capacity error, and MLP the lowest peak-count error (Table 4). No single metric orders the models consistently, which is why we report all four.

The input ablation separates shape from scale. Adding scalar high-rate capacity as an input raises profile RMSE from 0.00358 to 0.00376 but lowers capacity RMSE from 0.01664 to 0.01292 and

Table 5: Controls from Table 3 and selected autoencoders on 102 targets from held-out designs A–D. Alignment is charge-only. Classical rows report single evaluations; neural rows show mean ± sample standard deviation over seeds 0, 1, and 42. No A–D example is used for fitting or selection. Full results: Appendix Tables 9 and 18.

Method	Profile RMSE ↓	Cap RMSE ↓	Cap-aligned RMSE ↓	Peak count ↓
Training-target mean	0.01774	0.02932	0.02371	0.00000
High-rate identity	0.03741	0.06835	0.04563	1.00980
Pointwise affine	0.01857	0.02992	0.02400	0.00000
Ridge (V)	0.01269	0.03854	0.01347	0.01471
1-NN retrieval	0.01295	0.03920	0.01612	0.00000
Conv+LSTM AE	0.01228 ± 0.00158	0.04313 ± 0.00847	0.01624 ± 0.00216	0.145 ± 0.045
Conv AE	0.01195 ± 0.00228	0.03878 ± 0.00994	0.01533 ± 0.00227	0.162 ± 0.073
MLP AE	0.01074 ± 0.00031	0.05883 ± 0.01889	0.01737 ± 0.00232	0.007 ± 0.006
BiLSTM AE	0.01368 ± 0.00342	0.02919 ± 0.00032	0.01563 ± 0.00481	0.052 ± 0.091
Transformer AE	0.01270 ± 0.00093	0.03447 ± 0.00771	0.01533 ± 0.00123	0.000 ± 0.000

charge-aligned RMSE from 0.00598 to 0.00514 (Appendix Table 16). A ranking by profile error alone would miss this gain in capacity. Charge-aligned error covers only the common charge-capacity interval, so it is read together with capacity error.

Peak count is a coarser check. The training-target mean has peak-count error 0.04865 on Model E with profile RMSE 0.00998, and zero count error on A–D with profile RMSE 0.01774 (Table 5): a constant curve with the right number of peaks scores well on this metric. We therefore report peak count only together with profile and capacity errors. Small count differences are also sensitive to numerical details (Appendix B.3).

5.4 Observation, Context, and Component Controls

In the window ablation, $W{=}64$ has the lowest mean peak-count error and $W{=}16$ lower profile and capacity errors (Appendix B.5), so no window size is best on all metrics. In the fixed Conv setting, segment bias and differential loss together reduce peak-count error from 0.133 ± 0.011 to 0.086 ± 0.011; neither component alone does (Appendix B.8).

In practice, both sides of a reference are not always available. Reconstructing the reference at the start of a test leaves only the cycles that follow it, and estimating the reference at the latest measured point leaves only the cycles before it. We therefore train separate models with both-sided, previous-only, and next-only context. On the common 244-target interior subset, they give mean profile RMSE 0.00403, 0.00399, and 0.00386 and capacity RMSE 0.01671, 0.01655, and 0.01761, so either side alone performs about as well as both. At the natural boundaries, next-only models reach profile RMSE 0.00280 on the initial targets and previous-only models 0.00380 on the final targets; these subsets contain different targets and are not a paired comparison (Appendix B.6–B.7). One-sided context is therefore usable at both ends of a test history. Local features are less robust: previous-only context raises interior peak-count error from 0.192 to 0.345, and the final-boundary subset has peak-count error 0.464.

5.4.1 Partial-observation control

Following work on partial-charge reconstruction (Tian et al., 2021; Zhou et al., 2024), a separate $W{=}1$ Conv experiment compares the full high-rate charge branch with its first, middle, or last contiguous 50%. On Model E, the full branch gives lower profile, capacity, and charge-aligned errors than each half; on A–D the ordering differs (Appendix B.4). This is a single-seed ablation of observation coverage.

5.5 Held-out Product Designs

To test the setting in which a model trained on an established design serves new designs (Section 4.2), we apply the Model-E models to A–D without adaptation. Table 5 shows selected results; Appendix B.12 gives full pooled and per-design results.

Every method in Table 5 has higher pooled profile RMSE on A–D than on Model E. Ridge rises from 0.00429 to 0.01269 and Conv+LSTM AE from 0.00362 to 0.01228, about $3.0\times$ and $3.4\times$. The two test sets are different populations, so these ratios are descriptive, but they show that accuracy on unseen Model-E cells does not carry over to related designs.

Profile error alone again hides differences between outputs. MLP AE has the lowest held-out profile RMSE among the autoencoders (0.01074) but capacity RMSE 0.05883, against 0.02919 for BiLSTM AE and 0.02932 for the training-target mean. Per design, B has the lowest capacity error for all seven autoencoders. LSTM AE's pooled held-out capacity RMSE is 0.04450, 0.04254, and 0.13367 at seeds 0, 1, and 42, with the largest errors of the last run on C and D, although its Model-E results are stable across seeds.

Model-E training data already cover the main 25/45 °C and 0.33/0.5C groups (Appendix A.4), so the error increase is not explained by an unseen temperature or rate alone. Product design, capacity class, protocol mixture, and degradation coverage all change together in A–D, and this experiment does not separate them.

Supplementary reference-assisted reconstruction. Appendix B.13 adds a study in which one or two cells of each held-out design provide low-rate references, with testing on the remaining cells and an input-independent support-mean control. Because it uses target-design labels, it is reported separately from the zero-shot results. Simply averaging the references of one or two target cells lowers mean profile RMSE on the remaining cells from 0.01006 for the zero-shot Conv AE to about 0.0043, close to what the adapted models reach, whereas fine-tuning on the same references mainly improves capacity (0.01446 with two cells versus 0.02138 for the average). Much of the held-out profile gap thus reflects a design-level shift in the typical reference.

6 Discussion and Conclusion

BatteryOCV pairs repeated high-rate and low-rate measurements on the same cells and defines the task of reconstructing the slow reference profile and its capacities from fast measurements, with fixed cell-disjoint evaluation. It allows measuring, across observed stages of degradation, how much of a diagnostic reference fast measurements can recover.

On unseen Model-E cells, learned mappings bring profile and capacity errors well below those of copying the high-rate curve or predicting the training-target mean, and voltage-only ridge is competitive with the autoencoders. Profile error does not reflect errors in capacity scale, and equal peak counts do not imply equal peak positions, so the four metrics are reported together. On held-out designs, profile errors increase about threefold. Closing this gap matters in practice, because new designs are where slow references are scarcest.

Limitations. The target is a measured 0.05C profile rather than equilibrium OCV, and reconstruction accuracy is not a validated electrode-level diagnosis. The centered window makes the task retrospective. Model E contributes about 94% of cells, reflecting the intended imbalance between established and new designs. The four held-out designs change design, operating, and degradation-coverage factors together, do not form a balanced sample of battery products, and lack the heavily degraded cells present in Model E. Charge-aligned RMSE covers only the shared charge-capacity interval, and peak count does not assess peak positions or amplitudes. Three retraining seeds measure optimization variation under fixed splits. Limitations of the supplementary reference-assisted study are listed in Appendix B.13.

Research use. The paired references can be used to study measurement-fidelity reconstruction, sample-efficient learning, and physically informed corrections, with the measured low-rate reference as the common target. Natural extensions are transfer to designs with little paired data, shorter or more informative observation protocols, broader design and degradation coverage, and independent diagnostic validation.

AI USE STATEMENT

Generative AI tools assisted methodological discussion, code development and refactoring, consistency checks between the manuscript and archived experimental records, and editing for clarity. The dataset consists of physical battery measurements, and all reported numbers come from executed experimental pipelines. The authors are responsible for verifying the code, results, and claims, and for the final manuscript and all associated artifacts.

ETHICS STATEMENT

BatteryOCV contains industrial battery measurements, with no human subjects or personally identifiable information. Product identifiers are anonymized as A–E, nominal capacities are normalized, and proprietary design details are withheld (Section 3). Reconstructed references are not certified safety assessments or validated electrode-level diagnoses, and the cross-design errors and limitations in Section 6 apply to any use beyond the evaluated setting.

REPRODUCIBILITY STATEMENT

Sections 4–5 specify the low-to-high-fidelity profile-and-capacity reconstruction task, its retrospective observation rule, fixed cell-disjoint evaluations, and reported comparisons. Appendix A documents the data format, normalization, measurement protocol, and exceptional loader behavior. Appendix B.1 records training settings, checkpoint selection, the legacy Optuna objective, and three-seed retraining; Appendix B.2 distinguishes the reported metrics and their aggregation. Classical inputs, architecture definitions, pooled A–D results, and per-design results are specified in the remaining appendices. Appendix B.13 reports the supplementary reference-assisted analysis separately, including support/query separation, fine-tuning and refitting settings, support-mean controls, the label-using constant diagnostic, and reproduction checks. Fixed split metadata and archived reference tables accompany the code. The HDF5 measurements and split metadata are available for review at an anonymized link: `https://dataverse.harvard.edu/previewurl.xhtml?token=aae9111a-4310-4a44-a62b-0145a5775021`.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.

Anup Barai, Kotub Uddin, W. Dhammika Widanalage, Andrew McGordon, Paul Jennings, and James Marco. A comparison of methodologies for the non-invasive characterisation of commercial Li-ion batteries. *Journal of Power Sources*, 434:226238, 2019. doi: 10.1016/j.jpowsour.2019.226238.

Christoph Birkl and David Howey. Oxford battery degradation dataset 1, 2017. URL `https://ora.ox.ac.uk/objects/uuid:03ba4b01-cfed-46d3-9b1a-7d4a7bdf6fac`.

Christoph R. Birkl, Matthew R. Roberts, Euan McTurk, Peter G. Bruce, and David A. Howey. Degradation diagnostics for lithium ion cells. *Journal of Power Sources*, 341:373–386, 2017. doi: 10.1016/j.jpowsour.2016.12.011.

Yunhong Che, Joachim Schaeffer, Jinwook Rhyu, Liang Wu, Patrick A. Asinger, Minsu Kim, Jacob Sass, Rolf Findeisen, Martin Z. Bazant, William C. Chueh, and Richard D. Braatz. Battery aging assessment: from critical insights to enhanced diagnosis. *Energy & Environmental Science*, 19: 1954–1968, 2026. doi: 10.1039/D5EE06439B.

Marc Doyle, Thomas F. Fuller, and John Newman. Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. *Journal of the Electrochemical Society*, 140(6):1526–1533, 1993. doi: 10.1149/1.2221597.

Matthieu Dubarry, Cyril Truchot, and Bor Yann Liaw. Synthesize battery degradation modes via a diagnostic and prognostic model. *Journal of Power Sources*, 219:204–216, 2012. doi: 10.1016/j.jpowsour.2012.07.016.

Jincheng Hu, Pengyu Fu, Zhongbao Wei, Yanjun Huang, Juliana Early, Ashley Fly, and Yuanjian Zhang. Early prediction of lithium-ion battery degradation with a generative pre-trained transformer. *Nature Communications*, 17:126, 2026. doi: 10.1038/s41467-025-66819-0.

Scott J. Moura, Federico Bribiesca Argomedo, Reinhardt Klein, Anahita Mirtabatabaei, and Miroslav Krstic. Battery state estimation for a single particle model with electrolyte dynamics. *IEEE Transactions on Control Systems Technology*, 25(2):453–468, 2017. doi: 10.1109/TCST.2016.2571663.

John Newman and William Tiedemann. Porous-electrode theory with battery applications. *AIChE Journal*, 21(1):25–41, 1975. doi: 10.1002/aic.690210103.

Michael Pecht. CALCE battery group dataset, 2013. URL `https://calce.umd.edu/battery-data`.

Kristen A. Severson, Peter M. Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H. Chen, Muratahan Aykol, Patrick K. Herring, Dimitrios Fraggedakis, Martin Z. Bazant, Stephen J. Harris, William C. Chueh, and Richard D. Braatz. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4:383–391, 2019. doi: 10.1038/s41560-019-0356-8.

Ruifeng Tan, Weixiang Hong, Jiayue Tang, Xibin Lu, Ruijun Ma, Xiang Zheng, Jia Li, Jiaqiang Huang, and Tong-Yi Zhang. BatteryLife: A comprehensive dataset and benchmark for battery life prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5789–5800, 2025. doi: 10.1145/3711896.3737372.

Jinpeng Tian, Rui Xiong, Weixiang Shen, and Fengchun Sun. Electrode ageing estimation and open circuit voltage reconstruction for lithium ion batteries. *Energy Storage Materials*, 37:283–295, 2021. doi: 10.1016/j.ensm.2021.02.018.

Caihao Weng, Matthieu Dubarry, et al. Differential voltage analysis for battery manufacturing process control. *Frontiers in Energy Research*, 11, 2023. doi: 10.3389/fenrg.2023.1087269.

Han Zhang, Xiaofan Gui, Shun Zheng, Ziheng Lu, Yuqi Li, and Jiang Bian. BatteryML: An open-source platform for machine learning on battery degradation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2310.14714`.

Ziyou Zhou, Yonggang Liu, Chengming Zhang, Weixiang Shen, and Rui Xiong. Deep neural network-enabled battery open-circuit voltage estimation based on partial charging data. *Journal of Energy Chemistry*, 90:120–132, 2024. doi: 10.1016/j.jechem.2023.11.009.

A Dataset Details

A.1 Measurement Protocol

Low-rate reference profiles are collected using a standardized reference performance test (RPT) protocol designed to approximate near-equilibrium conditions. Each measurement is conducted at approximately 25 °C. Prior to the 0.05C cycle, the cell is discharged to the lower voltage limit under low current to minimize residual overpotential, followed by a full charge–discharge cycle at 0.05C. Each profile is manually verified by a trained operator; measurements exhibiting abnormal behavior are discarded and repeated.

High-rate cycling is performed under CC–CV charge to the upper voltage limit (cutoff at C/20) and CC discharge to the lower voltage limit at 0.33C or 0.5C, depending on cell model, at nominal chamber temperatures of 25 °C, 30 °C, or 45 °C. Instrumentation noise, contact instability, and transient interruptions are retained to preserve realistic operating conditions.

A.2 Data Format

Each raw cycle is uniformly resampled along the capacity axis at 128 points for charge and 128 points for discharge. Voltage, temperature, and current use the same format. A row therefore has 775 values: seven metadata fields (model, cell, cycle, charge/discharge capacity, and two scalar charge/discharge temperatures) followed by six profile channels of 128 points. The two scalar temperature fields are omitted from the schematic in Figure 2. The public HDF5 file stores the rows in `highrate_cycle` and `lowrate_cycle` datasets. The benchmark downsamples each voltage branch to 64 points and concatenates charge and discharge into the 128-point model profile.

In the released loader, two training targets have no exact-cycle high-rate row. Their 64-cycle inputs are all-zero windows; they are retained in the training target count. There are no such cases in validation, Model-E test, or A–D. The single-cycle loader instead selects the first available cycle at or after the target, clipped to the last row. This handling differs between window sizes, and reproductions should keep it unchanged.

A.3 Preprocessing and Filtering

We apply the following steps to ensure data quality. Cycles that do not traverse the full voltage window are removed, as are records with persistent equipment failure or corrupted signals. Low-rate targets are checked by an operator, and defective measurements are repeated. Transient high-rate disturbances are retained. All current and capacity values are normalized by nominal capacity.

A.4 Operating-Protocol Audit

The public file stores measured temperature and current. For this descriptive audit, we assign each high-rate record to the nearest nominal protocol in the measurement specification: 25/30/45 °C and 0.33/0.5C discharge. Table 7 shows Model-E training and each held-out design. Model-E training already spans both major temperature groups and both discharge rates. The held-out-design gap therefore does not stem from a single unseen operating condition; temperature mixture, capacity class, and product design vary together.

Table 6 summarizes observed degradation coverage. For each cell, capacity loss is one minus the lowest normalized low-rate discharge capacity divided by that of its first reference. A–D omit the more severely degraded tail present in Model E, but their histories are not uniformly shorter: C has a higher median last cycle and median capacity loss than E. The lower within-group profile variation reported in Appendix B.13 is consistent with this difference in coverage, although the audit cannot show that degradation range is its only cause. Neither the held-out-design results nor the adaptation study cover degradation beyond the observed range.

B Extended Experimental Details

Table 6: Degradation coverage of the low-rate references by design. Cycle is the target cycle index of the last reference per cell.

Design	Cells	References per cell (med.)	Last cycle (med.)	Capacity loss median	Capacity loss max	Cells with >10% loss
E	380	4	393	4.6%	22.9%	19.7%
A	8	6	455	4.4%	7.0%	0%
B	7	2	99	1.8%	3.0%	0%
C	4	4.5	763	6.8%	7.8%	0%
D	5	5	490	4.8%	5.8%	0%

Table 7: Observed high-rate operating-condition audit. Each record is assigned to the nearest nominal temperature using the mean of its charge and discharge temperatures, and to the nearest nominal discharge rate. Percentages are within each row. Capacity classes are the disclosed Small/Large groups.

Model	Class	Cycles	25 °C	30 °C	45 °C	Discharge 0.33C / 0.5C
E train	Small	116,488	49,376 (42.4%)	256 (0.2%)	66,856 (57.4%)	88,781 (76.2%) / 27,707 (23.8%)
A	Large	3,578	1,709 (47.8%)	0	1,869 (52.2%)	3,578 (100%) / 0
B	Small	1,837	0	0	1,837 (100%)	1,837 (100%) / 0
C	Large	3,062	1,515 (49.5%)	0	1,547 (50.5%)	3,062 (100%) / 0
D	Large	2,913	1,950 (66.9%)	10 (0.3%)	953 (32.7%)	2,913 (100%) / 0

B.1 Training Details

The original window, component, and input ablations use AdamW with learning rate 10^{-3}, $\beta = (0.9, 0.95)$, weight decay 10^{-2}, and up to 3,000 epochs. The tables report mean and standard deviation over three random seeds (0, 1, 42). Architecture and generative comparisons instead use separately selected Optuna hyperparameters. For both comparisons, one selected configuration per model is held fixed and retrained with seeds 0, 1, and 42; the tables report the mean and sample standard deviation. The search includes learning rate, weight decay, loss weights, and architecture parameters, so optimizer settings differ across models.

The profile decoder and capacity head are trained jointly using

$$\mathcal{L} = \lambda_p \,\mathrm{MSE}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_c \,\mathrm{SmoothL1}(\hat{\mathbf{c}}, \mathbf{c}) + \lambda_d \,\mathrm{MSE}(\Delta\hat{\mathbf{y}}, \Delta\mathbf{y}), \tag{3}$$

where Δ denotes consecutive profile differences. The fixed ablation weights are $(\lambda_p, \lambda_d, \lambda_c) = (100, 100, 5)$ when differential loss is enabled, and $\lambda_d = 0$ otherwise. Segment bias is an input encoding, separate from these loss terms.

The additional partial-observation experiment fixes seed 42, a single-cycle input, 128 profile outputs, latent dimension 128, batch size 32, and the same fixed AdamW settings. Segment bias and differential loss are disabled, leaving profile and capacity weights 100 and 5. It uses 3,000 epochs and selects the minimum validation profile RMSE. DVA evaluation records a smoothing window of 7 and polynomial order 2; the partial-observation call specifies prominence 0.02.

For the autoencoder, VAE, and GAN comparisons, training retains the checkpoint with the lowest Model-E validation profile RMSE. The four validation metrics from that checkpoint are then combined to score the trial's hyperparameter configuration, using fixed weights that rescale their different numerical ranges: $\mathcal{M}_{\text{cnt}}$ is measured in peak counts ($\sim$0.1), $\mathcal{M}_{\text{pos}}$ and $\mathcal{M}_{\text{cap}}$ in normalized capacity ($\sim$0.01–0.02), and $\mathcal{M}_{\text{profile}}$ in normalized voltage ($\sim$0.003–0.004):

$$\mathcal{O} = \mathcal{M}_{\text{cnt}} + 10 \cdot \mathcal{M}_{\text{pos}} + 5 \cdot \mathcal{M}_{\text{cap}} + 30 \cdot \mathcal{M}_{\text{profile}}. \tag{4}$$

These weights do not guarantee equal contributions on every dataset or prevent a large error in one metric from dominating the objective.

The original search used $\mathcal{M}_{\text{pos}}$, a matched-peak position error. Reference peaks are visited in capacity-axis order and matched to the nearest unused predicted peak. Distances are averaged over matched pairs, then available branches, then targets with a defined value. Unmatched peaks receive no penalty, and branches with no matches are excluded. This can give a small error despite missing or excess

peaks. We therefore omit it from the result tables, while retaining the actual selection objective here for reproducibility. The selected configurations and checkpoints are unchanged.

For the architecture comparison (Section 5.2.2) and the generative baselines (Section 5.2.3), all of lr, weight decay, $(\lambda_{\mathrm{p}}, \lambda_{\mathrm{d}}, \lambda_{\mathrm{c}})$, and the per-architecture model hyperparameters (e.g. latent dimension, hidden width, depth, dropout, number of attention heads, β_{KL} for VAEs, adversarial weight for GANs) are tuned with Optuna (TPE sampler, 20 trials per architecture, median pruner) using the objective $\mathcal{O}$ in Equation (4). Continuous parameters $\mathrm{lr} \in [10^{-4}, 10^{-2}]$ and weight decay $\in [10^{-4}, 10^{-1}]$ are searched on a log scale, and $\lambda_{\mathrm{p}}, \lambda_{\mathrm{d}} \in \{100, 300, 500\}$, $\lambda_{\mathrm{c}} \in \{5, 10, 20\}$ are categorical. The full per-architecture search ranges and selected values are released with the code.

For the three-seed architecture and generative reruns, Optuna is not repeated for each seed. Initialization, minibatch order, and other stochastic training operations change, while the selected configuration for each model remains fixed. Autoencoder, VAE, and GAN reruns retain the checkpoint with the lowest Model-E validation profile RMSE. The diffusion rerun instead retains the checkpoint minimizing its validation loss (the `loss_total` value returned by the diffusion epoch routine). Each selected checkpoint is evaluated on Model E and Models A–D. The reported seed rows contain 370 and 102 evaluation targets, respectively.

Generative-specific training. VAEs (Conv VAE, LSTM VAE, Transformer VAE) replace the deterministic latent with a Gaussian $q(\mathbf{z}|\mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, add a KL term $\beta_{\mathrm{KL}}\,\mathrm{KL}(q\|\mathcal{N}(0, I))$ to Equation (3), and use a linear KL warm-up over the first 200 epochs. At evaluation we use the deterministic latent $\mathbf{z} = \boldsymbol{\mu}$ rather than sampling, so reported numbers are not averaged over draws. Conv GAN keeps the autoencoder generator and adds a convolutional discriminator over the predicted profile; generator and discriminator are trained with separate AdamW optimizers ($\beta = (0.5, 0.999)$), label smoothing on the real labels, and an adversarial weight tuned by Optuna in $[10^{-2}, 1.0]$ (log scale).

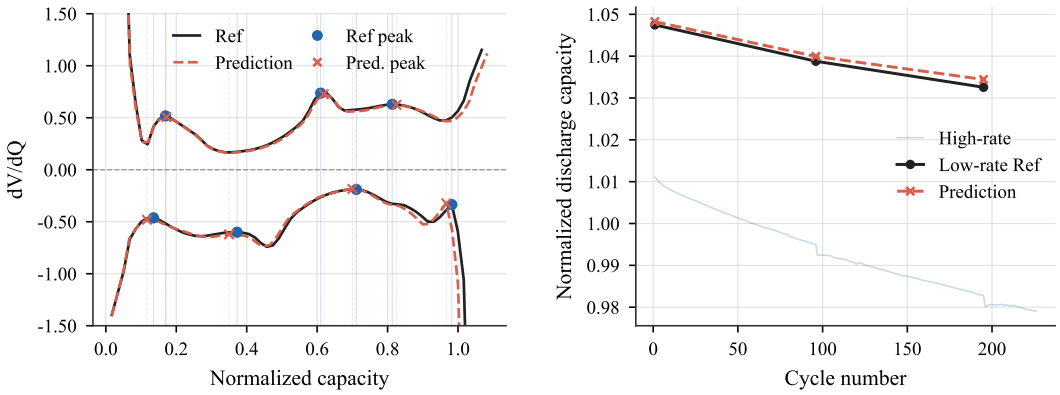
B.2 Metric Definitions and Aggregation

All error metrics are lower-is-better. Profile and capacity errors use the dimensionless normalization in Section 3.1; "Cap-aligned" denotes charge-branch alignment, not an average over both branches. Standard deviations in the neural and fixed-configuration comparisons describe training-seed variation on fixed evaluation sets. The supplementary adaptation tables report variation across support/query selections or design–selection splits, as stated in their captions.

Profile RMSE is the square root of mean squared error over all targets and all 128 profile points. Capacity RMSE is computed over all targets and both normalized capacity outputs. These two aggregate RMSEs are therefore not averages of per-sample RMSEs.

The saved evaluator computes capacity-aligned profile RMSE on the *charge branch only*. For each target, it interpolates the true and predicted charge curves to 128 uniformly spaced points on $[0, \min(c_{\mathrm{true}}, c_{\mathrm{pred}})]$, computes RMSE there, and averages the defined per-target values. A nonpositive common capacity interval gives an undefined value, which is excluded. The metric therefore excludes discharge alignment and errors beyond the shorter capacity endpoint, and we report it together with capacity RMSE.

Peak-count error averages the absolute count difference across charge and discharge, then across targets. It measures agreement in the number of detected peaks, not their locations or amplitudes. A zero count error can therefore coexist with an inaccurate curve. All result tables use the same four metrics: profile RMSE, capacity RMSE, capacity-aligned profile RMSE, and peak-count error.



(a) Cycle-level peak evaluation

(b) Cell-level degradation trajectory

Figure 3: Evaluation examples from the original Conv AE run, not three-seed averages. **(a)** Detected DVA peaks in the reference and reconstruction. **(b)** Measured and reconstructed capacities at observed cycles. These qualitative examples show local curve features and capacity changes; aggregate errors are reported in the result tables.

B.3 Classical Baseline Definitions

The training-target mean outputs the pointwise mean low-rate profile and the mean charge/discharge capacities from Model-E training cells. Every evaluation input receives the same output; it uses no input-dependent mapping. Identity copies the center high-rate cycle. Temporal mean averages the valid cycles in the $W{=}64$ window. Savitzky–Golay identity smooths the center cycle, with the window selected on Model-E validation data. Pointwise affine uses the center high-rate cycle and fits $\hat{y}_j = a_j x_j + b_j$ by ordinary least squares separately at each profile coordinate on paired Model-E training records. It also fits separate affine mappings from observed high-rate charge/discharge capacities to the corresponding low-rate capacities. These coefficients are learned from data and held fixed at evaluation; the method is a restricted linear mapping. The current-based control estimates one resistance-like voltage offset per branch and applies an affine capacity correction. Voltage-only ridge uses the center profile and the window mean, standard deviation, and first-to-last change; its regularization is selected on Model-E validation data. All targets used for fitting come from Model-E training cells.

The feature scaler is fitted only on training cells. Ridge alpha is selected by validation profile RMSE, with capacity RMSE as a tie-breaker; the voltage-summary model selects alpha 1 from the six tested values. Nearest-neighbor retrieval uses center and mean voltage features and selects $k = 1$ on validation cells.

Capacity outputs do not all receive the same inputs. Identity, temporal mean, and smoothing use observed high-rate capacity. Affine and current-dependent corrections fit capacity mappings from that observation. Voltage-summary Ridge and voltage-based retrieval do not receive observed capacity. Observed high-rate capacity is distinct from the low-rate target.

All fitted parameters use Model-E training data, and all method choices use Model-E validation data. The resulting controls are applied to A–D without refitting or adaptation. On A–D, voltage-summary Ridge has profile RMSE 0.01269, capacity RMSE 0.03854, capacity-aligned RMSE 0.01347, and peak-count error 0.01471. Retrieval has corresponding errors 0.01295, 0.03920, 0.01612, and 0.00000. These pooled results cover 102 targets from 24 held-out cells. Tables 8 and 9 include all 12 controls in the saved comparison. The three-seed architecture results are reported separately in Table 4.

The tables preserve the archived run values. A fresh CPU rerun with the environment's default numerical-library threading reproduced all archived aggregate values exactly. Restricting these libraries to one thread preserved the displayed profile and capacity errors but changed some peak counts in the voltage-only absolute and residual Ridge rows. For example, the absolute Ridge A–D peak-count error changed from 0.01471 to 0.01961, corresponding to one branch-count difference on one target. Small peak-count differences should therefore not be read as robust advantages.

The temporal median uses the same valid window as the temporal mean. The voltage-plus-capacity Ridge variants add the observed capacity's center, mean, standard deviation, and first-to-last change. Residual Ridge learns the low-rate minus center high-rate voltage and adds the center back at inference. Its capacity output is absolute in the voltage-only variant and residual relative to the observed center capacity in the voltage-plus-capacity variant.

Table 8: All classical controls on 370 Model-E targets. V denotes voltage summaries; Q denotes observed high-rate capacity. Alignment uses charge only. Lower is better.

Input / method	Profile	Capacity	Aligned (ch.)	Peak count
Training-target mean	0.00998	0.04779	0.01941	0.04865
High-rate identity	0.02746	0.04253	0.03770	0.51216
Temporal mean	0.02743	0.04288	0.03783	0.48919
Temporal median	0.02737	0.04273	0.03762	0.48919
SG identity	0.02747	0.04253	0.03765	0.43514
Pointwise affine	0.00972	0.03548	0.01652	0.04865
Current correction	0.01786	0.03548	0.02480	0.50405
Ridge, V	0.00429	0.01647	0.00670	0.05405
Ridge, V+Q	0.00487	0.01338	0.00614	0.06081
Residual Ridge, V	0.00429	0.01647	0.00670	0.06622
Residual Ridge, V+Q	0.00477	0.01198	0.00587	0.05946
1-NN, V	0.00433	0.01924	0.00661	0.04865

Table 9: All classical controls on 102 held-out-design targets. Fits use Model-E training data, method choices use Model-E validation data, and A–D evaluation uses no refitting or adaptation. Same definitions as Table 8. Lower is better.

Input / method	Profile	Capacity	Aligned (ch.)	Peak count
Training-target mean	0.01774	0.02932	0.02371	0.00000
High-rate identity	0.03741	0.06835	0.04563	1.00980
Temporal mean	0.03730	0.06820	0.04557	1.01471
Temporal median	0.03727	0.06805	0.04545	1.00980
SG identity	0.03747	0.06835	0.04560	1.11765
Pointwise affine	0.01857	0.02992	0.02400	0.00000
Current correction	0.02982	0.02992	0.02706	1.02941
Ridge, V	0.01269	0.03854	0.01347	0.01471
Ridge, V+Q	0.01176	0.03997	0.01464	0.10294
Residual Ridge, V	0.01265	0.03854	0.01346	0.27451
Residual Ridge, V+Q	0.01333	0.03561	0.01526	0.06373
1-NN, V	0.01295	0.03920	0.01612	0.00000

B.4 Partial-Observation Results

All conditions predict the full low-rate charge–discharge profile and both capacities. The charge-only conditions use either all 64 charge points or a contiguous 32-point interval: first (0:32), middle (16:48), or last (32:64). Each interval is linearly resampled to 128 input points. The full charge–discharge control uses all 128 observed voltage points. Each row is a separately trained $W{=}1$ Conv AE for the stated observation condition. Tables 10 and 11 evaluate the same five condition-specific checkpoints on the two evaluation sets. This removes points within an observed charge profile; it does not remove earlier or later cycles, as in the context experiment below. Full charge input has lower Model-E profile, capacity, and capacity-aligned errors than each charge half, but not the lowest peak-count error. On A–D the ranking changes. No A–D sample is used to train or select these models.

Table 10: Single-cycle partial-observation control on Model E. Each row is a separately trained Conv AE for the stated input condition. Seed 42; lower is better.

Input / method	Profile	Capacity	Aligned (ch.)	Peak count
Full charge + discharge	0.00372	0.01537	0.00555	0.10676
Full charge	0.00417	0.01776	0.00630	0.20811
First 50%	0.00452	0.01886	0.00668	0.14189
Middle 50%	0.00653	0.02197	0.00866	0.20676
Last 50%	0.00538	0.02361	0.00801	0.17568

Table 11: The five condition-specific Conv AE checkpoints from Table 10, evaluated on held-out designs A–D without adaptation. Seed 42; lower is better.

Input / method	Profile	Capacity	Aligned (ch.)	Peak count
Full charge + discharge	0.02004	0.04074	0.01903	1.66667
Full charge	0.01160	0.03828	0.01534	0.83824
First 50%	0.01473	0.04145	0.01782	0.27941
Middle 50%	0.01268	0.04541	0.02017	2.66176
Last 50%	0.01297	0.02549	0.01645	0.34314

B.5 Window Size Ablation

Table 12 reports the fixed-configuration three-seed window ablation comparing $W \in \{1, 16, 32, 64\}$. $W{=}64$ has the lowest mean peak-count error; $W{=}16$ has lower mean profile and capacity errors.

Table 12: Window ablation with voltage-only input, without segment bias or differential loss. Mean and standard deviation over seeds 0, 1, and 42; lower is better.

Configuration	Profile RMSE	Cap RMSE	Cap-aligned	Peak count
1	0.00393 ± 0.00085	0.01650 ± 0.00253	0.00619 ± 0.00072	0.140 ± 0.031
16	0.00356 ± 0.00038	0.01532 ± 0.00041	0.00583 ± 0.00014	0.280 ± 0.165
32	0.00362 ± 0.00046	0.01568 ± 0.00131	0.00601 ± 0.00025	0.147 ± 0.034
64	0.00362 ± 0.00030	0.01561 ± 0.00043	0.00598 ± 0.00035	0.133 ± 0.011

B.6 Controlled Context Availability

Comparing first, middle, and last targets directly would change both available context and degradation state. We instead hold the interior evaluation subset fixed (244 targets from 71 Model-E test cells) and change only the available side of a 64-slot input window. Both uses the 32 preceding records and 32 records starting at the target cycle; previous-only keeps the 32 preceding records and zero-pads the rest; next-only keeps the target and following records and zero-pads the preceding side. The Conv AE architecture (159,987 parameters) and hyperparameters are fixed across modes. Each mode is trained separately with seeds 0, 1, and 42, selecting checkpoints by Model-E validation profile RMSE. Each context condition thus has its own trained models; we do not mask the inputs of a single bidirectional checkpoint.

Table 13: Context availability on the common interior subset (244 targets, 71 cells). Mean $\pm$ sample standard deviation over three training seeds. All errors are lower-is-better; alignment uses charge only.

Context	Profile RMSE	Cap RMSE	Cap-aligned RMSE	Peak count
Both	0.00403 ± 0.00009	0.01671 ± 0.00054	0.00631 ± 0.00026	0.192 ± 0.140
Previous-only	0.00399 ± 0.00014	0.01655 ± 0.00135	0.00637 ± 0.00039	0.345 ± 0.123
Next-only	0.00386 ± 0.00012	0.01761 ± 0.00393	0.00621 ± 0.00086	0.161 ± 0.140

Using both sides gives no advantage in Table 13. Next-only has the lowest mean profile, capacity-aligned, and peak-count errors; previous-only has the lowest mean capacity error. Profile RMSE is similar in magnitude across modes, but previous-only has a higher mean peak-count error, so removing one side can hurt local-feature agreement. These are descriptive three-seed comparisons without significance testing.

B.7 Natural Boundary Performance

We separately report targets in the initial- and final-boundary subsets of the available high-rate history. Initial-boundary targets use the next-only models; final-boundary targets use the previous-only models from the controlled experiment. No boundary-specific checkpoint is selected. The reported subsets contain 70 targets from 68 cells and 56 targets from 56 cells, respectively. Together with the 244 interior targets, their target counts sum to the 370-target Model-E test set; cell counts overlap.

Table 14: Natural boundary subsets, evaluated separately from the controlled interior comparison. Mean $\pm$ sample standard deviation over seeds 0, 1, and 42; lower is better. The rows contain different targets and are not a paired comparison of context direction.

Subset / context	Profile RMSE	Cap RMSE	Cap-aligned RMSE	Peak count
Initial / next-only	0.00280 ± 0.00017	0.01526 ± 0.00201	0.00517 ± 0.00041	0.129 ± 0.031
Final / previous-only	0.00380 ± 0.00005	0.02117 ± 0.00343	0.00717 ± 0.00081	0.464 ± 0.181

Profile RMSE is 0.00280 on the initial and 0.00380 on the final subset, but final-boundary peak-count error is 0.464 ± 0.181 (Table 14). Because the subsets differ in target position and degradation state, the difference between rows cannot be attributed to context direction.

B.8 Loss Design Ablation

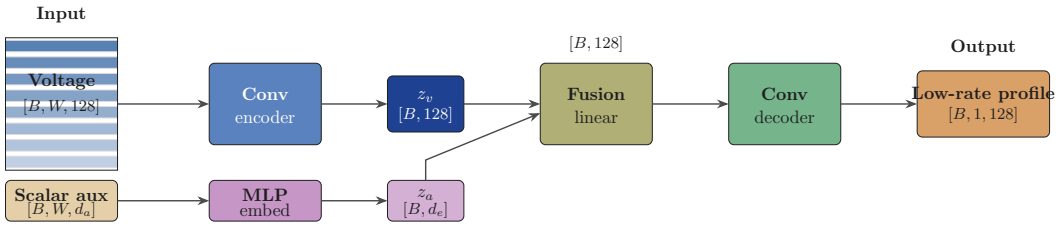
Table 15 shows that the combined segment-bias and differential-loss setting has lower peak-count error than the base setting. Neither component alone improves that metric over the base.

Table 15: Component ablation of segment bias and differential loss, $W=64$ and voltage-only input. Mean and standard deviation over seeds 0, 1, and 42; lower is better.

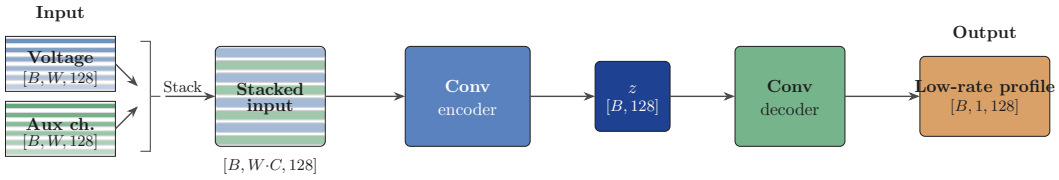
Configuration	Profile RMSE	Cap RMSE	Cap-aligned	Peak count
Base	0.00362 ± 0.00030	0.01561 ± 0.00043	0.00598 ± 0.00035	0.133 ± 0.011
Seg only	0.00360 ± 0.00024	0.01685 ± 0.00111	0.00604 ± 0.00043	0.151 ± 0.066
Diff only	0.00362 ± 0.00046	0.01587 ± 0.00101	0.00600 ± 0.00025	0.155 ± 0.069
Seg + Diff	0.00358 ± 0.00043	0.01664 ± 0.00069	0.00598 ± 0.00035	0.086 ± 0.011

B.9 Input Representation Ablation

We compare two families of auxiliary input beyond the voltage-only baseline. In the *auxiliary scalar* (Aux) mode, scalar summaries of temperature, current, or capacity are embedded through a small MLP and concatenated with the main latent vector before decoding. In the *profile channel* (Channel) mode, additional signal channels (e.g., temperature or current profiles) are stacked with the voltage profile as extra input channels, increasing the encoder input from $[B, W, 128]$ to $[B, W \cdot C, 128]$, where C is the number of profile channels. Figure 4 illustrates the two architectures.



(a) Aux mode: scalar auxiliary features embedded via MLP and fused with the voltage latent z_v.



(b) Channel mode: auxiliary profiles stacked as additional input channels; channel dimension is absorbed into the Conv encoder input width.

Figure 4: Input representation variants. **(a) Aux mode:** scalar auxiliary features (e.g. mean temperature, mean current) are extracted per cycle and embedded via a two-layer MLP into z_a, which is concatenated with the voltage latent z_v before a fusion linear layer. **(b) Channel mode:** auxiliary profiles are stacked with the voltage profile as additional input channels, yielding $[B, W, C, 128]$, which is reshaped to $[B, W{\cdot}C, 128]$ before the Conv encoder. In both cases, the decoder and output structure are identical.

Table 16 reports the fixed-configuration three-seed input ablation, including individual and combined auxiliary variables. Voltage-only input has the lowest mean peak-count error among these configurations, while some capacity-input settings have lower capacity error. Q denotes observed high-rate capacity, not the low-rate target.

Table 16: Input ablation, W=64 with segment bias and differential loss. Mean and standard deviation over seeds 0, 1, and 42; lower is better. V: voltage; T: temperature; I: current; Q: observed high-rate capacity.

Configuration	Profile RMSE	Cap RMSE	Cap-aligned	Peak count
V only	0.00358 ± 0.00043	0.01664 ± 0.00069	0.00598 ± 0.00035	0.086 ± 0.011
V+T scalar	0.00355 ± 0.00025	0.01651 ± 0.00072	0.00589 ± 0.00011	0.134 ± 0.046
V+I scalar	0.00368 ± 0.00036	0.01707 ± 0.00082	0.00618 ± 0.00054	0.089 ± 0.016
V+Q scalar	0.00376 ± 0.00033	0.01292 ± 0.00050	0.00514 ± 0.00028	0.110 ± 0.010
V+T channel	0.00361 ± 0.00045	0.01675 ± 0.00042	0.00588 ± 0.00013	0.104 ± 0.007
V+I channel	0.00390 ± 0.00024	0.01805 ± 0.00175	0.00641 ± 0.00033	0.121 ± 0.018
V+Q channel	0.00388 ± 0.00023	0.01209 ± 0.00002	0.00489 ± 0.00022	0.128 ± 0.039
V+T+I scalar	0.00386 ± 0.00027	0.01690 ± 0.00069	0.00613 ± 0.00027	0.137 ± 0.025
V+T+Q scalar	0.00374 ± 0.00025	0.01310 ± 0.00091	0.00528 ± 0.00012	0.101 ± 0.009
V+I+Q scalar	0.00373 ± 0.00025	0.01294 ± 0.00093	0.00520 ± 0.00011	0.179 ± 0.072
V+T+I+Q scalar	0.00375 ± 0.00019	0.01328 ± 0.00105	0.00511 ± 0.00004	0.112 ± 0.037
V+T+I channel	0.00375 ± 0.00031	0.01604 ± 0.00044	0.00590 ± 0.00020	0.101 ± 0.020
V+T+Q channel	0.00355 ± 0.00018	0.01198 ± 0.00060	0.00473 ± 0.00020	0.131 ± 0.031
V+I+Q channel	0.00380 ± 0.00038	0.01297 ± 0.00105	0.00506 ± 0.00013	0.147 ± 0.098
V+T+I+Q channel	0.00374 ± 0.00005	0.01327 ± 0.00137	0.00518 ± 0.00037	0.094 ± 0.016

B.10 Model Architectures

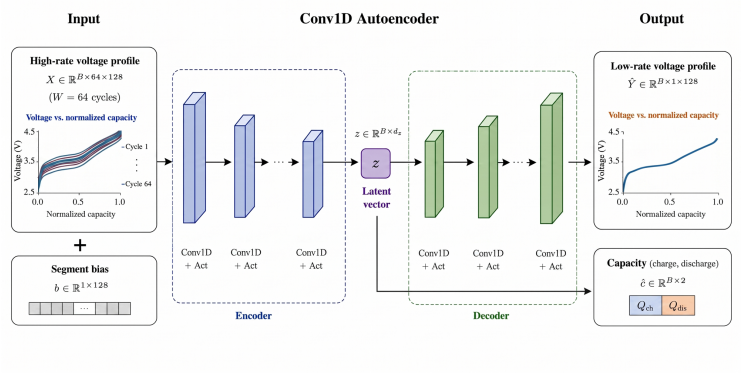


Figure 5: Reference Conv AE: W=64 neighboring high-rate profiles predict a matched 128-point low-rate profile and two capacities at the same life stage. The centered window is for retrospective reconstruction.

The autoencoders and their VAE/GAN variants use an encoder–latent–decoder structure: an encoder maps the windowed input $\mathbf{X} \in \mathbb{R}^{B\times W\times 128}$ to a latent $\mathbf{z} \in \mathbb{R}^{B\times 128}$, a profile decoder reconstructs $\hat{\mathbf{y}} \in \mathbb{R}^{B\times 1\times 128}$ from $\mathbf{z}$, and the capacity head described above predicts $(\hat{Q}_{\mathrm{ch}}, \hat{Q}_{\mathrm{dis}})$ from the same $\mathbf{z}$. Variants differ in (i) the encoder/decoder backbone and (ii) how the auxiliary input $\mathbf{x}_{\mathrm{aux}} \in \mathbb{R}^{W\times d_{\mathrm{aux}}}$ is fused.

Auxiliary input fusion. Three fusion patterns are used. *Late fusion* (Conv, MLP, UNet, Conv VAE, GAN, Diffusion) embeds $\mathbf{x}_{\mathrm{aux}}$ through a small MLP and concatenates the result with the main encoder's latent before a final fusion linear layer. *Early fusion* (LSTM, BiLSTM, Conv+LSTM, LSTM VAE) concatenates $\mathbf{x}_{\mathrm{aux}}$ to the per-cycle features at every time step before the recurrent encoder. *Token-wise conditioning* (Transformer, Transformer VAE) projects the per-cycle aux vector to the model dimension and adds it to each input token.

Discriminative variants. **Conv AE** treats the window axis W as the input channel dimension and applies three 1D conv blocks with stride 2 (channels $W\to 64\to 32\to 16$, kernel size 3, ReLU); a symmetric transposed-conv decoder reconstructs the profile. **MLP AE** flattens $\mathbf{X}$ to $W\cdot 128$ and uses a three-layer dense encoder and a three-layer dense decoder (hidden width $\in \{256, 512, 1024\}$). **LSTM AE** flattens each cycle's profile to a 128-vector, concatenates the per-cycle aux, and runs a unidirectional LSTM over the W time steps; the final hidden state is the latent. The decoder is a second LSTM that consumes the latent broadcast over 128 time steps and projects each step to a scalar

voltage. **BiLSTM AE** replaces the encoder with a bidirectional LSTM and projects the concatenated final forward and backward states back to the latent width (the decoder remains a unidirectional LSTM). **Transformer AE** projects each per-cycle profile to a token, prepends a learnable [CLS] token, adds a learnable positional embedding, and applies a Transformer encoder (2–5 layers, 2/4/8 heads, feed-forward $\in \{128, 256, 512\}$). The output [CLS] representation is the latent, and the decoder is a three-layer MLP that produces the full target profile. **UNet AE** uses a 1D UNet with three encoder levels (MaxPool stride-2 downsampling, two Conv-BN-ReLU blocks per level, base channel width $\in \{16, 32, 64\}$), a bottleneck, and a mirror-symmetric decoder with transposed-conv upsampling and skip connections between matching levels. **Conv+LSTM AE** embeds each cycle independently with three strided 1D conv blocks followed by global average pooling; the resulting per-cycle vectors are then aggregated over W by a unidirectional LSTM. The final hidden state is the latent, and the decoder is a three-layer MLP.

Generative variants. **Conv VAE**, **LSTM VAE**, and **Transformer VAE** share the encoder/decoder of their discriminative counterparts but predict a Gaussian latent $q(\mathbf{z}|\mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ via two parallel linear heads on top of the encoder output, and add a KL term to Equation (3) (training details in Appendix B.1). **Conv GAN** reuses the Conv AE as the generator and adds a 1D convolutional discriminator (four stride-2 conv blocks with LeakyReLU, BatchNorm after the first block, base channel width $\in \{8, 16, 32\}$, followed by a linear classifier) that takes a predicted or ground-truth profile and outputs a real/fake score. Generator and discriminator are optimized alternately with separate AdamW optimizers. The tested conditional-diffusion baseline (FiLM-conditioned 1D-UNet denoiser, DDIM sampling) has substantially higher peak-count error than the AE and VAE baselines (Table 17).

B.11 Generative Baseline Details and Full Results

Table 17 reports full primary Model-E and held-out-design results for the generative baselines. One Optuna-selected configuration per model is held fixed and retrained with seeds 0, 1, and 42. For VAE models, inference is deterministic at evaluation time ($\mathbf{z}{=}\mu$). Conv GAN uses standard adversarial training with a convolutional discriminator. Conv Diffusion uses its DDPM training-objective validation loss (`loss_total`) for checkpoint selection and DDIM sampling at inference, whereas the AE, VAE, and GAN models are selected by validation profile RMSE. On Model E, LSTM VAE has the lowest mean profile and peak-count errors among the generative models, while Transformer VAE has the lowest mean capacity and capacity-aligned errors. On the held-out designs, LSTM VAE has the lowest mean profile and capacity-aligned errors, while Conv GAN has the lowest mean capacity error. Conv VAE and Conv GAN have large peak-count variation across seeds.

Table 17: Generative model results. One Optuna-selected configuration per model is held fixed across three retraining seeds. Entries are mean ± sample standard deviation over seeds 0, 1, and 42. Top: primary Model-E test. Bottom: held-out designs A–D. **Bold**: lowest displayed mean per split and column.

Model	Profile RMSE ↓	Cap RMSE ↓	Cap-aligned ↓	Peak cnt ↓
Primary Model-E test				
Conv VAE	0.00868 ± 0.00207	0.03533 ± 0.01611	0.01458 ± 0.00590	2.364 ± 2.737
LSTM VAE	$\mathbf{0.00591 \pm 0.00012}$	0.02172 ± 0.00124	0.00845 ± 0.00022	$\mathbf{0.052 \pm 0.005}$
Transformer VAE	0.00597 ± 0.00011	$\mathbf{0.02020 \pm 0.00144}$	$\mathbf{0.00812 \pm 0.00032}$	0.067 ± 0.032
Conv GAN	0.00748 ± 0.00087	0.03565 ± 0.00918	0.01351 ± 0.00287	2.606 ± 4.077
Conv Diffusion	0.14939 ± 0.00029	0.02082 ± 0.00082	0.14234 ± 0.00043	9.155 ± 0.654
Held-out designs				
Conv VAE	0.01715 ± 0.00025	0.04130 ± 0.01496	0.02166 ± 0.00405	2.356 ± 2.783
LSTM VAE	$\mathbf{0.01291 \pm 0.00013}$	0.04002 ± 0.00887	$\mathbf{0.01534 \pm 0.00139}$	$\mathbf{0.000 \pm 0.000}^*$
Transformer VAE	0.01487 ± 0.00110	0.03511 ± 0.00461	0.01773 ± 0.00221	$\mathbf{0.000 \pm 0.000}^*$
Conv GAN	0.01669 ± 0.00306	$\mathbf{0.02793 \pm 0.00116}$	0.01864 ± 0.00314	2.624 ± 3.930
Conv Diffusion	0.14428 ± 0.00074	0.03549 ± 0.00794	0.14016 ± 0.00039	9.096 ± 0.579

* Zero peak-count error does not imply zero error on the other metrics.

B.12 Full Held-out-Design Results

Table 18 reports the full zero-shot held-out-design results. The ranking differs by metric and from the primary Model-E evaluation. Values are means and sample standard deviations over retraining seeds 0, 1, and 42. The held-out set is small and several design and protocol factors change together, so the transfer gap cannot be attributed to a single cause.

Table 18: Full held-out-design results (Models A–D, zero-shot). One Optuna-selected configuration per model is held fixed across three retraining seeds. Entries are mean ± sample standard deviation. **Bold**: lowest displayed mean per column.

Model	Profile RMSE ↓	Cap RMSE ↓	Cap-aligned ↓	Peak cnt ↓
Conv AE	0.01195 ± 0.00228	0.03878 ± 0.00994	$\mathbf{0.01533 \pm 0.00227}$	0.162 ± 0.073
MLP AE	$\mathbf{0.01074 \pm 0.00031}$	0.05883 ± 0.01889	0.01737 ± 0.00232	0.007 ± 0.006
LSTM AE	0.01314 ± 0.00364	0.07357 ± 0.05205	0.02211 ± 0.01156	0.108 ± 0.089
BiLSTM AE	0.01368 ± 0.00342	$\mathbf{0.02919 \pm 0.00032}$	0.01563 ± 0.00481	0.052 ± 0.091
Transformer AE	0.01270 ± 0.00093	0.03447 ± 0.00771	$\mathbf{0.01533 \pm 0.00123}$	$\mathbf{0.000 \pm 0.000}^*$
UNet AE	0.01365 ± 0.00215	0.03379 ± 0.00534	0.01633 ± 0.00364	0.523 ± 0.426
Conv+LSTM AE	0.01228 ± 0.00158	0.04313 ± 0.00847	0.01624 ± 0.00216	0.145 ± 0.045

* Zero peak-count error does not imply that the predicted curves or peak locations are correct.

B.12.1 Errors by Product Design

Table 19 separates the same 102 held-out targets into product designs A–D. The pooled result in Table 18 remains the primary held-out-design result because these groups are small and unequal in size. The checkpoints are unchanged and no adaptation is performed.

Table 19: Held-out-design errors separated by product design for the seven-architecture comparison. Entries are mean ± sample standard deviation over retraining seeds 0, 1, and 42.

Model	Profile RMSE↓	Cap RMSE↓	Cap-aligned↓	Peak cnt↓
Design A: 8 cells, 40 targets				
Conv AE	0.00981 ± 0.00269	0.03104 ± 0.00255	0.01083 ± 0.00391	0.221 ± 0.191
MLP AE	0.00862 ± 0.00084	0.04124 ± 0.00346	0.01219 ± 0.00184	0.017 ± 0.014
LSTM AE	0.00906 ± 0.00095	0.04854 ± 0.01360	0.01335 ± 0.00266	0.079 ± 0.081
BiLSTM AE	0.01410 ± 0.00469	0.02717 ± 0.00807	0.01720 ± 0.00701	0.079 ± 0.137
Transformer AE	0.01252 ± 0.00096	0.02913 ± 0.00284	0.01544 ± 0.00178	0.000 ± 0.000
UNet AE	0.01254 ± 0.00254	0.03889 ± 0.01034	0.01793 ± 0.00458	0.454 ± 0.369
Conv+LSTM AE	0.01241 ± 0.00090	0.03889 ± 0.01539	0.01453 ± 0.00159	0.154 ± 0.246
Design B: 7 cells, 15 targets				
Conv AE	0.00972 ± 0.00026	0.01620 ± 0.00417	0.00616 ± 0.00142	0.000 ± 0.000
MLP AE	0.00979 ± 0.00055	0.01900 ± 0.00431	0.00493 ± 0.00095	0.000 ± 0.000
LSTM AE	0.00814 ± 0.00058	0.01818 ± 0.00403	0.00769 ± 0.00170	0.000 ± 0.000
BiLSTM AE	0.01208 ± 0.00226	0.02455 ± 0.00682	0.01226 ± 0.00785	0.022 ± 0.038
Transformer AE	0.01098 ± 0.00097	0.02204 ± 0.00371	0.00848 ± 0.00584	0.000 ± 0.000
UNet AE	0.01016 ± 0.00081	0.01341 ± 0.00482	0.00976 ± 0.00376	0.022 ± 0.038
Conv+LSTM AE	0.00855 ± 0.00053	0.01557 ± 0.00253	0.00833 ± 0.00129	0.000 ± 0.000
Design C: 4 cells, 17 targets				
Conv AE	0.01363 ± 0.00430	0.05301 ± 0.01781	0.02337 ± 0.00890	0.147 ± 0.147
MLP AE	0.01237 ± 0.00046	0.07784 ± 0.02868	0.02651 ± 0.00586	0.000 ± 0.000
LSTM AE	0.01642 ± 0.00691	0.09763 ± 0.08634	0.03570 ± 0.02594	0.176 ± 0.281
BiLSTM AE	0.01326 ± 0.00334	0.03107 ± 0.01079	0.01559 ± 0.00178	0.059 ± 0.102
Transformer AE	0.01313 ± 0.00111	0.04284 ± 0.01176	0.01779 ± 0.00137	0.000 ± 0.000
UNet AE	0.01606 ± 0.00307	0.03319 ± 0.00195	0.01779 ± 0.00371	0.471 ± 0.328
Conv+LSTM AE	0.01294 ± 0.00345	0.05462 ± 0.01218	0.02153 ± 0.00532	0.333 ± 0.392
Design D: 5 cells, 30 targets				
Conv AE	0.01403 ± 0.00311	0.04401 ± 0.01976	0.02137 ± 0.00866	0.172 ± 0.160
MLP AE	0.01257 ± 0.00011	0.07588 ± 0.03023	0.02532 ± 0.00659	0.000 ± 0.000
LSTM AE	0.01660 ± 0.00641	0.09342 ± 0.07865	0.03330 ± 0.02204	0.161 ± 0.237
BiLSTM AE	0.01399 ± 0.00252	0.03013 ± 0.00763	0.01525 ± 0.00235	0.028 ± 0.048
Transformer AE	0.01347 ± 0.00108	0.03950 ± 0.01428	0.01722 ± 0.00085	0.000 ± 0.000
UNet AE	0.01493 ± 0.00253	0.03330 ± 0.00204	0.01665 ± 0.00284	0.894 ± 0.775
Conv+LSTM AE	0.01315 ± 0.00234	0.04688 ± 0.01706	0.01948 ± 0.00524	0.100 ± 0.087

The groups differ jointly in product design and protocol, so these errors do not isolate any single design factor. A zero peak-count error does not imply zero error on the other metrics.

Across all seven architectures, design B has the lowest capacity RMSE among the four held-out groups. Otherwise, the architecture ranking depends on both the group and the metric.

B.12.2 LSTM Seed Sensitivity

LSTM AE is stable on the primary Model-E test but has the largest held-out capacity variation: its capacity RMSE is 0.04450, 0.04254, and 0.13367 for seeds 0, 1, and 42. The seed-42 result drives the high held-out mean and variance. In that run, capacity RMSE reaches 0.19689 on C and 0.18361 on D, versus 0.06077 on A and 0.01692 on B. The failure is thus seed-dependent and concentrated in C and D, and these runs do not reveal its cause. Evaluating a single seed would have missed it.

B.13 Supplementary Reference-assisted Reconstruction

In this exploratory study, low-rate references from one or two support cells of each held-out design are available, and we measure how they affect reconstruction on the other cells of that design. It is reported separately from the 102-target zero-shot evaluation.

Support/query plan. For each held-out design, all two-cell combinations of its cells are shuffled with seed $2027 + d$ ($d = 0, \ldots, 3$ for A–D), and the first three combinations define three support selections. Within each pair, the member used for the one-cell condition is chosen at random. Both support candidates are removed from evaluation in every condition, including zero-cell and one-cell, so all three conditions of a selection share the same query cells. Support and query cells are disjoint in every split. Selections are not disjoint across repeats: in C, one cell appears in all three pairs, and in D the same cell serves as the one-cell support in two selections. Table 20 summarizes the resulting reference and query counts. A one-cell condition supplies all low-rate references of that cell. No

query reference is used for fitting, checkpoint selection, or hyperparameter selection. The query sets exclude both support candidates and therefore differ from the complete 102-target A–D evaluation.

Table 20: Support/query plan per held-out design. Ranges span the three support selections. A support cell contributes all of its low-rate references.

Design	Cells	References (1 cell)	References (2 cells)	Query cells	Query targets
A	8	6	9–12	6	28–31
B	7	2	4	5	11
C	4	4–5	7–8	2	9–10
D	5	4–7	7–10	3	20–23

Conv AE adaptation. The source model is the seed-42 Model-E Conv AE checkpoint from the three-seed architecture comparison, with its Optuna-selected hyperparameters. Before adaptation, this checkpoint reproduces its archived 102-target A–D results within 5×10^{-5} on profile, capacity, and charge-aligned RMSE; peak-count error differs by 0.0049, i.e., one of 102 targets crossing the discrete peak threshold across software builds. Each adapted condition reloads the same source weights and fine-tunes all parameters on the support references only, using the original training loss, AdamW ($\beta = (0.9, 0.95)$) with the source weight decay, learning rate 1.64×10^{-4} (one tenth of the Optuna-selected rate), batch size $\min(32, n_{\text{support}})$, gradient clipping at 1.0, and exactly 200 updates with torch seed 42. There is no early stopping; the final iterate is evaluated. The adaptation budget and learning-rate scale were fixed before examining the adaptation results. The Conv AE contains no normalization or dropout layers, and input voltages use the fixed operating-window normalization of Section 3.1, so no dataset-level statistics are re-estimated during Conv adaptation.

Ridge adaptation. Ridge uses the voltage-summary features of Appendix B.3 and keeps the regularization strength selected on Model-E validation. Before adaptation, it reproduces its archived A–D results within 5×10^{-5} on all reported metrics. Each adapted condition refits the feature standardizer and ridge coefficients on Model-E training records plus the support references. *Unweighted* appends the support references with unit weight. *Balanced* assigns weight 0.5 to each Model-E record and $0.5\, n_{\text{train}}/n_{\text{support}}$ to each support reference, so each group carries half of a total weight equal to n_{train}; the same weights are used when fitting the standardizer.

Support-mean control. For each split, the support mean predicts the element-wise average of the support references' normalized low-rate profiles and capacities for every query target, independent of its high-rate input. The zero-cell counterpart is the Model-E training-target mean on the same query cells; on all 102 A–D targets it reproduces the archived profile RMSE of 0.01774. No fitting is involved beyond averaging and query cells are never used. All methods use the same evaluation code. It measures the gain from knowing a target design's typical reference, which requires target-design low-rate measurements.

Average gains and their heterogeneity. Relative to matched zero-cell evaluation, Conv AE profile RMSE decreases by 49.5% with one cell and 59.0% with two; its capacity RMSE decreases by 47.2% with two. Balanced ridge reduces mean profile and capacity RMSE by 64.5% and 57.3% with two cells. However, an input-independent support mean already reaches profile RMSE 0.00431/0.00427 with one/two cells, compared with Conv AE's 0.00508/0.00412 and balanced ridge's 0.00536/0.00432 (Table 21). Much of the average profile gain thus comes from estimating the target design's typical reference.

With two cells, the more consistent aggregate advantage over the support mean appears in capacity (0.01446 for Conv and 0.01481 for ridge versus 0.02138) and charge-aligned error (0.00572 and 0.00617 versus 0.00832). Profile advantages differ by design: the two-cell Conv/support-mean profile RMSE ratio ranges from 0.36 on B to 1.27 on A. On B, Conv and ridge reach 0.00140 and 0.00167 versus 0.00383 for the support mean, whereas on C both have higher mean profile errors than that control. The similar aggregate profile errors thus average over designs where adaptation helps and designs where it does not. Relative to the support mean, Conv AE has lower profile, capacity, and

Table 21: Reference-assisted reconstruction on held-out designs. Entries are equally weighted means over 12 design–selection splits (4 designs × 3 support selections), each evaluated on query cells that exclude both support candidates. Zero-cell rows are recomputed on those query sets and are not comparable to the 102-target pooled errors of Table 5. Ridge rows use balanced weighting. Mean rows predict the average support reference for every query target; *Model-E training-target mean. Unweighted and per-design results are reported below.

Method	Support	Profile RMSE ↓	Cap RMSE ↓	Cap-aligned RMSE ↓	Peak count ↓
Conv AE	0 cells	0.01006	0.02742	0.01258	0.177
	1 cell	0.00508	0.02040	0.00747	0.218
	2 cells	0.00412	0.01446	0.00572	0.102
Ridge (V)	0 cells	0.01219	0.03465	0.01346	0.017
	1 cell	0.00536	0.01961	0.00756	0.084
	2 cells	0.00432	0.01481	0.00617	0.069
Mean (no input)	0 cells*	0.01721	0.02834	0.02337	0.000
	1 cell	0.00431	0.02216	0.00841	0.000
	2 cells	0.00427	0.02138	0.00832	0.000

charge-aligned RMSE in 7, 8, and 8 of 12 one-cell splits and in 8, 11, and 9 of 12 two-cell splits; balanced ridge has corresponding counts of 4, 7, and 7 and 7, 10, and 10.

Query-label-using constant diagnostic. For descriptive analysis only, the query-oracle constant predicts the mean of the query references themselves. It uses test labels and is not a baseline. Its profile, capacity, and charge-aligned RMSEs averaged over the 12 splits are 0.00384, 0.02021, and 0.00791. Per-design query-oracle profile RMSE ranges are A 0.00340–0.00360, B 0.00376–0.00389, C 0.00286–0.00483, and D 0.00378–0.00401. The analogous Model-E test-oracle profile RMSE is 0.00994, close to the training-target mean's 0.00998.

Under squared profile error, the query mean is the best predictor that gives every query the same profile; input-dependent predictors can do better, and it need not minimize the charge-aligned metric. The two-cell support mean comes close to this constant. Two-cell Conv AE has lower profile, capacity, and charge-aligned RMSE than the query constant in 8, 10, and 9 of 12 splits, and balanced ridge in 6, 10, and 8. The small variation around the query constant is a property of these cohorts; its physical cause is not identified.

Table 22: Input-independent means by held-out design: mean ± sample standard deviation over three support selections. The zero-cell row is the Model-E training-target mean on the same query cells.

Design	Support	Profile RMSE	Cap RMSE	Cap-aligned RMSE
A	0 cells	0.01813 ± 0.00004	0.03460 ± 0.00048	0.02481 ± 0.00015
	1 cell	0.00377 ± 0.00002	0.01999 ± 0.00092	0.00765 ± 0.00010
	2 cells	0.00374 ± 0.00035	0.01976 ± 0.00102	0.00766 ± 0.00057
B	0 cells	0.01458 ± 0.00002	0.02500 ± 0.00047	0.02297 ± 0.00010
	1 cell	0.00385 ± 0.00009	0.00970 ± 0.00069	0.00395 ± 0.00011
	2 cells	0.00383 ± 0.00008	0.00940 ± 0.00039	0.00391 ± 0.00004
C	0 cells	0.01742 ± 0.00032	0.03025 ± 0.00160	0.02224 ± 0.00080
	1 cell	0.00518 ± 0.00016	0.03065 ± 0.00259	0.01230 ± 0.00049
	2 cells	0.00494 ± 0.00013	0.03013 ± 0.00317	0.01183 ± 0.00048
D	0 cells	0.01871 ± 0.00012	0.02352 ± 0.00046	0.02345 ± 0.00027
	1 cell	0.00443 ± 0.00035	0.02830 ± 0.00203	0.00975 ± 0.00025
	2 cells	0.00457 ± 0.00026	0.02625 ± 0.00096	0.00988 ± 0.00048

Aggregation. Within each split (design, selection, condition), profile and capacity RMSE pool squared errors over all query targets and output coordinates before taking the square root. Charge-aligned and peak-count errors use the per-target aggregation specified in Appendix B.2. Tables 21 and 23 weight the 12 splits equally; their standard deviations therefore combine design heterogeneity with selection variation. Peak-position error is computed by the evaluation code but not reported, for the reason given in Section 4.3.

Table 23: All adaptation conditions: mean ± sample standard deviation over 12 design–selection splits.

Method	Support	Weighting	Profile RMSE ↓	Cap RMSE ↓	Cap-aligned RMSE ↓	Peak count ↓
Conv AE	0 cells	–	0.01006 ± 0.00133	0.02742 ± 0.00995	0.01258 ± 0.00306	0.177 ± 0.158
Conv AE	1 cell	fine-tune	0.00508 ± 0.00311	0.02040 ± 0.00936	0.00747 ± 0.00361	0.218 ± 0.178
Conv AE	2 cells	fine-tune	0.00412 ± 0.00309	0.01446 ± 0.00981	0.00572 ± 0.00377	0.102 ± 0.107
Ridge (V)	0 cells	–	0.01219 ± 0.00532	0.03465 ± 0.01512	0.01346 ± 0.00623	0.017 ± 0.025
Ridge (V)	1 cell	unweighted	0.00686 ± 0.00194	0.02524 ± 0.00950	0.00955 ± 0.00381	0.053 ± 0.096
Ridge (V)	1 cell	balanced	0.00536 ± 0.00272	0.01961 ± 0.00847	0.00756 ± 0.00348	0.084 ± 0.117
Ridge (V)	2 cells	unweighted	0.00579 ± 0.00179	0.02000 ± 0.00655	0.00809 ± 0.00293	0.057 ± 0.085
Ridge (V)	2 cells	balanced	0.00432 ± 0.00243	0.01481 ± 0.00573	0.00617 ± 0.00289	0.069 ± 0.099
Mean	0 cells	Model-E train	0.01721 ± 0.00166	0.02834 ± 0.00465	0.02337 ± 0.00104	0.000 ± 0.000
Mean	1 cell	support	0.00431 ± 0.00062	0.02216 ± 0.00870	0.00841 ± 0.00320	0.000 ± 0.000
Mean	2 cells	support	0.00427 ± 0.00056	0.02138 ± 0.00833	0.00832 ± 0.00310	0.000 ± 0.000

Table 24: Adaptation results by held-out design: mean ± sample standard deviation over three support selections. Ridge uses balanced weighting.

		Conv AE			Ridge (V), balanced		
Design	Support	Profile	Cap	Peak count	Profile	Cap	Peak count
A	0 cells	0.01054 ± 0.00005	0.03564 ± 0.00075	0.311 ± 0.018	0.00784 ± 0.00056	0.02964 ± 0.00071	0.000 ± 0.000
	1 cell	0.00676 ± 0.00185	0.02822 ± 0.00674	0.281 ± 0.058	0.00463 ± 0.00165	0.01832 ± 0.00439	0.012 ± 0.021
	2 cells	0.00478 ± 0.00167	0.02051 ± 0.01694	0.144 ± 0.030	0.00329 ± 0.00068	0.01313 ± 0.00426	0.006 ± 0.010
B	0 cells	0.00967 ± 0.00006	0.01183 ± 0.00066	0.000 ± 0.000	0.00654 ± 0.00005	0.02430 ± 0.00160	0.000 ± 0.000
	1 cell	0.00166 ± 0.00024	0.00838 ± 0.00207	0.030 ± 0.026	0.00198 ± 0.00043	0.01092 ± 0.00259	0.000 ± 0.000
	2 cells	0.00140 ± 0.00043	0.00638 ± 0.00060	0.045 ± 0.000	0.00167 ± 0.00011	0.00961 ± 0.00093	0.000 ± 0.000
C	0 cells	0.00858 ± 0.00162	0.03434 ± 0.00223	0.070 ± 0.061	0.01831 ± 0.00028	0.02609 ± 0.00781	0.037 ± 0.032
	1 cell	0.00723 ± 0.00383	0.02333 ± 0.01006	0.248 ± 0.227	0.00724 ± 0.00208	0.02022 ± 0.00262	0.130 ± 0.179
	2 cells	0.00583 ± 0.00511	0.01156 ± 0.00135	0.035 ± 0.031	0.00677 ± 0.00200	0.01615 ± 0.00359	0.111 ± 0.147
D	0 cells	0.01144 ± 0.00061	0.02788 ± 0.00018	0.326 ± 0.092	0.01607 ± 0.00014	0.05855 ± 0.00539	0.031 ± 0.027
	1 cell	0.00467 ± 0.00248	0.02166 ± 0.00257	0.314 ± 0.211	0.00761 ± 0.00159	0.02899 ± 0.01063	0.193 ± 0.060
	2 cells	0.00447 ± 0.00265	0.01941 ± 0.00616	0.183 ± 0.194	0.00557 ± 0.00215	0.02034 ± 0.00744	0.159 ± 0.065

Split-level consistency. Relative to the zero-cell value on the same query cells, Conv AE profile RMSE is lower in 11 of 12 splits with one cell and 11 of 12 with two; capacity RMSE is lower in 12 and 11 of 12. Conv peak-count error is higher than zero-cell in 7 of 12 one-cell splits and 3 of 12 two-cell splits. For every ridge variant, profile RMSE is lower in all 12 splits and capacity RMSE in 11 of 12, while peak-count error is higher in 4–6 of 12. Two cells give lower Conv profile and capacity RMSE than one cell in 9 of 12 splits. Selections reuse cells, so these counts are not independent replications.

Failures and limits of the supplementary analysis. Improvement is not uniform across outputs or support selections. In one C selection, two-cell Conv adaptation lowers capacity RMSE from 0.03681 to 0.01010 but raises profile RMSE from 0.00672 to 0.01174. The two-cell Conv profile RMSE across selections is 0.00140 ± 0.00043 on B and 0.00583 ± 0.00511 on C. In one A selection, two-cell capacity RMSE is 0.04003 versus 0.03650 before adaptation; in one D selection, two cells perform worse than one on profile, capacity, and charge-aligned errors. More support helps on average but not in every split.

Peak-count accuracy does not consistently follow curve and capacity improvements. Conv's aggregate count error changes from 0.177 to 0.218 and 0.102 with one and two cells, respectively; balanced ridge changes from 0.017 to 0.084 and 0.069. The support mean has zero count error on these query sets despite nonzero profile and capacity errors. As elsewhere, count agreement does not imply correct peak positions or amplitudes.

The Conv evaluation uses one source checkpoint and one adaptation seed. There is no input-dependent model trained only on target references, so the contribution of Model-E pretraining to adaptation remains unseparated. Support references come from each supplied cell's full observed history, and two-cell adaptation uses half of design C's four cells. The experiments do not determine how many

low-rate measurements a given diagnostic accuracy requires, or whether references from early life remain sufficient after further degradation.