Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Zero-Shot Transfer Fails: Parameter-Efficient Few-Shot Adaptation of a Selective State-Space Battery RUL Model to an Unseen LFP Fleet

Dikshant Dikshant

dikshant1103@gmail.com

Motilal Nehru National Institute of Technology https://orcid.org/0009-0006-9434-610X

Research Article

Keywords: remaining useful life, lithium iron phosphate, selective state-space models, transfer learning, parameter-efficient adaptation, conformal prediction, data leakage

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10713547/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Cross-chemistry zero-shot transfer has become a common claim in deep-learning battery prognostics: a model trained on one cell family is reported to generalise to another without retraining. The fragility of that claim, together with its repair, is characterised here on the HUST lithium iron phosphate (LFP) cycling dataset — 77 usable cylindrical cells and 127,179 prediction windows — using a selective state-space (Mamba) remaining-useful-life (RUL) model for which a zero-shot R^2 of 0.911 had previously been reported on the Oxford lithium nickel manganese cobalt oxide (NMC) pouch dataset. That earlier figure was subsequently found to be leakage-dependent (Section 1.2) and is quoted only as the claim whose limits this study documents. Applied zero-shot to HUST, the same checkpoint reaches a median per-cell R^2 of only 0.157, and 26 of 77 cells score below zero, which places them behind a cell-mean predictor; nominal 90% conformal intervals attain 41.2% empirical coverage. A parameter-efficient adaptation that freezes the three Mamba backbone blocks and updates 80,641 of 812,024 parameters (9.9%), trained on 61 cells and evaluated on 16 permanently held-out cells, raises the median per-cell R^2 to 0.610 under a strictly causal end-of-life (EOL) denominator that is available at inference. Because the adaptation normalises RUL by each cell's own EOL, converting a prediction back to cycles requires a denominator, and the choice of denominator materially changes the reported score. Three regimes are therefore ablated: an oracle EOL, which is not available at inference and yields 0.985; a constant fleet-median prior, which yields 0.886; and a causal EOL projected only from capacity observed up to the current cycle, which yields 0.610. The causal value is reported as the result of this work. The 0.985 figure is an upper bound that cannot be achieved in deployment and is stated only to bound the gap. Under the causal regime no held-out cell falls below $R^2 = 0.112$, against a zero-shot floor of -0.943 with 9 of 16 cells below zero. The oracle denominator is not a free win: supplied to the un-adapted model it lowers the median per-cell R^2 from -0.246 to -0.448, so the improvement is attributable to adaptation rather than to a leaked label, although adaptation and normalisation remain entangled in the headline number and must be reported separately. Conformal intervals recalibrated on the target fleet narrow from ± 236 to ± 145.3 cycles, while measured coverage moves from a dangerously optimistic 41.2% to a conservative 99.5%. Three contributions follow: a negative result that bounds when cross-chemistry zero-shot transfer may be claimed; a demonstration that roughly 10% of the parameters and 61 cells suffice to repair such a failure; and an explicit denominator ablation that is argued to belong with any RUL result obtained under per-cell target normalisation.

1. Introduction

1.1. Motivation and problem statement

A remaining-useful-life model is economically interesting only if it survives contact with cells it was not trained on. Battery fleets are heterogeneous by construction: a grid installation mixes production lots, a second-life pack mixes provenances, and a manufacturer's next cell generation is always a distribution shift. The field has responded with zero-shot transfer claims, in which a model trained on chemistry A is

reported to work on chemistry B without retraining. Such claims are attractive precisely because they promise to eliminate the per-fleet data collection that makes prognostics expensive to deploy.

The question addressed here is narrower than whether transfer is possible, and more practical: how much confidence does a single successful zero-shot transfer justify, and what does it cost to recover performance when that transfer does not hold?

1.2. Failure of zero-shot transfer

A claim of this kind has been made by the present authors. In prior work, a bidirectional selective state-space model with degradation-anchor cross-attention attained a zero-shot R^2 of 0.911 on the Oxford NMC pouch dataset under an Urban Artemis drive cycle, a 25× lifetime extrapolation beyond anything seen in training. That number was obtained with the v10-final checkpoint. A subsequent internal audit established that it is leakage-dependent: masking the cumulative-energy feature on the trained model collapses the Oxford result to $R^2 = -4.687$, and Oxford is absent from the corpus of the later v12 checkpoint, so the figure cannot be recovered from the clean model. It is quoted in this paper only as the prior claim whose fragility is being documented, and it should not be cited as a validated zero-shot result.

The present paper reports what happened when the same model met a second unseen fleet. On the HUST LFP dataset, under the same protocol, the checkpoint degrades to a median per-cell R^2 of 0.157. Twenty-six of seventy-seven cells score below zero, meaning that for roughly a third of the fleet the model is worse than a constant predictor.

Reporting this outcome, rather than quietly repairing it, is justified on two grounds. First, a single successful zero-shot transfer is weak evidence. The field currently rewards the positive instance and has no convention for reporting the negative one, which makes published transfer performance an optimistic sample. Second, the pattern of failure is itself informative: the model does not fail uniformly, but fails specifically on long-lived cells, which points towards a mechanism.

1.3. Research questions

Three questions organise the study:

RQ1. Does a checkpoint with a published successful zero-shot transfer generalise to a second unseen fleet of a different chemistry?

RQ2. If it does not, how much adaptation — measured in trainable parameters and in target-fleet cells — is required to repair the failure?

RQ3. How much of the reported post-adaptation performance is attributable to the model, and how much to the denominator used to convert per-cell-normalised predictions back into cycles?

1.4. Contributions

C1 — A bounded negative result. Zero-shot failure is quantified on a 77-cell LFP fleet for a model with a published successful zero-shot transfer, establishing that one success does not generalise: median per-cell R^2 0.157; 26 of 77 cells below zero; pooled RMSE 513.7 cycles (26.6% of the LFP RUL scale); nominal 90% conformal intervals achieving 41.2% coverage.

C2 — Parameter-efficient repair. Freezing the Mamba backbone and adapting 9.9% of the parameters on 61 cells lifts the median per-cell R^2 on 16 permanently held-out cells to 0.610 under a causal, inference-available denominator, with no cell below 0.112, against an oracle-regime ceiling of 0.985.

C3 — A denominator ablation, and an argument that it should be standard. Per-cell RUL normalisation requires an EOL denominator at inference. Performance is measured under oracle, fleet-median-prior and causal-projection denominators (0.985 / 0.886 / 0.610), and the un-adapted model is shown not to benefit from the oracle, which isolates the adaptation gain from the normalisation gain.

C4 — Recalibrated uncertainty. Conformal intervals recalibrated on the target fleet narrow from ± 236 to ± 145.3 cycles while measured coverage flips from a dangerously optimistic 41.2% to a conservative 99.5%.

1.5. Relationship to prior work

The backbone architecture, the leakage-audited feature pipeline and the online per-cell conformal protocol were established in earlier work by the same authors. No new architecture is proposed here. The contribution of this paper is the transfer study, the negative result, the adaptation protocol and the denominator ablation, none of which appear in that earlier work — and the first of which directly qualifies its central zero-shot claim.

2. Related Work

2.1. Data-driven battery RUL prediction

Data-driven prognostics rely heavily on a small number of public cycling corpora, among them the MIT–Stanford fast-charging release [2, 3], the Oxford degradation dataset [4], and the CALCE and NASA PCoE collections. Deep transfer learning has also been applied to personalised health-status prediction on operating cells [1]. A broader survey of data-driven RUL estimation is not attempted here; the present study is concerned with what happens to one trained model when the target fleet changes.

2.2. Cross-domain and cross-chemistry transfer

Transfer between datasets is usually posed either as domain adaptation with access to target data, or as zero-shot generalisation. Published cross-dataset results concentrate on CALCE, NASA PCoE and MIT–Stanford; reports of a single model spanning four chemistries remain rare, and reports of a transfer failure are rarer still. No prior study is known to the authors that publishes both a successful and an

unsuccessful zero-shot transfer for the same checkpoint, which is the comparison that makes the boundary of the claim visible.

2.3. Parameter-efficient adaptation

Freezing a pretrained backbone and training a small adapter is standard practice in language and vision [9, 10], and rests on the broader transfer-learning framework surveyed in [11]. In battery prognostics the pretrained-backbone regime is newer, because until recently no battery model had been trained on enough heterogeneous data to be worth freezing. The result reported here — 9.9% of the parameters, 61 cells — suggests that this regime has now arrived for battery RUL models.

2.4. Selective state-space models

Mamba-family models replace the quadratic cost of attention with an input-dependent linear-time recurrence [5], which suits per-cycle degradation sequences that run to thousands of steps. The architecture is treated as given in this work and is not modified.

2.5. Target normalisation and denominator-induced evaluation issues

The battery literature has a documented history of features that encode the label, cumulative discharge energy being the clearest case: under constant-current cycling it is a monotone function of cycle index and therefore acts as a cycle counter. The denominator ablation reported here concerns the same class of problem moved from the feature side to the target side. To the authors' knowledge this variant has not been examined in the battery prognostics setting.

2.6. Uncertainty quantification and conformal prediction

Conformal prediction supplies distribution-free predictive intervals under an exchangeability assumption [6, 7, 8]. It is used in this work in split form, with residuals obtained by leave-one-cell-out folds; the procedure is described in Section 4.7. What matters for the transfer question is not the guarantee itself but whether the calibration transfers with the model, which Section 5.5 addresses directly.

3. Dataset and Experimental Setting

3.1. HUST LFP dataset

The HUST battery dataset comprises cylindrical LFP cells cycled to end of life under varied discharge protocols. After the loading and quality pipeline described in Section 4.1, 77 **cells** yield usable sequences and **127,179** 30-cycle prediction windows.

3.2. Dataset characteristics

Table 1 summarises the properties of the fleet.

Table 1
Characteristics of the HUST LFP fleet used in this study.

Property	Value
Chemistry	$LiFePO_4$ (LFP), cylindrical
Cells (usable)	77
Windows	127,179
Cycle life range	1,042–2,159 cycles (16 held-out cells)
Median cycle life	1,616 cycles (held-out); 1,694 cycles (training fleet)
EOL definition	First cycle with capacity < 80% of initial
Temperature	Not logged per cycle; constant 30°C assumed per dataset description

Two properties of this fleet matter for transfer. The cells are long-lived relative to the training distribution, which was dominated by CALCE LCO cells living a few hundred cycles, so the model is required to extrapolate in absolute RUL. In addition, temperature is not measured, so the temperature input of the model is a constant. Any degradation signal that the backbone had learned to read from thermal variation is unavailable here — a plausible partial explanation for the zero-shot collapse, although it cannot be tested without thermal logs.

3.3. Source-model context

The source checkpoint is a selective state-space RUL model trained on a heterogeneous multi-chemistry corpus in which short-lived LCO cells predominate, and evaluated zero-shot on Oxford NMC as described in Section 1.2. Its feature normalisation statistics are stored with the checkpoint and are reused unchanged on the target fleet (Section 4.1), so no target-fleet information enters the input scaling.

3.4. Train/validation/test protocol

The 77 cells are shuffled once under a fixed seed and partitioned by cell: 61 cells for training and **16 cells permanently held out**. No window from a held-out cell appears in training, in validation, or in conformal calibration. Conformal residuals are obtained from 15 leave-one-cell-out folds over the 61 training cells only. The validation split used for early stopping is a 12% window-level split drawn from the training cells.

4. Methodology

4.1. Feature extraction

Nine base channels are extracted per discharge half-cycle: discharge capacity (Ah), charge time (s), mean voltage, voltage at 75% depth of discharge, discharge energy (Wh), temperature, capacity-fade slope, an internal-resistance proxy (maximum voltage minus mean voltage), and a chemistry code. Capacity is smoothed with a Savitzky–Golay filter (window 11, order 3), and a kinetic filter removes

cycles whose capacity exceeds that of the previous cycle by more than 5%, which suppresses rest-recovery artefacts. Four derived channels are then appended — capacity percentage relative to the first five cycles, per-cycle capacity delta, normalised cumulative energy, and internal-resistance delta — giving a **13-channel representation**.

The first nine channels are z-scored using the *source-model* statistics stored with the checkpoint. These statistics are never recomputed on the target fleet, so no target information enters the normalisation. Sequences are cut into windows of **30 cycles** with a non-overlapping stride, and each window is labelled with the RUL at its final cycle.

Note on cumulative energy. Channel 12 is normalised cumulative energy, a feature that the authors' own prior work identifies as a within-cell cycle counter. It is retained here only because the frozen backbone was trained with it, and because removing it would change the transfer question being asked. Since every comparison in this paper uses the identical feature set, the ablations remain internally valid; the absolute values, however, inherit whatever optimism that feature carries. A clean-feature replication is therefore the first item of future work (Section 7).

4.2. Target normalisation

The adaptation normalises the RUL of each cell by that cell's own EOL cycle,

$$y_{\text{norm}}(t) = RUL(t) \,/\, EOL_{\text{cell}} \quad (1)$$

so that the model predicts the fraction of life remaining, $y_{\text{norm}} \in [0, 1]$. The choice is deliberate. A single global scale compresses short-lived cells into a narrow band and destabilises training across a 2× spread in cycle life.

The cost of this choice is that recovering a prediction in cycles requires a denominator D,

$$\widehat{RUL}(t) = \hat{y}_{\text{norm}}(t) \times D \quad (2)$$

and D is a modelling decision that materially changes the reported score.

4.3. EOL denominator regimes

Three denominator regimes are evaluated; they are defined in Table 2.

Table 2
The three de-normalisation regimes, and whether each denominator can be obtained at inference on a live cell.

Regime	Denominator D	Available at inference?
A — Oracle	The cell's true EOL cycle	No; it is the quantity being predicted
B — Median prior	1,694 cycles (median life of the 61 training cells)	Yes; constant
C — Causal projection	EOL extrapolated by linear regression on the last ≤ 30 observed capacities, falling back to regime B when the fit is non-decreasing or implausible	Yes; strictly causal

Regime A is what the original evaluation used, and it is what produces the 0.985 figure. It is reported here for continuity and as an upper bound. Regime C is regarded as the deployable number.

4.4. Source model and parameter-efficient adaptation

The backbone is a selective state-space RUL model with three Mamba blocks followed by a degradation-anchor cross-attention head, **812,024 parameters** in total. Adaptation freezes the three Mamba blocks and updates only the input projection, the positional encoding, the cross-attention, the final normalisation and the MLP head. The resulting budget is given in Table 3.

Table 3
Parameter budget of the adaptation. Only the adapter set receives gradients.

Component	Parameters	Share
Frozen Mamba backbone	731,383	90.1%
Trainable adapter set	80,641	9.9%
Total	**812,024**	**100%**

4.5. Training procedure

Training uses AdamW with a learning rate of 2×10^{-5}, weight decay 10^{-4}, batch size 256, and a Huber loss with $\delta = 0.05$. The learning rate follows cosine annealing over a 40-epoch budget, with early stopping on a 12% validation split at patience 6, and gradient-norm clipping at 1.0. Each training cell is subsampled to at most 100 windows so that long-lived cells do not dominate the objective. Seed 42 is used throughout.

4.6. Causal post-processing

Predictions are post-processed with a running minimum (np.minimum.accumulate) that enforces a monotonically non-increasing RUL trajectory. The operation is causal, since it uses only predictions at or

before the current cycle, and it encodes a physical constraint rather than a fitted correction.

4.7. Conformal prediction

Conformal half-widths are obtained from 15 leave-one-cell-out folds over the 61 training cells, yielding 25,911 calibration residuals. Two settings are compared: the interval transferred with the source model, and an interval recalibrated on the target fleet. Coverage is measured on the 25,763 windows of the 16 held-out cells, which contribute to neither training nor calibration.

4.8. Evaluation metrics

Performance is reported per cell as R^2, RMSE and MAE, and summarised across cells by the median, since the distribution over cells is skewed by catastrophic cases. Pooled metrics computed over all windows are also given for the zero-shot setting. Counts of cells above $R^2 = 0.9$, above 0.8 and below 0 are reported alongside the medians, because for fleet monitoring the tail matters more than the average. Coverage is the empirical fraction of held-out windows whose true RUL falls inside the predicted interval.

5. Results

5.1. Zero-shot transfer performance

The unmodified source checkpoint was applied to all 77 HUST cells. Table 4 reports the outcome, and Fig. 1 shows the per-cell distribution behind it.

Table 4
Zero-shot transfer of the source checkpoint to the full HUST fleet.

Metric	Value
Cells / windows	77 / 127,179
Median per-cell R^2	**0.157**
Cells with $R^2 > 0.9$	3 / 77
Cells with $R^2 > 0.8$	9 / 77
Cells with $R^2 < 0$	**26 / 77**
Median per-cell RMSE	427.7 cycles
Median RMSE (% of 1,934-cycle scale)	22.1%
Pooled R^2	0.046
Pooled RMSE / MAE	513.7 / 396.3 cycles
Transferred 90% conformal coverage	**41.2%** (nominal 90%)

Three observations follow from Table 4. The median cell is barely better than a constant predictor; a third of the fleet is worse than one; and the transferred uncertainty intervals are badly miscalibrated, since 41.2% empirical coverage at a nominal 90% level means that an operator trusting the intervals would be surprised more than half the time. A model in this state is not merely inaccurate. It is unsafe to act on, because it gives no indication that it is wrong.

5.2. Parameter-efficient adaptation

Table 5 compares the adapted model on the 16 held-out cells (25,763 windows) with the zero-shot model on those same cells. Both columns of Table 5 use the oracle denominator; the deployable regimes are reported in Section 5.3.

Table 5
Matched comparison on the 16 permanently held-out cells: identical cells, windows and features, with only the weights differing.

Metric	Zero-shot (same 16 cells)	Adapted (oracle)
Median per-cell R^2	**−0.246**	**0.985**
Per-cell R^2 range	−0.943 … 0.886	0.939 … 0.997
Interquartile range	[− 0.587, 0.614]	[0.980, 0.992]
Cells with $R^2 > 0.9$	0 / 16	**16 / 16**
Cells with $R^2 < 0$	9 / 16	**0 / 16**
Median per-cell RMSE	509.6 cycles	59.4 cycles

The adapted column uses the oracle EOL denominator and is therefore not a deployable figure; deployable denominators are reported in Table 6.

Every held-out cell exceeds $R^2 = 0.9$ after adaptation, and the interquartile range contracts from 1.20 to 0.012. The adapted model is therefore not only better on average but considerably more consistent across cells, which is the property a fleet operator needs.

The 16 held-out cells are harder than the fleet as a whole: the zero-shot median is − 0.246 on this subset against 0.157 across all 77 cells. The split is a fixed seeded draw rather than a selection, and the comparison in Table 5 is deliberately matched, so the difference between the columns is attributable to the weights alone. The oracle denominator, however, is common to both columns of the adapted result and is unpacked next.

5.3. EOL denominator ablation

This ablation is the result considered most important here, and it qualifies Section 5.2. Table 6 reports all five combinations of model and denominator, and Fig. 2 shows the corresponding per-cell distributions.

Table 6
Denominator ablation on the 16 held-out cells.

#	Model	Denominator	Deployable	Median R^2	$R^2 > 0.9$	$R^2 > 0.8$	$R^2 < 0$	Median RMSE
1	Zero-shot	Global 1,934	Yes	−0.246	0	1	9	509.6
2	Zero-shot	Oracle EOL	**No**	**−0.448**	0	0	15	—
3	Adapted	Oracle EOL	**No**	**0.985**	16	16	0	59.4
4	Adapted	Median prior (1,694)	Yes	0.886	8	10	1	200.0
5	Adapted	Causal projection	**Yes**	**0.610**	0	4	**0**	252.6

Rows 2 and 3 use the oracle EOL denominator and are therefore not deployable. Row 5 is the principal deployable configuration. Median RMSE values are in cycles; cell counts are over the held-out cells.

Rows 2 and 3 should be read together. Supplying the oracle denominator to the un-adapted model makes it worse, from − 0.246 to − 0.448, with 15 of 16 cells below zero. The oracle is therefore not a label leak in the ordinary sense, since it hands no model the answer; it pays off only in combination with a model trained to emit a fraction of life, because only then is the product meaningful. The 0.985 figure is thus a genuine adaptation result multiplied by an unavailable constant, and the accurate statement is that adaptation and normalisation are entangled within it.

Rows 4 and 5 are what a deployed system would actually achieve, and both remain far above every zero-shot row. Row 5 has a property that neither row 1 nor row 4 possesses: no cell below zero, with a worst case of $R^2 = 0.112$. Row 4 attains a higher median but fails one cell badly (− 1.101), and its per-cell R^2 correlates at + 0.719 with true cell life, which is to say that the fleet-median prior works when a cell resembles the fleet median and degrades as the cell departs from it. Row 5 correlates at − 0.549 with cell life, indicating that the linear capacity extrapolation systematically under-projects EOL for the longest-lived cells — a specific and, in principle, fixable weakness (Section 7).

For deployment, row 5 is the row to report: median R^2 0.610, worst cell 0.112, and no catastrophic failures, against a zero-shot median of − 0.246 with a worst cell of − 0.943.

5.4. Per-cell behaviour

Adaptation is not uniformly beneficial, and the exceptions are reported here rather than aggregated away. Figure 4 shows per-cell R^2 before and after adaptation under the causal denominator.

Three of the 16 cells lose accuracy under adaptation

cell 4 − 3 (+ 0.89 → +0.58), cell 8 − 1 (+ 0.69 → +0.58) and cell 7 − 4 (+ 0.68 → +0.62). All three are cells on which zero-shot transfer already worked well. The gain is therefore concentrated on the cells that were failing, and the procedure trades a little accuracy on the already-good cells for the removal of every catastrophic one. For a fleet-monitoring application, in which the cost of a badly wrong cell dominates the benefit of a slightly better good one, that trade is judged worthwhile — but it is a trade rather than a free improvement, and a deployment whose value is concentrated in the best cells should weigh it differently.

5.5. Uncertainty recalibration

Conformal half-widths were computed from 15 leave-one-cell-out folds over the training cells (25,911 calibration residuals), with coverage measured on the 25,763 held-out windows. Table 7 reports the four settings evaluated, and Fig. 5 shows the corresponding half-widths and coverages.

Table 7
Conformal interval half-widths and measured coverage on the held-out windows.

Prediction setting	Interval	Nominal coverage	Measured coverage
Zero-shot	Transferred, ± 236 cycles	90%	**41.2%**
Adapted	Transferred, ± 236 cycles	90%	100%
Adapted, target-recalibrated	**± 145.3 cycles**	90%	**99.5%**
Adapted, target-recalibrated	± 174.7 cycles	95%	100%

All values in this table are computed under the oracle regime.

The failure mode changes completely between the first and third rows of Table 7. Applied to zero-shot predictions, the transferred interval covers 41.2% at a nominal 90% level, which is dangerously optimistic and the worst direction of error for an operational alarm. After adaptation and recalibration the interval is 38% narrower and covers 99.5%, so it has become conservative rather than optimistic.

That 99.5% is not the nominal level, and the over-coverage is stated explicitly. It is also expected: calibration residuals come from leave-one-cell-out folds whose models saw 60 cells, whereas the production model saw 61 and predicts the held-out cells more accurately, so the calibration residuals are pessimistic. The interval is consequently safe but wider than necessary, and a practitioner requiring a true 90% band should recalibrate against production-model residuals. All of the values in Table 7 are oracle-regime numbers; residuals under regime C are larger and would require their own calibration, which has not been performed (Section 7).

6. Discussion

6.1. Why zero-shot transfer fails

Three candidate mechanisms are offered below, ordered by decreasing confidence. They are labelled to separate what was measured from what is inferred and from what remains untested.

Mechanism 1 — absolute-RUL extrapolation. *Observation*: HUST cells live 1,042–2,159 cycles, whereas the training distribution of the source model was dominated by cells living a few hundred; per-cell R^2 for the zero-shot model is worst on the longest-lived cells (Fig. 3). *Interpretation*: under a global denominator the model is asked to emit values far outside its output range, and the trend across cell life is consistent with saturation of that output.

Mechanism 2 — missing thermal signal. *Observation*: HUST provides no per-cycle temperature, so that input channel is a constant. *Interpretation*: whatever the backbone had learned to read from thermal variation is simply absent, and a constant input may shift a learned representation off-manifold rather than merely removing information. *Hypothesis*: this cannot be tested with the present data, since no thermal logs exist for the target fleet.

Mechanism 3 — LFP voltage-plateau geometry. *Observation*: mean voltage and voltage at 75% depth of discharge are among the base channels, and LFP exhibits a flat discharge plateau. *Interpretation*: voltage-derived features are far less informative here than in the sloped LCO/NMC curves that dominated training, so those channels carry less degradation signal. *Hypothesis*: isolating this effect would require an ablation over voltage-derived channels on both chemistries, which has not been performed.

These mechanisms can be distinguished only partially with the present data, and no claim is made that the cause has been isolated. What can be stated is that adapting 9.9% of the parameters — notably including the input projection, which is where a shifted input manifold would be corrected — suffices to repair the failure, an outcome that is at least consistent with mechanisms 2 and 3.

6.2. What adaptation actually repairs

The matched comparison in Table 5 holds cells, windows and features fixed, so the improvement is attributable to the weights. Two features of the improvement are worth separating. The median rises, but the tail improves more: the interquartile range contracts from 1.20 to 0.012 under the oracle regime, and under the causal regime every cell that had been below zero is lifted above it. Adaptation therefore acts less as a uniform accuracy gain than as the removal of catastrophic per-cell failure, which is consistent with the observation in Section 5.4 that the three cells losing accuracy are precisely those on which zero-shot transfer already worked.

6.3. Oracle versus deployable performance

Per-cell target normalisation is increasingly common and is well motivated, since it stabilises training across a wide spread of cycle life. It also, however, silently converts the output of the model from cycles into a fraction of life, and converting back requires a quantity that is unknown at inference. A paper reporting per-cell-normalised RUL without stating how the denominator is obtained at test time is not reporting a deployable number, and the gap here is large: R^2 0.985 against 0.610 on identical predictions. The ablation of Section 5.3 is therefore recommended as routine practice. It is worth noting that the authors' own pipeline required exactly this audit.

6.4. Implications for battery prognostics

One zero-shot success is not evidence of zero-shot capability. The same checkpoint that transfers to Oxford NMC at a reported R^2 of 0.911 — itself later found to be leakage-dependent — fails on HUST LFP at 0.157. Had testing stopped at Oxford, a general claim would have been published that the present paper falsifies. Cross-chemistry transfer claims should therefore be accompanied by at least two unseen target fleets, with failures reported alongside successes; otherwise published transfer performance is a filtered sample.

Few-shot adaptation beats zero-shot transfer at an affordable cost. Sixty-one cells and 9.9% of the parameters converted an unusable model into a fleet-consistent one. For a commercial telemetry deployment this reframes the economics: instead of promising chemistry-agnostic prediction out of the box, a vendor promises a short target-fleet calibration campaign, which is a schedule item rather than a research risk.

6.5. Interpretation of uncertainty results

The uncertainty results carry a practical asymmetry. Before adaptation the intervals are optimistic, covering 41.2% at a nominal 90% level, which is the failure direction that would silently suppress alarms. After target-fleet recalibration they are conservative at 99.5%, which costs width but does not hide error. The direction of miscalibration therefore matters as much as its magnitude, and the transferability of a calibration should be checked separately from the transferability of a model — the two do not move together, as the second row of Table 7 shows, where the transferred interval applied to adapted predictions over-covers at 100%.

7. Limitations and threats to validity

Cumulative energy retained. Channel 12 is a feature that the authors' prior work identifies as a within-cell cycle counter (Section 4.1). It is held constant across every comparison, so the ablations are internally valid, but the absolute values inherit its optimism. A clean-feature replication is the highest-priority follow-up and should be completed before the absolute numbers reported here are quoted as fleet performance.

Single seed, single split. All results use seed 42 and a single 61/16 partition. The held-out set is harder than the fleet mean (zero-shot − 0.246 against 0.157), which suggests that split variance is not

negligible. Repeated splits with confidence intervals are needed before the reported medians are treated as stable estimates.

Conformal intervals calibrated under the oracle regime, and over-covering. The half-widths in Section 5.5 are not valid for the deployable regimes, since residuals under regime C would be wider. The recalibrated band also measures 99.5% rather than its nominal 90%, so it is conservative rather than tight; a true 90% band requires production-model residuals.

Reproducibility defect, disclosed. The stored feature statistics of the released checkpoint show that it was trained *before* an Arrhenius temperature encoding was added to the training script, so the script in the repository no longer reproduces the released weights. All results in this paper were regenerated by matching the checkpoint's own stored statistics, and the regenerated oracle-regime numbers reproduce the original evaluation exactly (median R^2 0.985, 16/16 cells above 0.9, median RMSE 59.4 cycles). The script must be reverted, or the model retrained, before release.

Causal EOL projection is crude. Linear regression on the last ≤ 30 capacities under-projects long-lived cells ($r = -0.549$ with cell life). A nonlinear or model-based EOL projector would likely close much of the 0.610 → 0.886 gap and is the clearest route to improving the deployable number.

No temperature measurements. The target fleet provides no per-cycle temperature, so the conclusions drawn about thermal mechanisms in Section 6.1 remain conjecture and cannot be tested with these data.

8. Conclusion

A selective state-space RUL model with a published successful zero-shot transfer fails on a second unseen fleet: the median per-cell R^2 is 0.157 across 77 HUST LFP cells, 26 cells are worse than a constant predictor, and nominal 90% intervals cover 41.2%. Adapting 9.9% of its parameters on 61 cells repairs the failure, giving a median per-cell R^2 of 0.985 on 16 permanently held-out cells under the oracle de-normalisation used by the original evaluation, and 0.610 with every cell above $R^2 = 0.112$ under a strictly causal denominator available at inference. Conformal intervals recalibrated on the target fleet narrow from ± 236 to ± 145.3 cycles while measured coverage moves from a dangerously optimistic 41.2% to a conservative 99.5%.

The practical conclusion is that zero-shot cross-chemistry transfer should be treated as a hypothesis to be tested per fleet rather than as a property of a model, and that a short target-fleet calibration is a reasonable price to pay for reliability. The methodological conclusion is that per-cell target normalisation carries a denominator obligation, and that reporting the oracle number alone overstates deployable performance by a wide margin — here, 0.985 against 0.610.

Reproducibility, Data and Code Availability

Table 8 lists the scripts and result artefacts behind each reported number.

Table 8
Artefacts supporting the reported results.

Artefact	Path
Zero-shot evaluation	validate_hust_v10.py
Adaptation and LOOCV conformal	finetune_hust_loocv.py
Denominator ablation (Section 5.3)	reeval_deployable.py + _ft_lib.py
Per-cell zero-shot results	results/hust_cell_results.csv
Per-window adapted predictions	results/test_results.csv
Per-cell ablation table	results/reeval_per_cell.csv
Ablation summary	results/reeval_summary.json
Conformal half-widths	results/conformal_hust.json
Figures (4)	make_paper4_figures.py → figures/

Seed 42; a single NVIDIA GPU; PyTorch. The disclosed script/checkpoint divergence is described in Section 7.

References

1. Ma G, Xu S, Jiang B, Cheng C, Yang X, Shen Y, Yang T, Huang Y, Ding H, Yuan Y (2022) Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. Energy Environ Sci 15:4083–4094. 10.1039/D2EE01676A
2. Severson KA, Attia PM, Jin N, Perkins N, Jiang B, Yang Z, Chen MH, Aykol M, Herring PK, Fraggedakis D, Bazant MZ, Harris SJ, Chueh WC, Braatz R (2019) D. Data-driven prediction of battery cycle life before capacity degradation. Nat Energy 4:383–391. 10.1038/s41560-019-0356-8
3. Attia PM, Grover A, Jin N, Severson KA, Markov TM, Liao Y-H, Chen MH, Cheong B, Perkins N, Yang Z, Herring PK, Aykol M, Harris SJ, Braatz RD, Ermon S, Chueh WC (2020) Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature 578:397–402. 10.1038/s41586-020-1994-5
4. Birkl CR (2017) Oxford Battery Degradation Dataset 1. University of Oxford Research Archive 10.5287/bodleian:KO2kdmYGg
5. Gu A, Dao T, Mamba (2023) Linear-Time Sequence Modeling with Selective State Spaces. *arXiv:2312.00752*
6. Vovk V, Gammerman A, Shafer G (2005) Algorithmic Learning in a Random World. Springer. 10.1007/b106715
7. Angelopoulos AN, Bates S (2021) A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv:2107.07511*

8. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution-free predictive inference for regression. J Am Stat Assoc 113(523):1094–1111. 10.1080/01621459.2017.1307116
9. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations* arXiv:2106.09685
10. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S (2019) Parameter-Efficient Transfer Learning for NLP. *International Conference on Machine Learning* arXiv:1902.00751
11. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. 10.1109/TKDE.2009.191
12. Pedregosa F et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

Figures

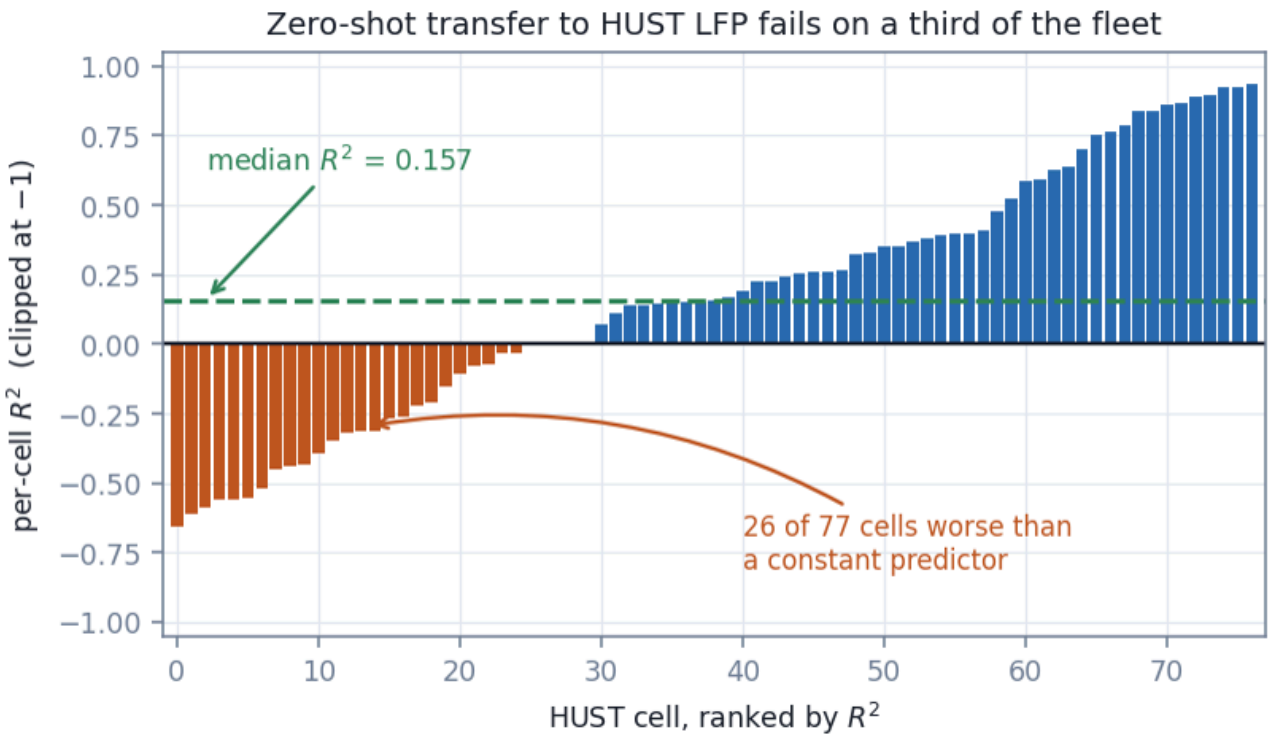


Figure 1

Per-cell R² for all 77 cells, ranked. The 26 cells below zero are the result of interest: for roughly a third of the fleet, the transferred model performs worse than a constant predictor.

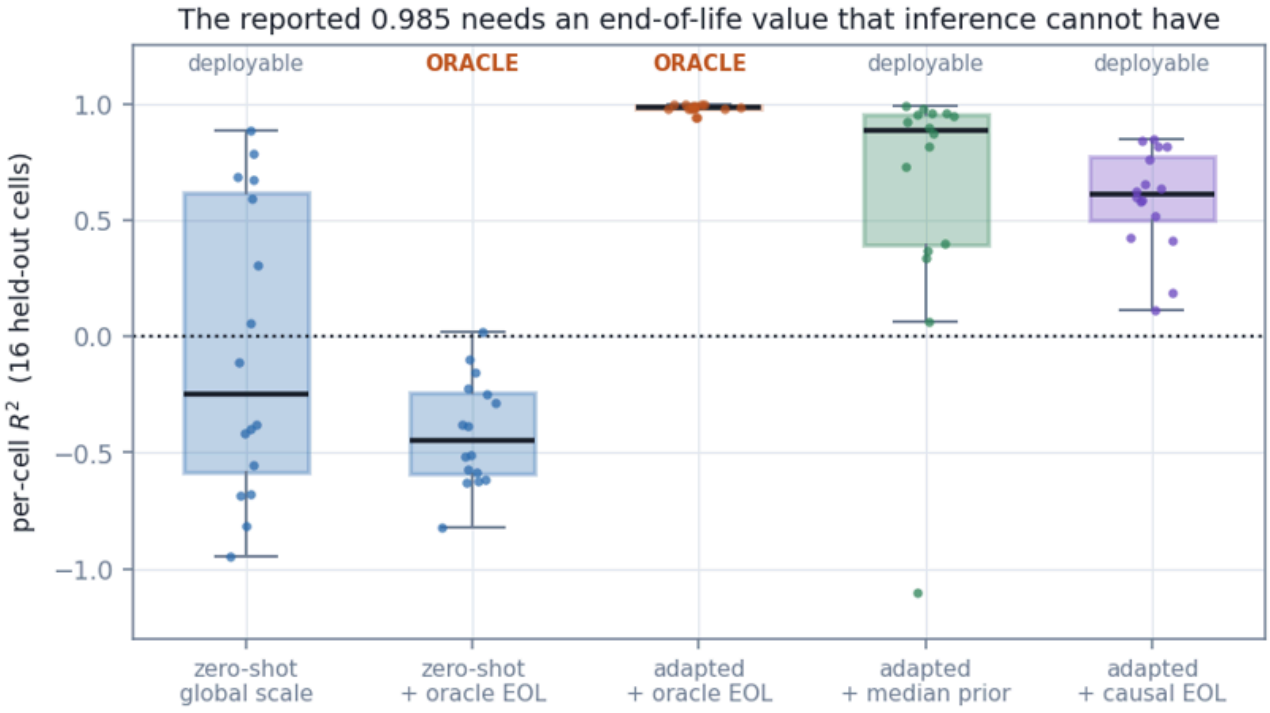


Figure 2

The five rows of Table 6 as per-cell distributions, with the two oracle regimes flagged.

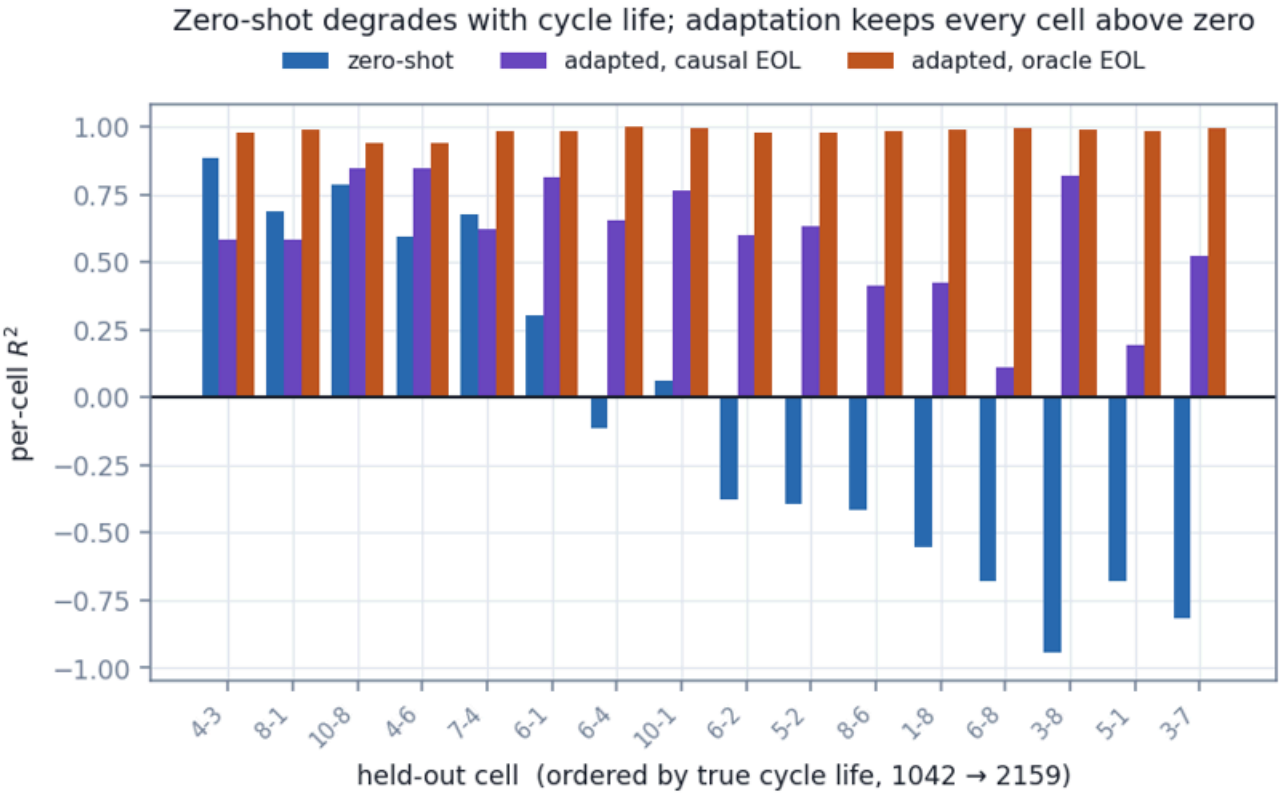


Figure 3

The same 16 cells ordered by true cycle life, showing zero-shot performance collapsing as cell life grows while both adapted variants remain above zero.

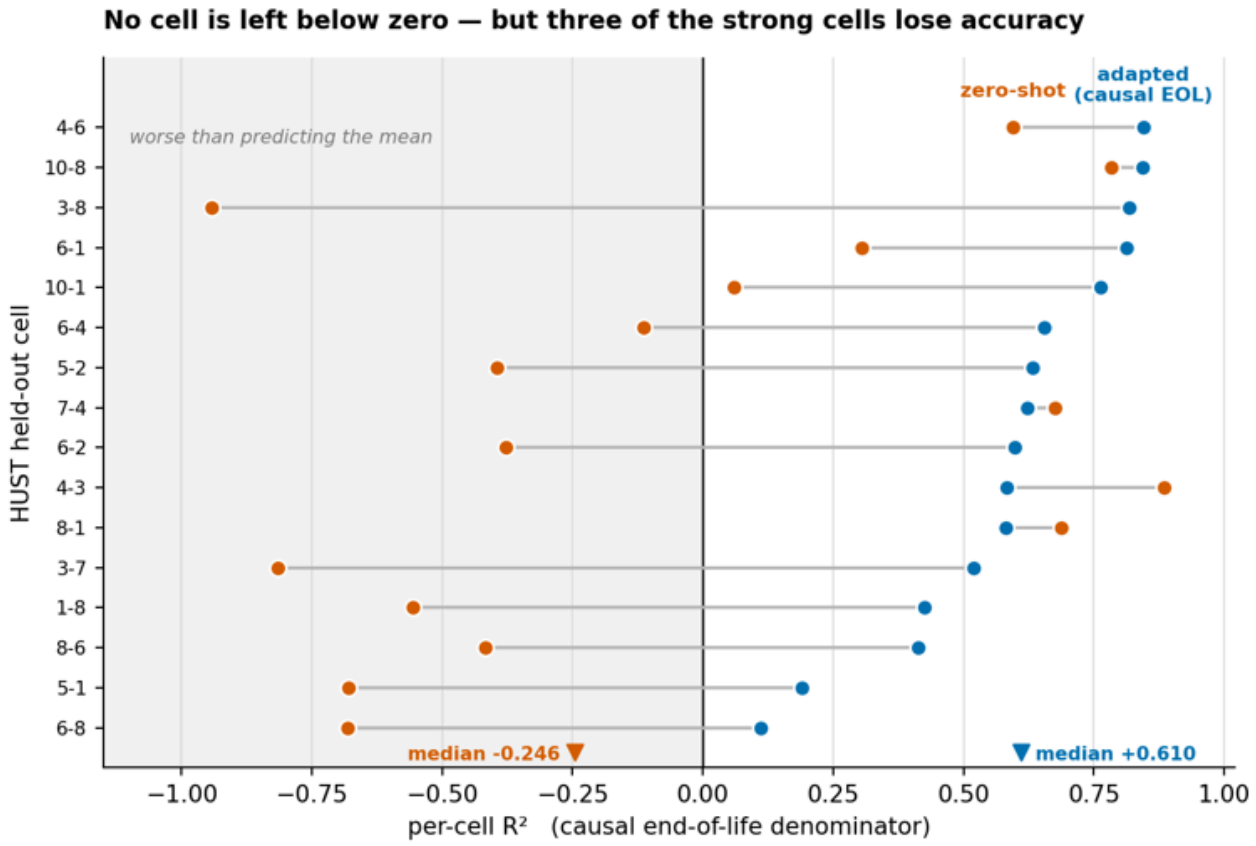


Figure 4

Per-cell R^2 before and after adaptation under the causal EOL denominator, the regime reported as the result of this work. The zero-shot median is −0.246, with 9 of 16 cells below zero; the adapted median is +0.610, with no cell below zero and a minimum of +0.112.

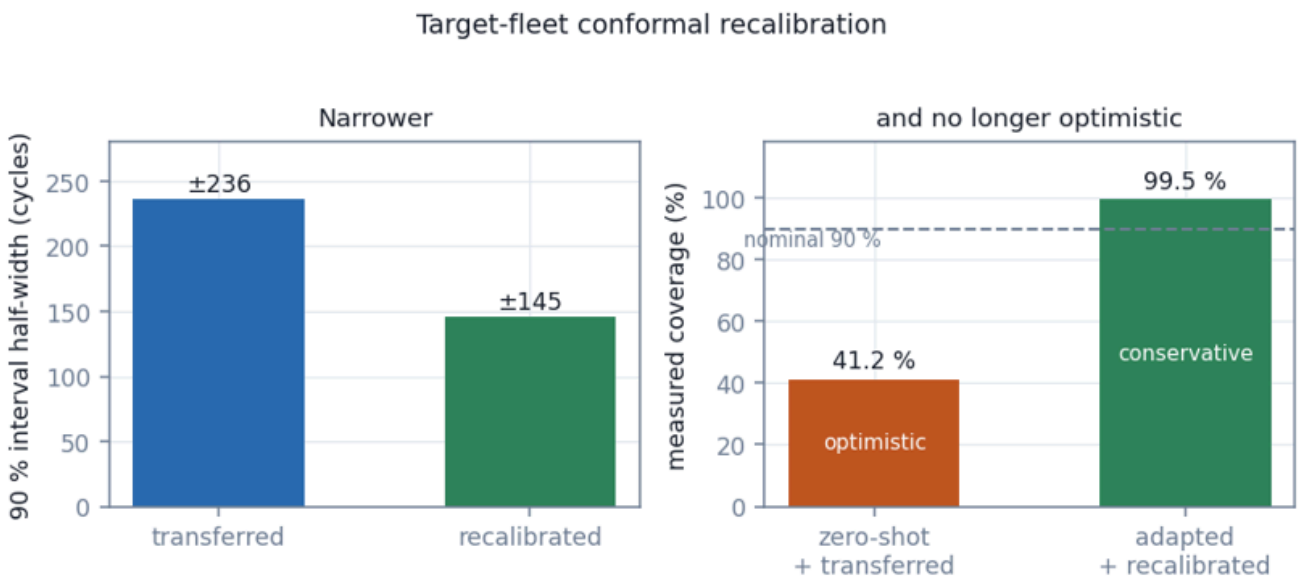


Figure 5

Half-width and measured coverage, before and after target-fleet recalibration.