Energy Conversion and Management 369 (2026) 122086

Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Research paper

Continuous ageing trajectory representations across heterogeneous battery datasets: Knee analysis and protocol-held-out early-life prediction

Agnieszka Pregowska *, Stefan Marynowicz

Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawinskiego 5B, Warsaw, 02-106, Poland

ARTICLE INFO

Keywords:
Lithium-ion batteries
Battery ageing
Implicit neural representations
State of health
Remaining useful life
Knee-point detection
Cross-protocol validation
Reliability analysis

ABSTRACT

Accurate assessment of lithium-ion battery ageing is complicated by cell-to-cell variability, heterogeneous cycling protocols, and limited transferability of data-driven predictors. This study develops a continuous ageing-trajectory framework in which coordinate-based multilayer perceptron (MLP) and sinusoidal representation network (SIREN) models represent state-of-health and voltage–capacity trajectories as differentiable functions of normalized trajectory coordinates. The framework combines continuous reconstruction, derivative-based knee descriptors, equivalent-full-cycle (EFC) harmonization, protocol-held-out early-life prediction, perturbation analysis, and conformal uncertainty quantification. Early-life prediction was evaluated in four protocol-held-out folds, repeated for three random seeds, using the first $N \in \{5, 10, 20\}$ available degradation observations rather than assuming that one observation represented one physical cycle. The pooled association between robust consensus knee location and the first crossing of the 80% SOH threshold was strong ($n = 172$, Spearman $\rho_S = 0.834$, 95% bootstrap CI [0.738, 0.907], Holm-adjusted $p = 3.36 \times 10^{-45}$). Raw adapted cINR predictions exhibited substantial architecture- and window-dependent bias. A leakage-safe bias-corrected MLP-protocol blend achieved MAEs of 113.57, 107.79, and 112.82 EFC for $N = 5$, 10, and 20, respectively, compared with 116.66, 115.57, and 125.02 EFC for the protocol-median baseline. These numerical reductions were not statistically significant after multiplicity correction, although the blend significantly improved over raw MLP cINR predictions for $N = 5$ and $N = 10$. Fold-external conformal intervals at the nominal 90% level attained empirical coverages of 90.3%, 91.0%, and 95.5%, but remained wide, reflecting substantial inter-cell uncertainty. Knee detection was comparatively stable under missing observations but sensitive to aggressive downsampling and measurement noise. The results support continuous trajectory representations as useful differentiable descriptors and components of calibrated hybrid predictors, but do not establish a uniform predictive advantage over a strong protocol-informed baseline.

1. Introduction

The rapid deployment of lithium-ion batteries in electric vehicles, stationary storage and consumer electronics has intensified the need for reliable and interpretable ageing models [1,2]. Despite extensive research, predicting battery degradation remains challenging due to the interplay of multiple mechanisms, including loss of cyclable lithium, impedance growth and thermal effects, which evolve nonlinearly over time [3,4]. Accurate estimation of state of health (SOH) and remaining useful life (RUL) is therefore essential for safe operation and lifecycle optimization [5].

Existing approaches to battery ageing modelling can be broadly divided into physics-based and data-driven methods [6–9]. Physics-based models, such as equivalent circuit and electrochemical models, provide physical interpretability but are often limited by parameter identifiability and computational complexity [10,11]. Data-driven methods, including machine learning and deep learning models, offer flexibility but typically rely on discretized time series or hand-crafted features, which limits transferability across datasets and operating conditions [12–16].

A persistent limitation of battery-prognostics studies is not the absence of accurate model families, but the lack of a common representation and validation protocol that permits degradation trajectories collected under different sampling densities and stress conditions to be compared without treating every recorded observation as an equivalent physical cycle. Feature-based models compress each trajectory into a fixed set of hand-crafted statistics, sequence models learn directly from discretized histories, Gaussian-process models provide continuous

* Corresponding author.
E-mail address: aprego@ippt.pan.pl (A. Pregowska).

https://doi.org/10.1016/j.enconman.2026.122086
Received 14 April 2026; Received in revised form 5 August 2026; Accepted 19 August 2026
Available online 31 August 2026

probabilistic interpolation, and physics-informed models introduce explicit electrochemical or dynamical constraints. These approaches are valuable but differ substantially in their sampling requirements, computational assumptions, and ability to provide stable derivative-based descriptors [2,17,18].

The present study does not claim architectural novelty for the MLP or SIREN networks themselves. Its methodological contribution is the integration of five elements within one leakage-controlled framework: differentiable coordinate-based trajectory representations; EFC-based interpretation of heterogeneous observation indices; curvature- and change-point-based knee extraction; protocol-held-out early-life prediction repeated across folds and seeds; and explicit sensitivity, censoring, statistical-significance, and conformal-coverage analyses. In contrast to a purely feature-based pipeline, the continuous representation permits the trajectory and its derivatives to be evaluated at arbitrary coordinates. In contrast to a physics-informed neural network, it does not require specification of a particular electrochemical differential equation.

Throughout this study, "interpretable" refers to the outputs derived from the continuous representation-including slope, curvature, plateau extent, knee location, and EFC-based lifetime-rather than to full transparency of the neural-network parameters. These descriptors have a direct geometrical or degradation-related meaning, whereas the learned network weights and latent coordinates are not interpreted individually.

The study therefore evaluates whether continuous trajectory representations can support consistent ageing analysis across heterogeneous sources and whether cINR-derived information adds predictive value beyond strong protocol-informed baselines. The contribution is assessed conservatively: numerical improvement is distinguished from statistically supported improvement, and robustness is evaluated explicitly rather than inferred from performance on clean data.

This study progresses SDG 7 (Affordable and Clean Energy) and SDG 9 (Industry, Innovation and Infrastructure) by improving lithium-ion battery health prediction and lifetime assessment, enabling more reliable, efficient, and sustainable energy storage systems through interpretable, transferable battery ageing analytics [19,20].

The main contributions of this study are fourfold. First, we establish a common coordinate-continuous representation for analysing heterogeneous battery-ageing trajectories on an EFC scale. Second, we extract geometrically interpretable descriptors, including slope, curvature, plateau extent, and consensus knee location, and explicitly evaluate their sensitivity to missing observations, noise, and reduced sampling resolution. Third, we assess, rather than assume, whether cINR-derived information provides incremental prognostic value beyond a strong protocol-informed baseline under protocol-held-out validation. Fourth, we integrate trajectory reconstruction, lifetime analysis, leakage-safe post-processing, paired statistical comparisons, and conformal uncertainty quantification within one auditable framework. The contribution therefore lies in the unified evaluation and interpretation protocol rather than in claiming that a particular neural architecture universally outperforms conventional battery-prognostics models.

2. Materials and methods

2.1. Implicit Neural Representations

Implicit Neural Representations constitute a class of continuous, coordinate-based neural functions that have recently emerged as a powerful alternative to classical interpolation and regression models [21–25]. Instead of predicting over a discrete grid or relying on predefined basis expansions, an INR directly parameterizes an unknown signal, such as a voltage curve, ageing trajectory, or multidimensional electrochemical response, as a differentiable function

$$g_\theta : \mathbb{R}^d \to \mathbb{R}^c, \quad \theta \in \mathbb{R}^p \tag{1}$$

where the input coordinate $\mathbf{x} \in \mathbb{R}^d$ encodes the model domain (e.g., time, current, temperature, cycle index, or auxiliary metadata), and the output $\mathbf{y} \in \mathbb{R}^c$ corresponds to the reconstructed electrochemical response (voltage, derivative dV/dQ, capacity-normalized features, or latent space parameters). The function g_θ is implemented as a multilayer perceptron whose trainable parameters θ implicitly encode the full continuous signal [26].

In the analyses reported here, a fixed input dimensionality was used within each representation task. For capacity-ageing trajectories, the principal coordinate was the normalized EFC or source trajectory coordinate, and the response was normalized SOH. For voltage–capacity reconstruction, the coordinate was normalized discharged capacity and the response was voltage. Dataset-specific metadata such as protocol group, charge rate, discharge rate, and depth of discharge were retained for grouping, fold construction, or conditional priors; missing metadata were not replaced by additional variable-dimensional neural inputs. Consequently, the same coordinate-to-response mapping was applied within a task even when the source datasets differed in the amount of available auxiliary information.

To further formalize the continuous representation, we consider the INR model as a function approximator trained by minimizing a data-fitting objective over observed electrochemical measurements:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left\| g_\theta(\mathbf{x}_i) - y_i \right\|^2, \tag{2}$$

where $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denotes the set of coordinate–response pairs obtained from battery cycling data. This formulation allows the model to learn a continuous mapping from the input domain (e.g., capacity or cycle index) to the measured signal (e.g., voltage or normalized capacity).

A key advantage of this representation is that differential operators can be applied directly to the learned function. For example, the first and second derivatives with respect to capacity Q,

$$\frac{\partial V}{\partial Q}, \qquad \frac{\partial^2 V}{\partial Q^2}, \tag{3}$$

can be computed analytically via automatic differentiation. The first derivative $\mathrm{d}V/\mathrm{d}Q$ provides a differential-voltage descriptor, whereas incremental-capacity analysis is based on the inverse derivative $\mathrm{d}Q/\mathrm{d}V$. Together with the second derivative, these quantities can be used to identify changes in local slope, curvature, and inflection behaviour along the voltage–capacity trajectory.

Unlike grid-based interpolants, polynomial expansions, or kernel regressors, INRs operate in continuous coordinate space and do not require predefined sampling locations. Their capacity to represent high-frequency and nonlinear behaviour depends not on the discretization, but on the network architecture (e.g., sinusoidal activations in SIREN networks or random Fourier feature mappings). This makes them well suited for electrochemical systems, where voltage curves exhibit sharp transitions (plateau boundaries, phase-change regions, knee points) and strongly nonlinear ageing dynamics. INRs support dense evaluation at arbitrary coordinates and direct computation of first- and second-order derivatives through automatic differentiation. These properties facilitate reconstruction, interpolation, and extraction of trajectory-shape descriptors without applying finite differences directly to irregularly sampled measurements. In the present study, however, the learned network weights and latent coordinates are not interpreted as uniquely identifiable electrochemical parameters or as direct fingerprints of specific degradation mechanisms. Similarly, cross-dataset invariance is treated as an empirical question evaluated through held-out protocols and perturbation experiments rather than as an inherent property of coordinate-based learning. From a scientific machine learning perspective, INRs complement physics-informed neural networks and neural ODEs by providing flexible, continuous surrogates for strongly nonlinear electrochemical dynamics, while remaining agnostic to a specific partial differential equation model [18].

While the INR model is data-driven, partial physical consistency is ensured through the structure of the training data and the choice of target variables. Physical consistency was treated differently for reconstruction and lifetime extrapolation. The general voltage–capacity and trajectory-reconstruction INRs were not subject to a hard monotonicity constraint and could therefore reproduce short-term capacity-recovery effects or measurement fluctuations.

For early-life lifetime extrapolation, the dedicated prognostic decoder represented SOH as a latent-dependent intercept minus the integral of a strictly positive degradation-rate field. At each quadrature point, the normalized ageing coordinate was concatenated with the cell-specific latent vector. The unconstrained rate-network output was transformed using $\mathrm{softplus}(\cdot) + 10^{-6}$ and subsequently integrated using trapezoidal quadrature during training and cumulative trapezoidal integration during dense trajectory generation. Consequently, the decoded SOH trajectory was non-increasing with respect to the ageing coordinate by construction. This monotonicity constraint was applied only to the prognostic branch.

Predicted EOL was additionally constrained not to precede the last observed landmark, $\widehat{\mathrm{EOL}} \geq a_N$, thereby preventing negative predicted RUL. This output constraint was separate from the monotone trajectory construction.

The extracted trajectory descriptors, including slope, curvature and plateau extent, have a direct geometrical interpretation and may be associated with changes in degradation behaviour. Such changes can be influenced by processes including impedance growth, loss of active material, loss of cyclable lithium, and transport limitations. The descriptors are nevertheless phenomenological and are not treated as uniquely identifiable signatures of individual electrochemical mechanisms.

From a functional perspective, the INR representation can be interpreted as a neural field that approximates an underlying smooth degradation manifold. Assuming sufficient regularity, the learned function g_θ belongs to a Sobolev space $W^{k,2}$, which guarantees the existence of weak derivatives up to order k. This functional interpretation permits weak or automatic derivatives to be defined for the learned representation, but it does not guarantee their numerical stability under measurement noise or sparse sampling. Because second-order derivatives can amplify local perturbations, the stability of curvature- and knee-related descriptors was evaluated explicitly in the sensitivity experiments.

The Sobolev-space interpretation is relevant operationally because knee and curvature descriptors depend on first- and second-order derivatives. A differentiable surrogate permits these quantities to be evaluated on a dense coordinate grid without applying finite differences directly to irregular raw samples. This does not guarantee immunity to noise; derivative stability was therefore assessed empirically through the perturbation experiments reported in Section 7.4.

3. State-of-health and remaining-useful-life models

The early-life task was formulated as prediction of cell EOL in equivalent full cycles. The primary EOL coordinate was estimated by linear interpolation between the last observation at or above the operational SOH threshold and the first observation below that threshold. Let (a_{j-1}, s_{j-1}) denote the last observation satisfying $s_{j-1} \geq 0.80$ and (a_j, s_j) the first observation satisfying $s_j < 0.80$. The interpolated EOL coordinate was defined as

$$\mathrm{EOL}_{0.80} = a_{j-1} + \frac{0.80 - s_{j-1}}{s_j - s_{j-1}} \left(a_j - a_{j-1}\right). \tag{4}$$

No interpolation beyond the two observations bracketing the 80% SOH threshold was performed.

For a cell observed up to landmark a_N, remaining useful life was defined as

$$\mathrm{RUL}_N = \mathrm{EOL}_{0.80} - a_N. \tag{5}$$

The input window comprised the first $N \in \{5, 10, 20\}$ available degradation observations. A cell–window pair was retained in the prospective prediction benchmark only when the operational threshold had not been crossed within the input prefix:

$$a_N < \mathrm{EOL}_{0.80}, \qquad \mathrm{SOH}(a_j) \geq 0.80, \quad j = 1, \ldots, N. \tag{6}$$

Pairs for which the 80% SOH threshold had already been reached within the first N observations were excluded before prediction, bias correction, blending, statistical comparison, and conformal calibration. The term "observation" is used deliberately because one source observation did not represent the same amount of electrochemical throughput across datasets.

Cell–window pairs for which the 80% SOH threshold had already been crossed within the input prefix were excluded before model fitting, baseline construction, bias estimation, blending, statistical comparison, and conformal calibration. Consequently, the target always corresponded to a future EOL coordinate relative to the last available input observation. The observation-window definition was not interpreted as a common physical ageing landmark across datasets. In particular, $N = 5$, 10, or 20 specifies the amount of source information made available to the model, but not an equal number of physical cycles or an equal fraction of cell lifetime. For this reason, the EFC coordinate of the last input observation, a_N, was retained for every prediction, and its distribution was reported separately by validation domain and observation window. Cross-window comparisons were interpreted jointly with differences in both cohort composition and the physical EFC extent of the available prefix.

Two related INR uses were distinguished. First, reconstruction INRs represented an observed trajectory as a continuous coordinate-to-response function and supplied slope, curvature, area, and knee-related descriptors. Second, the conditional prognostic INR (cINR) used a decoder trained on source cells and an adapted cell-level latent code to extrapolate the SOH trajectory beyond the observed prefix. For a held-out cell, decoder weights were frozen and only the cell-specific representation was adapted using its first N observations. The evaluation target and all later observations of that cell remained unavailable during adaptation.

The principal neural architectures were a three-hidden-layer MLP and a SIREN. Both decoders used three hidden layers of width 64 and an eight-dimensional cell-specific latent representation. The SIREN used sinusoidal activations with $\omega_0 = 15$. Final decoder training used 1800 optimization steps and a batch size of 128. Models were trained in four protocol-held-out folds and repeated for seeds 2026, 2027, and 2028. For each architecture, window, and fold, model fitting, protocol priors, adaptation parameters, bias estimates, and combination weights were derived without using the held-out fold. Mean aggregation across the three seeds was used as the primary summary, while median aggregation was retained as a sensitivity analysis.

The decoder parameters were optimized for 1800 steps using Adam with a learning rate of 2×10^{-4}, a batch size of 128, and no explicit weight-decay term. During adaptation to a held-out cell, all decoder parameters remained frozen and only the cell-specific latent representation was optimized. For the MLP decoder, latent adaptation used 300 optimization steps, a learning rate of 2×10^{-3}, a latent-prior weight of 0.02, and three independent restarts. The seed-dependent SIREN adaptation settings are reported in Table 2, whereas the complete predefined candidate space used for source-domain configuration selection is reported in Table 1. For both architectures, the restart yielding the lowest adaptation objective was retained.

A protocol-median predictor was used as a strong reference because protocol group contains substantial information about expected lifetime. For each outer-test cell, the protocol-median EOL estimate was calculated exclusively from outer-training cells assigned to the same predefined protocol category. No EOL value from the corresponding outer-test fold contributed to the protocol median. When the outer-training partition contained no cell from the required protocol category, the median EOL of the complete outer-training partition was used as the prespecified fallback. The protocol categories and fallback rule were

Table 1
Source-domain hyperparameter search space for held-out-cell latent adaptation. Candidate configurations were evaluated using source-domain or inner-fold information only; held-out target-cell EOL values were not used for selection. "Selected" indicates that the candidate was retained in at least one final source-only selection run.

Architecture	Candidate	Prior weight	Steps	Learning rate	Restarts	Selected
MLP	A1	0.300	100	0.0030	2	No
MLP	A2	0.200	150	0.0030	3	Yes
MLP	A3	0.100	250	0.0020	3	No
MLP	A4	0.050	300	0.0020	3	No
MLP	A5	0.020	300	0.0020	3	Yes
SIREN	A1	0.300	100	0.0030	2	Yes
SIREN	A2	0.200	150	0.0030	3	No
SIREN	A3	0.100	250	0.0020	3	No
SIREN	A4	0.050	300	0.0020	3	Yes
SIREN	A5	0.020	300	0.0020	3	Yes

fixed before evaluation. Let $\widehat{E}_{\mathrm{cINR}}$ denote the seed-aggregated cINR prediction, $\widehat{b}_{-f}$ the prediction bias estimated using only training folds other than outer fold f, and $\widehat{E}_{\mathrm{prot}}$ the corresponding protocol-median prediction. The leakage-safe hybrid prediction was

$$\widetilde{E}_{\mathrm{cINR}} = \widehat{E}_{\mathrm{cINR}} - \widehat{b}_{-f}, \tag{7}$$

$$\widehat{E}_{\mathrm{hyb}} = w_{-f}\widetilde{E}_{\mathrm{cINR}} + (1 - w_{-f})\widehat{E}_{\mathrm{prot}}, \qquad 0 \le w_{-f} \le 1, \tag{8}$$

where w_{-f} was selected using only the outer-training units. Final predictions were constrained by $\widehat{E}_{\mathrm{hyb}} \ge a_N$, which prevented negative predicted RUL.

3.1. Data splitting and evaluation protocol

Units, rather than individual observations, formed the independent evaluation entities. Protocol groups were partitioned into four outer folds so that all observations from a held-out cell and its evaluation protocol remained outside model fitting and post-processing. Each configuration was repeated for three fixed seeds. The resulting outer-fold predictions were retained at cell level and used to calculate MAE, median absolute error, RMSE, bias, negative-RUL frequency, paired significance tests, and conformal calibration.

The primary inferential family comprised six predefined cell-level comparisons. The mean-aggregated hybrid cINR–MLP predictor was compared with ProtocolMedianEOL and with the raw cINR–MLP predictor separately at each of the three observation windows, $N = 5$, 10, and 20. Comparisons were based on paired absolute EOL errors for cells available to both methods.

Two-sided Wilcoxon signed-rank and Monte Carlo sign-flip tests were calculated for each comparison. For the sign-flip test, the null distribution of the mean paired difference in absolute error was approximated using 100,000 random sign assignments with a fixed reproducibility seed. The six raw Wilcoxon p-values were adjusted jointly using the Holm procedure. Independently, the six raw sign-flip p-values were adjusted across the same predefined six-comparison family. Tests involving other architectures, aggregation rules, perturbation conditions, or post-processing variants were not included in this primary multiplicity family.

Bootstrap confidence intervals for correlation estimates were calculated using 10,000 cell-level resamples. For domain-specific Spearman correlations based on fewer than ten cells, statistical significance was evaluated using an exact two-sided permutation test rather than the large-sample asymptotic approximation. For the NASA-randomized scope, all $5! = 120$ permutations of the EOL ranks relative to the observed knee-location ranks were enumerated. The exact two-sided p-value was defined as the proportion of permutations yielding an absolute Spearman correlation at least as large as the observed value. The resulting raw p-value was included together with the pooled and ISU–ILCC correlation p-values in the three-scope Holm adjustment.

Full-trajectory knee descriptors were not included in the strict early-life predictors; only information available within the first N observations was permitted.

3.2. Implementation details and training setup

The final cINR experiments used four protocol-held-out folds, three fixed random seeds, and observation windows $N \in \{5, 10, 20\}$. Hyperparameter and latent-adaptation configurations were selected using source-domain validation data only. No held-out-cell EOL values were used for decoder fitting, configuration selection, latent adaptation, bias correction, seed aggregation, protocol blending, or conformal calibration. Continuous coordinates were normalized within the corresponding training representation, whereas all reported lifetime predictions and errors were transformed back to EFC.

MLPs are parametric universal function approximators and were used here as continuous neural-field parameterizations. Local-linear extrapolation and ProtocolMedianEOL constituted the principal structurally different reference methods. Random Forest was used only in the exploratory feature-attribution analysis, whereas Fourier-feature, RBF, spline, and related models were evaluated separately in the continuous reconstruction benchmark.

The finite source-domain search space used for held-out-cell latent adaptation is reported in Table 1. Candidate configurations were evaluated without access to held-out target-cell EOL values.

3.3. Exploratory Random-Forest attribution

Tree-based models have been applied to battery SOH estimation because they can represent nonlinear feature interactions while retaining comparatively straightforward feature-attribution mechanisms [12–14]. SHAP-based interpretation has also been used to quantify the contribution of measured battery-health indicators to data-driven SOH estimates [27]. An auxiliary retrospective attribution analysis was conducted using 21 leakage-safe descriptors calculated from the first $N = 5$ available degradation observations. Variables containing EOL, RUL, knee, censoring, analysis-time, identifiers, or other full-trajectory information were excluded from the predictor set. Missing descriptor values were imputed using medians calculated from the corresponding training fold.

A Random-Forest regressor with 500 trees, square-root feature sampling, and a minimum leaf size of two was evaluated using five-fold out-of-fold prediction. TreeSHAP values were calculated only for cells held out from the corresponding model-fitting fold and were aggregated as mean absolute SHAP contributions expressed in EFC.

The descriptor table contained 244 cells with five available observations. Seven of these cells had already reached the operational 80% SOH threshold within the five-observation prefix and therefore did not satisfy the prospective eligibility criterion used in the strict early-life benchmark. The Random-Forest analysis is consequently treated as a retrospective descriptor-attribution analysis and is not included in the prospective protocol-held-out performance comparison reported in Table 4.

No reliable protocol or validation-domain grouping variable was retained in the merged descriptor table. Consequently, the Random-Forest analysis used internal K-fold validation and was not interpreted

Table 2
Implemented architecture, decoder-training, held-out-cell adaptation, and trajectory-generation settings for the prognostic cINR models. All configuration choices were made without access to held-out-cell EOL values.

Component	Implemented setting
MLP decoder	Three hidden layers, width 64 per layer, and an eight-dimensional cell-specific latent vector.
SIREN decoder	Three hidden layers, width 64 per layer, and an eight-dimensional cell-specific latent vector; sinusoidal activations with $\omega_0 = 15$.
Decoder optimization	Adam optimizer; learning rate 2×10^{-4}; 1800 optimization steps; batch size 128; no explicit weight-decay term.
Coordinate conditioning	Normalized ageing coordinate concatenated with the cell-specific latent vector.
Monotone prognostic output	Latent-dependent intercept minus the trapezoidal integral of the positive degradation-rate field $\mathrm{softplus}(r) + 10^{-6}$.
Numerical integration	16 quadrature points during decoder training and 32 quadrature points during prediction.
MLP held-out-cell adaptation	300 optimization steps; learning rate 2×10^{-3}; latent-prior weight 0.02; three independent restarts for seeds 2026, 2027, and 2028.
SIREN held-out-cell adaptation	For seeds 2026 and 2028: 100 optimization steps, learning rate 3×10^{-3}, latent-prior weight 0.30, and two restarts. For seed 2027: 300 optimization steps, learning rate 2×10^{-3}, latent-prior weight 0.05, and three restarts.
Adaptation selection	Decoder parameters remained frozen and only the cell-specific latent vector was optimized. The restart with the lowest adaptation objective was retained.
Dense trajectory query	5000-point coordinate grid with an extrapolation-extension factor of 1.25.
EOL output constraint	The predicted EOL coordinate was constrained by $\widehat{\mathrm{EOL}} \geq a_N$, preventing negative predicted RUL.
Seed aggregation	Arithmetic mean across the three seeds as the primary aggregation; median aggregation as a sensitivity analysis.
Repetition and validation	Four protocol-held-out folds; seeds 2026, 2027, and 2028; observation windows $N \in \{5, 10, 20\}$.

as a cross-domain or cross-protocol experiment. For the same reason, no Random-Forest train–test transfer heatmap was constructed.

4. Uncertainty quantification

MC dropout has been interpreted as an approximate Bayesian approach to estimating model uncertainty, whereas predictive uncertainty may contain distinct epistemic and aleatoric components [28,29]. Uncertainty-aware RUL prediction and post-hoc regression calibration have also been investigated as mechanisms for improving the reliability of predictive risk reporting [30,31]. Uncertainty was evaluated using two complementary procedures: MC-dropout as a diagnostic of stochastic model dispersion and fold-external conformal intervals as the principal empirical calibration method. These procedures address different questions. MC-dropout quantifies variability induced by stochastic network realizations, whereas conformal calibration uses held-out residuals to quantify empirical prediction uncertainty.

4.1. MC-dropout diagnostic

The MC-dropout analysis was an auxiliary retrospective diagnostic of stochastic prediction dispersion. It used the same 21 leakage-safe descriptors calculated from the first $N = 5$ available degradation observations for the full 244-cell descriptor cohort. Seven of these cells had already reached the operational 80% SOH threshold within the input prefix. The MC-dropout results are therefore not part of the strict prospective early-life benchmark, which retained 237 cells at $N = 5$.

A multilayer perceptron with hidden widths of 64, 64, and 32 units and a dropout probability of 0.10 was evaluated using internal four-fold out-of-fold prediction. Missing features were imputed using training-fold medians and standardized using training-fold statistics. The analysis was repeated for seeds 2026, 2027, and 2028. For each trained model and held-out cell, 200 stochastic forward passes were performed with dropout active, yielding 600 stochastic predictions per cell across the three seeds.

The resulting dispersion estimates characterize the behaviour of the auxiliary descriptor model on the retrospective 244-cell cohort. They were not used as prediction intervals for the prospective cINR benchmark or for its fold-external conformal calibration.

$$\hat{y}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} f_{\theta_m}(\mathbf{x}), \qquad \hat{\sigma}^2_{\mathrm{MC}}(\mathbf{x}) = \frac{1}{M-1}\sum_{m=1}^{M}\left[f_{\theta_m}(\mathbf{x}) - \hat{y}(\mathbf{x})\right]^2. \tag{9}$$

The association between MC-dropout standard deviation and absolute prediction error was evaluated using Spearman rank correlation. Gaussian mean ± 1.645SD and mean ± 1.960SD bands were examined only as diagnostics of stochastic dispersion. They were not treated as calibrated prediction intervals.

4.2. Fold-external conformal intervals

Fold-external conformal intervals were constructed from absolute residuals obtained without using the corresponding outer test fold. For nominal miscoverage α, the interval was

$$C_\alpha(\mathbf{x}) = \left[\max\left(a_N, \hat{E}(\mathbf{x}) - q_{1-\alpha}\right), \hat{E}(\mathbf{x}) + q_{1-\alpha}\right], \tag{10}$$

where $q_{1-\alpha}$ is the finite-sample conformal quantile of the fold-external calibration residuals. For a calibration set containing m absolute residuals $r_1, \ldots, r_m$, the finite-sample quantile was defined as the

$$k_\alpha = \lceil (m+1)(1-\alpha) \rceil \tag{11}$$

order statistic of the sorted residuals, with k_α truncated to m when necessary. Coverage uncertainty was summarized using exact binomial 95% confidence intervals calculated from the number of covered outer-test targets and the corresponding evaluation sample size. Empirical coverage and median interval width were evaluated for nominal 90% and 95% levels.

These intervals quantify empirical residual uncertainty within the present protocol-held-out study design. They should not be interpreted as external-validation guarantees for new chemistries, fleets, or battery-management systems.

4.3. Reliability analysis

To place the INR-based SOH and RUL predictions into a reliability-engineering context, we performed a classical lifetime analysis of the empirical EOL distribution on the EFC scale. For each validation domain, the lifetime coordinate t denotes equivalent full cycles rather than the source observation index. We estimated the empirical Kaplan–Meier survival function $S(t)$ and fitted descriptive two-parameter Weibull and lognormal lifetime models commonly used for lithium-ion batteries [32,33]. Related battery-RUL studies have also formulated lifetime prediction using Weibull accelerated-failure-time regression and deep survival models [34,35]. The Weibull model,

$$S_{\text{Weibull}}(t) = \exp\left[-\left(\frac{t}{\lambda}\right)^{k}\right], \tag{12}$$

with shape k and scale λ, is widely employed to characterize wear-out and defect-dominated failure modes, while the lognormal distribution is often adequate when the degradation process can be modelled as a multiplicative accumulation of random effects [32,36].

The corresponding hazard function, which characterizes the instantaneous failure rate, is given by

$$h(t) = \frac{f(t)}{S(t)} = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}. \tag{13}$$

For $k > 1$, the hazard rate increases with time, indicating a wear-out failure regime consistent with cumulative degradation processes in lithium-ion cells.

Weibull, lognormal, and Kaplan–Meier models were fitted separately for each validation domain and for the pooled sample. The primary 80% SOH analysis contained 271 observed events and no right-censored cells: 12 CALCE cells, 251 ISU-ILCC cells, and eight NASA-randomized cells. Fitting was performed by maximum likelihood using the `lifelines` implementation.

For the Weibull distribution, the reported parameters were the shape k and scale λ. For the lognormal distribution, the location μ, logarithmic scale σ, and equivalent natural-scale parameter $\exp(\mu)$ were retained. Model fit was summarized using Akaike's information criterion (AIC). AIC values were compared only between distributions fitted to the same analysis scope. Because the validation domains represented heterogeneous protocols and cell populations, pooled parameters were interpreted descriptively rather than as estimates of a universal battery-lifetime distribution.

4.4. Cross-domain lifetime comparison

Cross-domain differences in observed $\text{EOL}_{0.80}$ were evaluated on the EFC scale. A classical one-way ANOVA was reported as a conventional parametric comparison. Because the domain sample sizes were strongly unbalanced, homogeneity of variance was evaluated using the median-centred Brown–Forsythe test. Welch one-way ANOVA was treated as the primary parametric test when variances were heterogeneous. The Kruskal–Wallis test provided a rank-based non-parametric comparison.

Omnibus effect sizes were reported as η^2 and ω^2 for the classical ANOVA decomposition and ϵ^2 for the Kruskal–Wallis test. Pairwise domain comparisons used Welch two-sample tests and Mann–Whitney tests. Multiplicity across the three pairwise comparisons was controlled using the Holm procedure.

Pairwise magnitude was quantified using Cohen's d, bias-corrected Hedges' g, and Cliff's δ, signed as the first domain minus the second domain. Ninety-five-percent confidence intervals for Cohen's d and Cliff's δ were calculated using 5000 independent cell-level bootstrap resamples. The analyses describe differences between heterogeneous laboratory populations and do not identify a causal effect of dataset membership.

5. Input data

The study is based on a set of complementary, publicly available databases on lithium-ion cell ageing and material properties [37]; see Table 3. Time-series data from charge–discharge cycles comes primarily from the NASA Li-ion Battery Aging Datasets [38], recorded at the NASA Ames PCoE forecasting facility, where cylindrical Li-ion cells were operated in three profiles (charge, discharge, EIS) at various temperatures and current loads until their defined EOL. This data was supplemented with CALCE battery data (University of Maryland), including continuous and partial cycles, storage tests, dynamic driving profiles, open-circuit voltage and impedance measurements for cells of various formats (cylindrical, prismatic, pouch) and chemistries (including LCO, LFP, NMC), allowing for the study of degradation under various operating conditions [39]. The third dataset is the ISU-ILCC Battery Aging Dataset [40,41], which contains long-term tests of 251 18 650 cells designed to analyse the effect of three load factors, charge rate, discharge rate, and depth of discharge, on capacity loss. Additionally, the database "A database of battery materials auto-generated using ChemDataExtractor" (figshare 11888115) was used for an auxiliary materials-level analysis. It contains more than 290,000 literature-mined experimental records describing relationships between electrode-material composition and properties such as specific capacity, voltage, conductivity, Coulombic efficiency, and energy [42].

The ChemDataExtractor battery-materials database was used exclusively for an auxiliary materials-level analysis. Records describing gravimetric capacity or voltage were retained when they could be assigned unambiguously to one of five major cathode-material families: LCO, LFP, NMC, NCA, or LMO. Gravimetric-capacity values reported in mAh g^{-1} or the numerically equivalent Ah kg^{-1} were harmonized to mAh g^{-1}. Voltage values were expressed in volts. Records with unresolved units, values outside the accepted physical ranges, explicit correctness warnings, or ambiguous material assignments were excluded.

Because individual publications frequently reported multiple measurements, the primary descriptive analysis was balanced at the publication level. A median value was first calculated within each DOI–material-family–property combination, and the resulting DOI-level observations were subsequently used to calculate material-family medians and interquartile ranges. The ChemDataExtractor records were not linked to individual CALCE, NASA, or ISU–ILCC cells and were not used for SOH, knee, EOL, RUL, or model-training analyses.

Additional public datasets focused on rapid SOH assessment and second-life grading (e.g., Warwick rapid SOH dataset [43]) are not used in this work but represent promising candidates for future cross-dataset validation.

6. Preprocessing and feature extraction

Raw experiments were reduced to ordered degradation observations containing capacity, SOH, the source ageing coordinate, and the available operating metadata. Duplicate source coordinates were aggregated using the median for numerical variables. State of health was defined as

$$\text{SOH}_j = \frac{Q_j}{Q_0}, \tag{14}$$

where Q_0 denotes the initial reference capacity used for the corresponding cell.

The primary end-of-life coordinate was defined as the linearly interpolated first crossing of the operational SOH = 0.80 threshold. Let (a_{j-1}, s_{j-1}) denote the last observation satisfying $s_{j-1} \geq 0.80$ and (a_j, s_j) the first subsequent observation satisfying $s_j < 0.80$. The primary EOL coordinate was calculated as

$$\text{EOL}_{0.80} = a_{j-1} + \frac{0.80 - s_{j-1}}{s_j - s_{j-1}}\left(a_j - a_{j-1}\right). \tag{15}$$

Table 3
Overview of the public datasets used in this study. The table summarizes the source, order-of-magnitude data volume and the role of each dataset in the proposed analysis.

Dataset	Brief description
NASA Li-ion Battery Aging [38]	Source: NASA Prognostics Center of Excellence (NASA Open Data). Cylindrical 18650 Li-ion cells cycled under several charge/discharge profiles and ambient temperatures until EOL (80% of initial capacity). High-frequency time series of voltage, current, temperature and time; per-cycle discharge capacity. Used for unified SOH/RUL definition, knee-point analysis and INR-based curve reconstruction.
CALCE battery data [39]	Source: CALCE, University of Maryland. Mixed cell formats (cylindrical, prismatic, pouch) and chemistries (LCO, LFP, NMC). Contains continuous and partial cycling, storage (calendar-ageing) tests, dynamic driving profiles, open-circuit voltage, and impedance measurements. Typical data volume: dozens of cells with up to $\mathcal{O}(10^3)$ cycles per campaign. Used for EFC lifetime auditing, voltage–capacity reconstruction, knee and reliability analyses, and descriptive cross-domain lifetime comparison.
ISU-ILCC battery aging (22582234) [40,41]	Source: Iowa State University DataShare (ISU-ILCC Battery Aging Dataset). A total of 251 cylindrical NMC/graphite 18650 cells subjected to long-term cycling with systematic variation of charge rate, discharge rate, and depth of discharge. The number and physical spacing of the available degradation observations depended on the experimental protocol and preprocessing stage. The dataset was used for protocol-held-out early-life EOL prediction, exploratory descriptor analysis, knee-point analysis, and reliability modelling.
ChemDataExtractor materials DB (11888115) [42]	Source: Huang & Cole, Sci. Data 7, 260 (2020); figshare 11888115. Automatically mined database of battery materials properties built from the scientific literature using ChemDataExtractor. Includes 292 313 records and 214 617 unique composition–property relations for 17 354 distinct compounds (capacity, voltage, conductivity, Coulombic efficiency, energy). Used to provide materials-level context for the time-series datasets and to compare capacity and voltage distributions across chemistries.

Interpolation was restricted to the interval bounded by the two observations bracketing the threshold; no extrapolation beyond this interval was performed. The first observed measurement below SOH = 0.80 was retained only for the sampling-resolution sensitivity audit and was not used as the primary EOL coordinate.

Cross-source lifetime reporting used equivalent full cycles (EFC), derived from the source throughput coordinate normalized by the corresponding reference capacity. The analysis did not assume that the fifth, tenth, or twentieth recorded observation represented exactly 5, 10, or 20 physical charge–discharge cycles. Accordingly, early-life windows were defined as the first N available degradation observations. When an explicit source-provided physical-cycle mapping was unavailable, no physical cycle count was inferred retrospectively.

Cells without an observed pair of measurements bracketing the selected SOH threshold were treated as right censored at their last available EFC coordinate. Regression metrics were calculated only for cells with an observed EOL, whereas Kaplan–Meier, Weibull, and lognormal lifetime analyses retained the censoring indicator. At the primary 80% SOH threshold, EOL was observed for all 12 CALCE cells, all 251 ISU–ILCC cells, and all eight NASA-randomized cells included in the EFC audit. At the more stringent 75% SOH threshold, two ISU–ILCC cells remained right censored.

To characterize the shape of the voltage–capacity trajectory we extract additional cycle-level descriptors inspired by earlier work on physics-informed features for data-driven ageing models [44–46]. These include the initial ohmic voltage drop ΔV_{IR}, the slope of the end-of-discharge region, plateau length in capacity coordinates, curvature of the mid-discharge regime and the energy throughput

$$E_k = \int_{t_k^{\mathrm{start}}}^{t_k^{\mathrm{end}}} V(t)\, I(t)\, \mathrm{d}t. \tag{16}$$

In addition to scalar descriptors, the continuous formulation enables defining shape-related quantities directly in functional form. For instance, the curvature of the capacity trajectory $Q(k)$ with respect to cycle index k can be expressed as

$$\kappa(k) = \frac{Q''(k)}{\left(1 + (Q'(k))^2\right)^{3/2}}, \tag{17}$$

which captures deviations from linear degradation. High curvature values are typically associated with the onset of accelerated ageing, including the transition to the knee region.

In addition, temporal statistics such as mean current and mean temperature per cycle are computed, yielding a compact feature vector that summarizes both operating conditions and degradation signatures. The resulting per-cycle table constitutes a unified representation across all public datasets considered, despite differences in protocol, operating window and sampling frequency. This harmonized preprocessing enables joint analysis of capacity fade, knee-onset behaviour and reliability across heterogeneous sources, without imposing any task-specific parametric model at this stage. Our choice of cycle-level descriptors is aligned with recent work on health-indicator based SOH/RUL models, where a moderate number of physically interpretable features extracted from voltage and current profiles is preferred over raw time-series inputs [44,45,47].

While advanced sequence models such as LSTM and Transformer architectures have demonstrated strong performance in battery prognostics, they typically require large homogeneous datasets and do not provide explicit continuous representations of degradation trajectories.

In contrast, the proposed INR-based framework emphasizes robustness under heterogeneous conditions and interpretability of trajectory-level features, which are critical for engineering applications.

6.1. Consensus knee detection

Knee points in battery capacity-fade trajectories are commonly interpreted as transitions from an approximately gradual degradation regime to accelerated ageing. Previous studies have investigated algorithmic knee detection, the interpretation of battery-ageing knees, and knee-aware trajectory prediction under different operating conditions [48–51].

Consensus knee location was estimated independently of the strict early-life prediction task. The detector combined a curvature-based criterion with a two-segment piecewise-linear criterion. Both components were applied to the complete available SOH-EFC trajectory, and the median of the valid estimates was returned as the consensus knee.

The consensus knee location defined here was used consistently in the knee–EOL association, knee-descriptor summaries, and knee-sensitivity analyses. Knee descriptors were not used as covariates in the Kaplan–Meier, Weibull, or lognormal reliability models, which were based exclusively on the EOL coordinate and the corresponding event or censoring indicator.

6.2. Association between knee location and EOL

Algorithm 1 Consensus knee detection from an SOH-EFC trajectory

Require: Ordered EFC coordinates $\{a_j\}$ and SOH observations $\{s_j\}$
Ensure: Consensus knee estimate a_{knee}
Smooth the trajectory using a Savitzky–Golay filter with window length 9 and polynomial order 3
Use a centred three-point median smoother when the trajectory is too short for the Savitzky–Golay window
Normalize the available coordinate range to $u_j \in [0,1]$
Compute $s'(u)$, $s''(u)$ and

$$\kappa(u) = \frac{|s''(u)|}{\left(1 + [s'(u)]^2\right)^{3/2}}$$

Restrict curvature search to $0.10 \le u \le 0.95$
Set the curvature threshold to the 90th percentile of κ within the search interval
Select the earliest point satisfying both the curvature threshold and a slope not greater than the median search-interval slope
Independently fit all admissible two-segment linear models with at least four points per segment
Accept the piecewise knee only when the two-segment fit reduces SSE by at least 5% and the post-knee slope is lower by more than 10^{-4}
Return the median of the valid curvature and piecewise-linear knee estimates

6.3. Exploratory early-life descriptor embedding

Exploratory unsupervised analysis was performed using descriptors calculated from the first $N = 5$ available degradation observations. Cells with a complete useable descriptor vector were retained, yielding 244 cells. Seven of these cells had already crossed the operational 80% SOH threshold within the five-observation prefix. The embedding and clustering analyses are therefore retrospective descriptions of the five-observation descriptor space and are not part of the strict prospective early-life prediction benchmark.

Numeric descriptors with more than 40% missing values, near-zero variance, identifier fields, and variables containing EOL, RUL, knee, censoring, analysis-time, or other full-trajectory information were excluded. The remaining 21 descriptors were median-imputed and standardized using the full exploratory sample. K-means solutions with $k \in \{2,3,4,5,6\}$ were compared using the silhouette score. Clustering was conducted in the complete standardized descriptor space, with random seed 2026, and not in the two-dimensional embeddings. PCA and t-SNE were used only for visualization. The t-SNE projection used PCA initialization, perplexity 30, automatic learning-rate selection, and 1500 optimization iterations.

Algorithm 2 Continuous battery ageing analysis pipeline

Require: Raw battery datasets containing time-series measurements $\{I(t), V(t), T(t)\}$
Ensure: Lifetime statistics, degradation descriptors, and early-life prediction outputs
Harmonize heterogeneous datasets into a cycle-resolved representation
Segment charge–discharge records into individual cycles
Compute per-cycle capacity Q_k, state of health SOH_k, and energy E_k
Construct continuous voltage–capacity and capacity-cycle representations using INR models
Extract trajectory-level descriptors such as slope, curvature, plateau length and knee-related features
Estimate EOL and knee-cycle statistics from full trajectories
Perform reliability analysis using Kaplan–Meier, Weibull and lognormal models
Train early-life predictors using the first $N \in \{5, 10, 20\}$ available degradation observations
Construct fold-external conformal intervals for the strict prospective early-life predictor
Conduct retrospective MC-dropout and out-of-fold Random-Forest diagnostics on the 244-cell descriptor cohort
Conduct retrospective PCA, t-SNE, and K-means analysis of the five-observation descriptor space

A compact overview of the full computational pipeline is provided in Fig. 1. The diagram highlights the integration of continuous representations with reliability analysis and predictive modelling under heterogeneous conditions.

Due to filtering steps, missing measurements and the merging of trajectory-level descriptors (e.g., knee point) with EOL statistics, the effective number of cells varies slightly between analyses. These differences do not affect the overall conclusions.

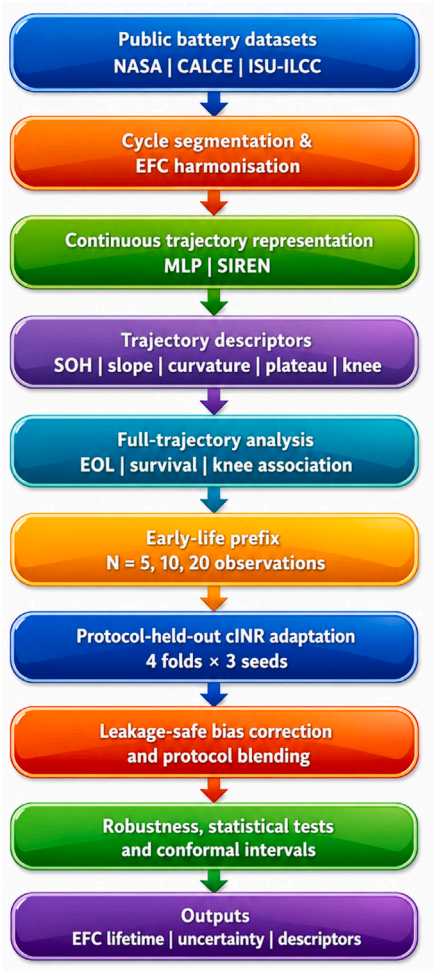


Fig. 1. Workflow of the proposed battery ageing analysis framework. Heterogeneous datasets are transformed into continuous trajectory representations, enabling feature extraction, knee/EOL analysis, reliability modelling and early-life prediction with uncertainty quantification.

The effective sample size differed across analyses because each task required a distinct combination of available early-life observations, observed EOL, complete descriptor values, or a valid consensus knee. The primary EFC lifetime audit included 271 cells, the exploratory early-life descriptor analysis included 244 cells, and the knee-EOL association included 172 cells. In the strict early-life prediction benchmark, the sample size decreased from 237 cells at $N = 5$ to 167 at $N = 10$ and 44 at $N = 20$ because progressively longer complete observation prefixes were required.

7. Results

7.1. Early-life prediction performance

All reported results are obtained under a harmonized evaluation protocol, ensuring consistent comparison across heterogeneous datasets and avoiding data leakage between early-life prediction and full-trajectory analysis. Table 4 summarizes the early-life RUL prediction performance for all evaluated models and input windows. Prospective landmark eligibility was verified before calculation of the prediction metrics. In the ISU–ILCC cohort, 244, 232, and 120 cells contained at least 5, 10, and 20 available degradation observations, respectively.

Table 4
Protocol-held-out early-life EOL prediction at the 80% SOH threshold. Errors are reported in EFC. Neural results use mean aggregation across three seeds. "Hybrid" denotes leakage-safe outer-fold bias correction and blending with ProtocolMedianEOL.

N observations	Model	n	MAE	Median AE	RMSE	Bias
5	ProtocolMedianEOL	237	116.656	94.275	149.391	−4.124
5	LocalLinear	234	189.590	–	255.313	−76.088
5	Raw cINR-MLP	237	319.192	–	390.880	−316.174
5	Raw cINR-SIREN	237	133.929	–	176.179	−12.724
5	Hybrid cINR-MLP	237	**113.571**	90.730	**148.635**	−15.471
5	Hybrid cINR-SIREN	237	116.003	94.268	148.848	−4.825
10	ProtocolMedianEOL	167	115.572	–	153.998	−1.972
10	LocalLinear	166	123.571	–	180.162	77.673
10	Raw cINR-MLP	167	176.582	–	246.899	−168.987
10	Raw cINR-SIREN	167	150.964	–	211.363	51.562
10	Hybrid cINR-MLP	167	**107.786**	76.911	**147.528**	−25.557
10	Hybrid cINR-SIREN	167	114.237	84.689	154.591	11.275
20	ProtocolMedianEOL	44	125.021	–	167.425	15.727
20	LocalLinear	43	212.301	–	258.091	200.031
20	Raw cINR-MLP	44	130.574	–	167.797	−90.149
20	Raw cINR-SIREN	44	196.850	–	324.186	23.786
20	Hybrid cINR-MLP	44	**112.823**	84.637	**149.007**	−4.659
20	Hybrid cINR-SIREN	44	134.685	79.997	194.005	30.831

After requiring the final input landmark to precede the interpolated 80% SOH crossing, 237 cells were retained for $N = 5$, 167 for $N = 10$, and 44 for $N = 20$. Thus, 7 cells at $N = 5$, 65 at $N = 10$, and 76 at $N = 20$ were excluded because the operational EOL threshold had already been reached within the corresponding input prefix. The retained proportions were therefore 97.1%, 72.0%, and 36.7%, respectively. All prediction results in Table 4 were calculated exclusively from cells that remained at risk at the corresponding landmark.

Table 4 shows that the raw cINR predictors did not exhibit uniform superiority. The raw cINR-MLP model strongly underestimated EOL, with mean biases of −316.17, −168.99, and −90.15 EFC at $N = 5$, 10, and 20, respectively. The raw cINR-SIREN predictor was less biased at $N = 5$, but its error and bias varied substantially across observation windows.

Leakage-safe outer-fold bias correction and protocol blending reduced the cINR-MLP MAE from 319.19 to 113.57 EFC at $N = 5$, from 176.58 to 107.79 EFC at $N = 10$, and from 130.57 to 112.82 EFC at $N = 20$. These changes correspond to absolute reductions of 205.62, 68.80, and 17.75 EFC, or relative reductions of 64.4%, 39.0%, and 13.6%, respectively.

Compared with ProtocolMedianEOL, the hybrid cINR–MLP reduced MAE by 3.09, 7.79, and 12.20 EFC for $N = 5$, 10, and 20, respectively. These numerical differences are interpreted together with the paired tests reported below and are not treated as evidence of model superiority. All hybrid predictions satisfied the non-negative RUL constraint.

The stored outer-fold postprocessing parameters show that the cINR component retained a non-zero contribution after bias correction and protocol blending. Median w_{-f} values were 0.100, 0.295, and 0.560 for $N = 5$, 10, and 20, respectively. These weights quantify the contribution of the bias-corrected neural estimate and should be interpreted together with the paired comparisons, which did not establish superiority over ProtocolMedianEOL. At $N = 20$, the bias-corrected cINR–MLP prediction alone achieved a lower MAE than the final protocol-blended hybrid (109.37 versus 112.82 EFC). Thus, although protocol blending improved the raw neural prediction and remained better than the protocol-median reference, it was not uniformly beneficial after bias correction. This result further supports interpreting the blending stage as a calibration mechanism rather than as a guaranteed source of additional accuracy.

Table 6 separates numerical MAE differences from statistically supported paired improvements. After correction across the six predefined comparisons, the hybrid cINR–MLP predictor improved over the raw cINR–MLP predictor according to both paired tests at $N = 5$ and $N = 10$, but not at $N = 20$. None of the three comparisons with ProtocolMedianEOL was supported by both adjusted tests. The hybrid is therefore interpreted as a leakage-safe calibration of the raw neural extrapolation and as numerically complementary to, rather than uniformly superior to, the protocol-informed reference.

Fig. 2 illustrates the distinction between raw neural extrapolation and calibrated hybrid prediction. The raw MLP showed substantial underestimation for $N = 5$ and $N = 10$, whereas leakage-safe calibration reduced its MAE to a level close to or numerically below ProtocolMedianEOL. The lowest observed MAE was 107.79 EFC for the hybrid MLP at $N = 10$. Nevertheless, the paired differences between the hybrid MLP and ProtocolMedianEOL were not statistically significant after multiplicity correction; the figure should therefore be interpreted as evidence of numerical complementarity rather than model superiority.

7.1.1. Matched-cohort and seed-aggregation sensitivity analyses

To separate the effect of observation-window length from the progressive change in cohort composition, an additional matched-cohort analysis retained only cells with valid prospectively eligible predictions at all three landmarks, $N = 5$, 10, and 20. This procedure yielded a common evaluation cohort of $n = 44$ cells. Consequently, differences reported in Table 7 reflect changes in the amount of early trajectory information available for the same cells rather than differences caused by the exclusion of shorter-lived or incompletely observed cells.

Table 7 shows the performance obtained when all observation windows were evaluated on the same 44 cells. The hybrid cINR–MLP MAE decreased from 201.929 EFC at $N = 5$ to 195.728 EFC at $N = 10$ and 112.823 EFC at $N = 20$. The corresponding ProtocolMedianEOL errors were 185.598, 182.684, and 125.021 EFC, respectively. The hybrid predictor therefore had a higher MAE than ProtocolMedianEOL within the matched cohort at $N = 5$ and $N = 10$, but a lower MAE by 12.198 EFC at $N = 20$. The bias-corrected cINR–MLP estimate alone attained the lowest $N = 20$ MAE, 109.373 EFC. These results should be distinguished from the primary-window benchmark in Table 4, for which the eligible cohort decreased from 237 to 167 and 44 cells as the observation window increased.

The paired window comparisons in Table 8 directly evaluated whether additional observations reduced prediction error for the same cells. For the hybrid cINR–MLP predictor, the change in MAE from $N = 5$ to $N = 10$ was −6.200 EFC, whereas the changes from $N = 10$ to $N = 20$ and from $N = 5$ to $N = 20$ were −82.906 and −89.106 EFC, respectively. After Holm correction, statistically supported improvements were observed for the comparisons between

Table 5
Outer-fold contribution and prediction error of the mean-aggregated cINR–MLP hybrid. The fold weight w_{-f} multiplies the bias-corrected cINR prediction; IQR denotes the interquartile range across outer folds. Errors are expressed in EFC. A dash indicates that a corrected-only aggregate was not retained in the source table.

N	n	Folds	Median w_{-f} [IQR]	Corrected cINR MAE	Hybrid MAE	Protocol MAE	Raw cINR MAE
5	237	4	0.100 [0.075–0.105]	180.00	113.57	116.66	319.19
10	167	4	0.295 [0.255–0.323]	136.17	107.79	115.57	176.58
20	44	4	0.560 [0.470–0.603]	109.37	112.82	125.02	130.57

Table 6
Predefined paired comparisons of the mean-aggregated hybrid cINR–MLP predictor. ΔMAE is the hybrid MAE minus the reference MAE, so negative values favour the hybrid. The six Wilcoxon p-values were adjusted jointly using the Holm procedure, and the six Monte Carlo sign-flip p-values were adjusted separately across the same predefined family.

N	Model 1	Model 2	n_{paired}	ΔMAE	Wilcoxon p_{raw}	Wilcoxon p_{Holm}	Sign-flip p_{Holm}
5	Hybrid cINR–MLP	ProtocolMedianEOL	237	−3.084	0.1303	0.2606	0.6114
5	Hybrid cINR–MLP	Raw cINR–MLP	237	−205.621	<0.001	<0.001	<0.001
10	Hybrid cINR–MLP	ProtocolMedianEOL	167	−7.787	0.0463	0.1852	0.5003
10	Hybrid cINR–MLP	Raw cINR–MLP	167	−68.796	<0.001	<0.001	<0.001
20	Hybrid cINR–MLP	ProtocolMedianEOL	44	−12.198	0.2162	0.2606	0.6114
20	Hybrid cINR–MLP	Raw cINR–MLP	44	−17.751	0.0623	0.1868	0.5003

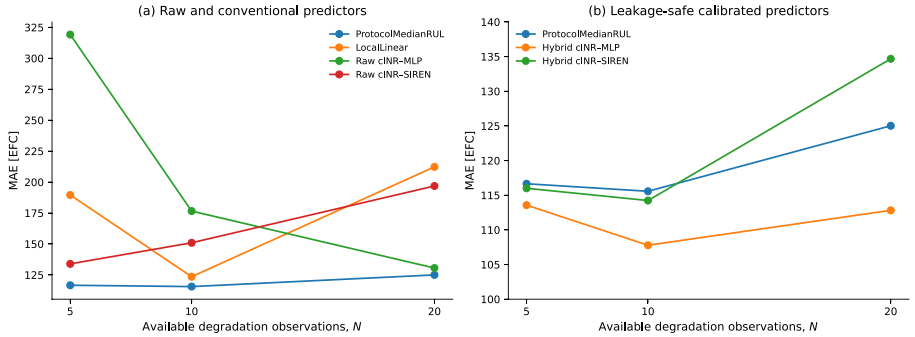


Fig. 2. Protocol-held-out early-life EOL prediction performance. Panel (a) compares the protocol-median predictor, local-linear extrapolation, and raw cINR predictors. Panel (b) compares the leakage-safe hybrid cINR predictors with ProtocolMedianEOL. Errors are reported in EFC, and lower MAE indicates better performance.

Table 7
Matched-cohort early-life EOL prediction performance. Only cells with valid prospectively eligible predictions at all three observation windows were retained. Errors are expressed in EFC.

N	Model	n	MAE	Median AE	RMSE	Bias
5	ProtocolMedianEOL	44	185.598	152.840	232.300	−71.837
5	Raw cINR–MLP	44	626.784	621.293	659.994	−621.224
5	Bias-corrected cINR–MLP	44	381.739	372.528	426.147	−360.618
5	Hybrid cINR–MLP	44	201.929	160.779	247.168	−136.227
10	ProtocolMedianEOL	44	182.684	150.160	229.025	−76.417
10	Raw cINR–MLP	44	385.216	386.318	422.409	−383.461
10	Bias-corrected cINR–MLP	44	285.217	277.990	328.325	−276.337
10	Hybrid cINR–MLP	44	195.728	168.833	239.762	−155.615
20	ProtocolMedianEOL	44	125.021	79.997	167.425	15.727
20	Raw cINR–MLP	44	130.574	108.425	167.797	−90.149
20	Bias-corrected cINR–MLP	44	109.373	84.369	144.614	−19.096
20	Hybrid cINR–MLP	44	112.823	84.637	149.007	−4.659

$N = 5$ and $N = 20$ and between $N = 10$ and $N = 20$. No statistically supported improvement was found between $N = 5$ and $N = 10$.

The ProtocolMedianEOL reference exhibited a similar window-dependent pattern. The reduction observed at $N = 20$ should therefore not be attributed uniquely to the cINR component.

The sensitivity analysis Table 9 compared arithmetic-mean and median aggregation across the three random seeds. Mean aggregation yielded hybrid cINR–MLP MAEs of 113.571, 107.786, and 112.823 EFC for $N = 5$, $N = 10$, and $N = 20$, respectively, whereas median aggregation yielded 114.003, 110.777, and 111.081 EFC. The corresponding median-minus-mean differences were 0.431, 2.992, and −1.741 EFC. These differences were small, indicating that the principal conclusions were robust to the choice of seed-aggregation rule. Arithmetic-mean aggregation was retained for the primary results because it was specified as the main aggregation procedure before the final comparative analysis.

Fig. 3 shows that the principal weakness of the raw cINR-MLP predictor was systematic EOL underestimation. Its mean bias was −316.17, −168.99, and −90.15 EFC for $N = 5$, 10, and 20, respectively. Leakage-safe correction and protocol blending reduced the corresponding MLP biases to −15.47, −25.56, and −4.66 EFC. The SIREN bias was smaller at $N = 5$ but changed direction across observation windows, illustrating that raw cINR extrapolation was both architecture- and window-dependent.

Table 8
Paired comparison of observation windows in the matched cohort. ΔMAE is the MAE at N_2 minus the MAE at N_1; negative values favour the longer observation window. Holm correction was applied across the three window comparisons separately for each model.

Model	N_1	N_2	n	ΔMAE	Wilcoxon p_{Holm}	Sign-flip p_{Holm}
ProtocolMedianEOL	5	10	44	−2.913	0.230	0.414
ProtocolMedianEOL	5	20	44	−60.577	0.001	<0.001
ProtocolMedianEOL	10	20	44	−57.663	0.002	<0.001
Raw cINR–MLP	5	10	44	−241.568	<0.001	<0.001
Raw cINR–MLP	5	20	44	−496.211	<0.001	<0.001
Raw cINR–MLP	10	20	44	−254.642	<0.001	<0.001
Bias-corrected cINR–MLP	5	10	44	−96.522	<0.001	<0.001
Bias-corrected cINR–MLP	5	20	44	−272.365	<0.001	<0.001
Bias-corrected cINR–MLP	10	20	44	−175.843	<0.001	<0.001
Hybrid cINR–MLP	5	10	44	−6.200	0.428	0.383
Hybrid cINR–MLP	5	20	44	−89.106	<0.001	<0.001
Hybrid cINR–MLP	10	20	44	−82.906	0.001	<0.001

Table 9
Sensitivity of the bias-corrected cINR–MLP protocol blend to aggregation across the three random seeds. ΔMAE is median-aggregated MAE minus mean-aggregated MAE. Errors are expressed in EFC.

N	n	Mean-aggregated MAE	Median-aggregated MAE	ΔMAE	Mean-aggregated RMSE	Median-aggregated RMSE	Mean-aggregated bias	Median-aggregated bias
5	237	113.571	114.003	0.431	148.635	149.390	−15.471	−15.315
10	167	107.786	110.777	2.992	147.528	150.211	−25.557	−23.374
20	44	112.823	111.081	−1.741	149.007	148.643	−4.659	−11.789

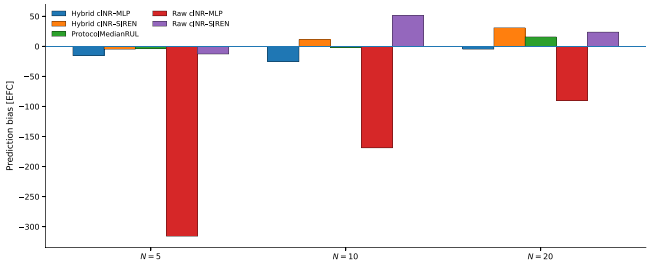


Fig. 3. Prediction bias for the protocol-informed baseline, raw cINR predictors, and leakage-safe hybrid predictors at $N = 5$, 10, and 20 available degradation observations. Negative values indicate systematic EOL underestimation, whereas positive values indicate overestimation.

Table 10
Fold-external conformal performance of the mean-aggregated hybrid cINR-MLP predictor. Exact binomial 95% confidence intervals are reported for empirical coverage. Interval widths are expressed in EFC.

N	n	Nominal	Covered	Empirical	95% CI	Median width
5	237	0.90	214/237	0.903	[0.858, 0.937]	506.23
5	237	0.95	229/237	0.966	[0.935, 0.985]	642.92
10	167	0.90	152/167	0.910	[0.856, 0.949]	488.84
10	167	0.95	162/167	0.970	[0.932, 0.990]	580.92
20	44	0.90	42/44	0.955	[0.845, 0.994]	488.43
20	44	0.95	43/44	0.977	[0.880, 0.999]	605.20

7.2. Conformal uncertainty of early-life prediction

As summarized in Table Table 10, empirical coverage was close to or above the nominal level for all observation windows. However, the confidence intervals were wide for $N = 20$ because only 44 cells remained prospectively eligible. In particular, the 95% confidence interval for nominal 90% coverage extended from 0.845 to 0.994. Moreover, median interval widths remained between approximately 488 and 643 EFC. The conformal results therefore support empirical calibration within the present cohort but also reveal substantial uncertainty in cell-specific lifetime prediction.

As summarized in Table Table 12, observed EOL differed significantly between the three validation domains. The Brown–Forsythe test indicated heterogeneous variances ($F(2268) = 8.58$, $p = 2.45 \times 10^{-4}$); therefore, Welch ANOVA and the Kruskal–Wallis test were treated as the primary inferential analyses. Both tests identified a significant domain effect ($F_{Welch}(2, 19.94) = 167.02$, $p = 3.52 \times 10^{-13}$; $H(2) = 41.57$, $p = 9.42 \times 10^{-10}$). The corresponding effect-size estimates ($\eta^2 = 0.137$, $\omega^2 = 0.131$, and $\varepsilon^2 = 0.148$) indicate substantial between-domain lifetime heterogeneity.

Pairwise analyses reported in Table Table 13 confirmed large differences in observed lifetime between all three validation domains. CALCE cells had substantially lower EOL than ISU-ILCC cells (Cohen's $d = -1.717$, 95% CI $[-1.921, -1.525]$; Hedges' $g = -1.713$; Cliff's $\delta = -0.934$, 95% CI $[-0.973, -0.886]$). The corresponding Welch and Mann–Whitney comparisons remained significant after Holm adjustment ($p_{Holm} = 2.30 \times 10^{-20}$ and 1.43×10^{-7}, respectively).

CALCE cells also had lower EOL than NASA-randomized cells ($d = -2.433$, 95% CI $[-5.111, -1.575]$; $g = -2.330$; $\delta = -0.938$, 95% CI $[-1.000, -0.771]$). Although this contrast was large, its Cohen's d confidence interval was comparatively wide because the comparison involved only 12 CALCE and eight NASA-randomized cells. Both adjusted tests nevertheless remained significant (Welch $p_{Holm} = 2.37 \times 10^{-4}$; Mann–Whitney $p_{Holm} = 1.86 \times 10^{-4}$).

ISU-ILCC cells had higher EOL than NASA-randomized cells ($d = 1.077$, 95% CI $[0.843, 1.300]$; $g = 1.074$; $\delta = 0.743$, 95% CI $[0.569, 0.871]$). This difference also remained significant after Holm adjustment (Welch $p_{Holm} = 1.32 \times 10^{-6}$; Mann–Whitney $p_{Holm} = 1.86 \times 10^{-4}$).

The consistency between standardized mean differences, rank-based effect sizes, and adjusted hypothesis tests demonstrates pronounced cross-domain lifetime heterogeneity. These effects describe differences between heterogeneous laboratory populations and should not be interpreted as causal effects of dataset membership.

As summarized in Table Table 11, the fitted Weibull shape parameter exceeded one in every validation domain, correspond. The estimated shape was 1.813 for CALCE, 2.144 for ISU-ILCC, and 3.627 for NASA randomized. The NASA estimate should be interpreted cautiously because it was based on only eight cells.

AIC comparisons favoured the Weibull distribution for the pooled sample and the ISU-ILCC domain. The Weibull and lognormal fits were nearly indistinguishable for CALCE, whereas the lognormal model had a slightly lower AIC for NASA randomized. Because the validation

Table 11
Observed EOL summaries and fitted Weibull and lognormal lifetime parameters on the EFC scale. AIC values are comparable only between models fitted to the same scope. No cells were right censored at the primary 80% SOH threshold. The pooled fit is descriptive because the validation domains represent heterogeneous lifetime populations.

Scope	n	Mean	SD	Median	KM median	k	λ	Weibull AIC	μ	σ	exp(μ)	Lognormal AIC
Pooled	271	607.21	323.40	555.04	555.04	1.925	682.15	3882.92	6.216	0.721	500.50	3964.88
CALCE	12	116.13	71.13	98.34	91.75	1.813	131.14	136.49	4.582	0.622	97.73	136.62
ISU-ILCC	251	640.20	311.42	575.82	575.82	2.144	721.19	3582.26	6.310	0.629	550.13	3651.66
NASA randomized	8	308.85	90.49	296.88	298.74	3.627	341.11	98.28	5.699	0.254	298.68	95.94

Table 12
Omnibus comparison of observed EOL between validation domains on the EFC scale. Because the Brown–Forsythe test indicated heterogeneous variances, Welch ANOVA and the Kruskal–Wallis test were treated as the primary inferential analyses.

Test	Statistic	df_1	df_2	p	Effect size
Classical one-way ANOVA	21.339	2	268.00	2.51×10^{-9}	$\eta^2 = 0.137$; $\omega^2 = 0.131$
Welch one-way ANOVA	167.017	2	19.94	3.52×10^{-13}	–
Kruskal–Wallis	41.566	2	–	9.42×10^{-10}	$\epsilon^2 = 0.148$
Brown–Forsythe	8.579	2	268.00	2.45×10^{-4}	–

Table 13
Pairwise comparisons of observed EOL on the EFC scale. Cohen's d, bias-corrected Hedges' g, and Cliff's δ are signed as the first domain minus the second domain. Confidence intervals were obtained using 5000 cell-level bootstrap resamples. The reported p-values were adjusted across the three pairwise comparisons using the Holm procedure.

Comparison	n_A	n_B	Cohen's d [95% CI]	Hedges' g	Cliff's δ [95% CI]	Welch p_{Holm}	Mann–Whitney p_{Holm}
CALCE vs. ISU-ILCC	12	251	−1.717 [−1.921, −1.525]	−1.713	−0.934 [−0.973, −0.886]	2.30×10^{-20}	1.43×10^{-7}
CALCE vs. NASA randomized	12	8	−2.433 [−5.111, −1.575]	−2.330	−0.938 [−1.000, −0.771]	2.37×10^{-4}	1.86×10^{-4}
ISU-ILCC vs. NASA randomized	251	8	1.077 [0.843, 1.300]	1.074	0.743 [0.569, 0.871]	1.32×10^{-6}	1.86×10^{-4}

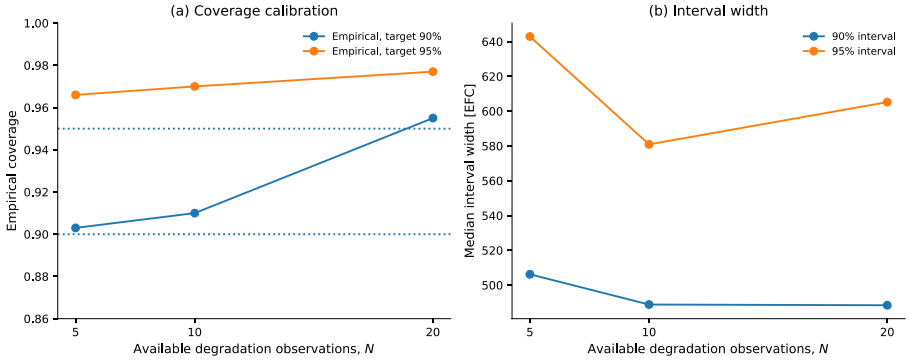


Fig. 4. Fold-external conformal performance of the mean-aggregated hybrid cINR-MLP predictor: (a) empirical coverage relative to the nominal 90% and 95% levels, and (b) median prediction-interval width expressed in EFC.

domains differed strongly in sample size and lifetime range, pooled parameters were treated as descriptive summaries rather than as evidence of a universal battery-lifetime distribution.

As shown in Fig. 4, empirical coverage was close to or above the corresponding nominal level for every observation window. This calibration was obtained with broad intervals: median widths ranged from approximately 488 to 643 EFC. The intervals therefore quantify substantial cell-specific lifetime uncertainty and should not be interpreted as high-precision early-life forecasts.

The same number of available degradation observations represented markedly different amounts of electrochemical throughput across validation domains. For example, among cells that remained at risk at $N = 20$, the median final-input coordinate was 19.69 EFC for CALCE, 817.01 EFC for ISU–ILCC, and 300.26 EFC for NASA randomized. Therefore, N defines the amount of source information supplied to the model but not a common physical ageing stage. Prediction errors were consequently reported in EFC, and cross-window comparisons were interpreted jointly with changes in landmark position and cohort composition.

The large domain-dependent differences in Table 14 explain why the terms "first 5 cycles" or "first 20 cycles" are physically misleading for the harmonized analysis. All revised predictive statements therefore refer to the first N available degradation observations and report errors on the EFC scale.

Fig. 5 confirms that the observation index is not a physically comparable cycle coordinate across the evaluated domains. Consequently, all early-life windows are defined by the number of available observations, whereas prediction errors and lifetime landmarks are reported in EFC.

7.3. Sensitivity of early-life predictions

Sensitivity analysis included two- and four-fold downsampling, contiguous and random missing observations, and additive SOH noise. The effect was architecture- and window-dependent. For example, the SIREN MAE at $N = 5$ changed from 133.93 EFC under clean input to 140.63 EFC at noise standard deviation 0.01, whereas four-fold downsampling yielded 122.75 EFC. At $N = 20$, SIREN degraded from 196.85 EFC under clean input to 229.01 EFC under four-fold downsampling and to 216.21 EFC under 20% random missingness.

Some perturbations numerically reduced the MAE of the raw MLP model. This should not be interpreted as a beneficial effect of data corruption: the clean MLP predictions exhibited substantial systematic

Table 14
Median EFC coordinate of the final input observation among cells that remained above the 80% SOH threshold at the corresponding landmark. Numbers in parentheses indicate the number of prospectively eligible cells.
The effect of threshold-crossing interpolation differed substantially between validation domains because of their different observation densities. For CALCE, the median interpolated EOL was 98.340 EFC, compared with 98.352 EFC for the first observed point below 80% SOH; the median displacement was only 0.066 EFC. For ISU–ILCC, the corresponding values were 575.821 and 603.226 EFC, with a median displacement of 25.492 EFC. For NASA randomized, the median interpolated and first-below-threshold coordinates were 296.877 and 309.501 EFC, respectively, giving a median displacement of 10.028 EFC. The median widths of the observed threshold-crossing brackets were 0.788, 46.962, and 24.559 EFC for CALCE, ISU–ILCC, and NASA randomized, respectively. These differences demonstrate that using the first observed point below the threshold would systematically delay EOL, particularly in the more sparsely sampled domains.

Validation domain	$N = 5$	$N = 10$	$N = 20$
CALCE	4.998 (12)	9.942 (12)	19.687 (12)
ISU-ILCC	178.343 (237)	426.104 (167)	817.011 (44)
NASA randomized	83.755 (8)	154.324 (8)	300.258 (3)

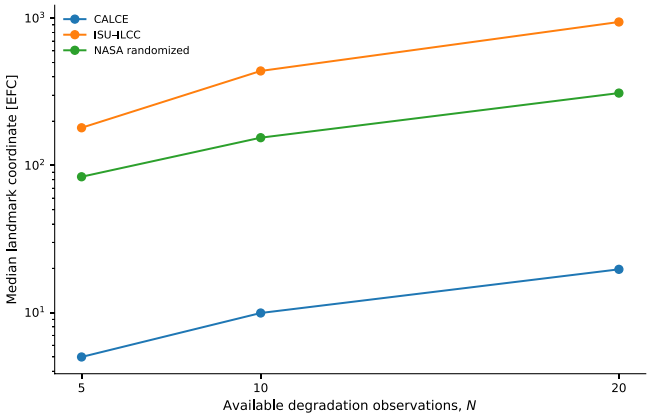


Fig. 5. Median EFC coordinate represented by the first 5, 10, and 20 available degradation observations in the CALCE, ISU-ILCC, and NASA-randomized domains. The logarithmic vertical scale highlights the substantial differences in electrochemical throughput represented by the same number of source observations.

underestimation, and perturbations could partially offset that bias. ProtocolMedianEOL was unchanged under input perturbations because it did not use the perturbed cell prefix; its invariance is therefore constructional and not evidence of empirical noise robustness.

End-of-life distributions and survival functions were recalculated directly on the EFC scale. At the primary 80% SOH threshold, median observed EOL was 98.34 EFC for CALCE, 575.82 EFC for ISU-ILCC, and 296.88 EFC for the NASA-randomized subset. These differences demonstrate that the source observation index cannot be interpreted as a common physical cycle count.

Fig. 6 shows substantial lifetime heterogeneity between the evaluated domains. Because all cells included in the primary 80% EOL audit reached the threshold, censoring had limited influence on the corresponding survival curves. The survival formulation remains relevant for alternative SOH thresholds and for future operational datasets containing incomplete lifetime records.

For the knee-EOL association analysis, a cell was included when an observed $EOL_{0.80}$ value and a valid consensus knee returned by both the implemented detection and stability filters were available. Knee location was not constrained to precede $EOL_{0.80}$, because the 80% SOH crossing represents an operational end-of-life threshold rather than the end of the recorded experimental trajectory. Applying these criteria yielded 172 cells, comprising 167 ISU-ILCC cells and five NASA-randomized cells (see Table 15).

The pooled association was strong ($\rho_S = 0.834$, 95% bootstrap CI $[0.738, 0.907]$) and remained highly significant after Holm correction. The ISU–ILCC result was similar to the pooled estimate ($n = 167$, $\rho_S = 0.828$).

Table 15
Association between robust consensus knee location and the first crossing of the 80% SOH threshold on the EFC scale. The pooled confidence interval was calculated using 10,000 cell-level bootstrap resamples. Large-sample Spearman inference was used for the pooled and ISU-ILCC scopes, whereas the NASA-randomized p-value was calculated using an exact two-sided permutation test over all 5! = 120 rank permutations. The three resulting raw p-values were adjusted jointly using the Holm procedure.

Scope	n	Spearman ρ_S	p_{Holm}
Pooled	172	0.834	3.36×10^{-45}
ISU-ILCC	167	0.828	5.28×10^{-43}
NASA randomized	5	0.900	8.33×10^{-2}

The NASA-randomized estimate was numerically high ($n = 5$, $\rho_S = 0.900$), but the exact two-sided permutation test was not statistically significant after Holm correction ($p_{Holm} = 0.0833$). Because the estimate was based on only five cells, it is reported as a descriptive domain-specific result. The principal inference is therefore supported by the pooled and ISU–ILCC analyses rather than by the small NASA-randomized subset.

These results establish a strong full-trajectory association between knee position and the operational EOL threshold, but do not by themselves demonstrate prospective early-life prediction.

Fig. 7 shows that cells with a later robust consensus knee generally reached the 80% SOH threshold at a later EFC coordinate. Across 172 cells, the pooled Spearman correlation was $\rho_S = 0.834$, with a 95% bootstrap confidence interval of $[0.738, 0.907]$ and a Holm-adjusted p-value of 3.36×10^{-45}. The association was similarly strong in the ISU–ILCC cohort ($n = 167$, $\rho_S = 0.828$).

The NASA-randomized result was based on only five cells ($\rho_S = 0.900$) and was not statistically significant under exact permutation inference ($p_{Holm} = 0.0833$); it should therefore be interpreted descriptively.

For 23 of the 172 cells, the robust consensus knee occurred at or after the first 80% SOH crossing. This was possible because the experimental trajectory continued after the operational EOL threshold had been reached. Consequently, the observed relationship represents a full-trajectory geometrical association and does not establish that the detected knee necessarily provides prospective advance warning of EOL.

7.3.1. Sensitivity analyses of the knee-EOL association

After excluding the 23 cells for which the detected knee occurred at or after operational EOL, the remaining $n = 149$ cells yielded a Spearman correlation of $\rho_S = 0.820$ with a 95% bootstrap CI of $[0.714, 0.903]$. The partial rank correlation controlling for validation domain and maximum recorded EFC was $\rho_{partial} = 0.687$ ($p = 7.01 \times 10^{-22}$). When knee detection was repeated using trajectories truncated at the first 80% SOH crossing, a valid knee was obtained for 157 cells and its association with EOL was $\rho_S = 0.501$ (95%

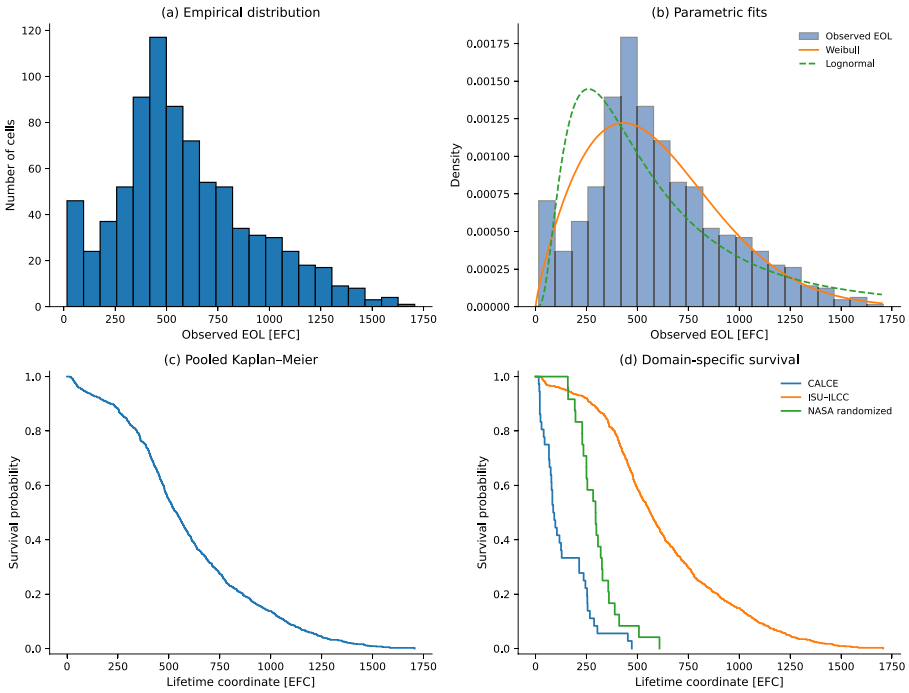


Fig. 6. End-of-life distribution and survival analysis on the EFC scale: (a) empirical distribution of observed EOL; (b) descriptive Weibull and lognormal fits; (c) pooled Kaplan–Meier survival function; and (d) domain-specific Kaplan–Meier survival functions. Parametric fits are included as descriptive lifetime summaries and do not imply that one universal lifetime distribution governs all evaluated chemistries and operating protocols.

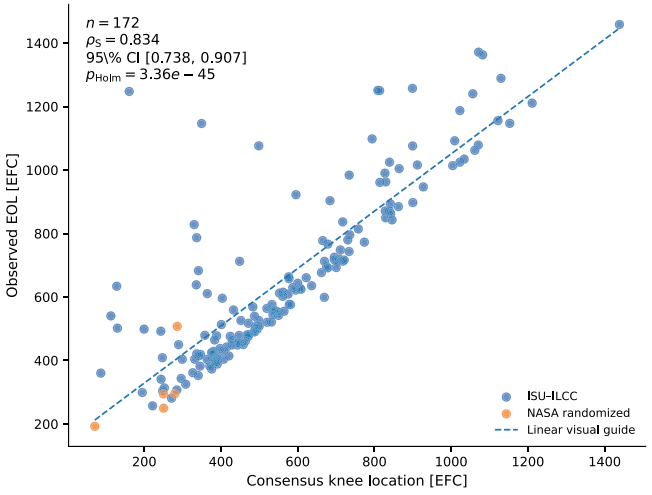


Fig. 7. Association between robust consensus knee location and the first crossing of the 80% SOH end-of-life threshold on the EFC scale. Knee locations were estimated from the available full trajectories and were not constrained to precede the 80% SOH crossing. The dashed regression line is included only as a visual guide. Statistical inference is based on Spearman rank correlation and 10,000 cell-level bootstrap resamples ($n = 172$, $\rho_S = 0.834$, 95% CI $[0.738, 0.907]$, Holm-adjusted $p = 3.36 \times 10^{-45}$).

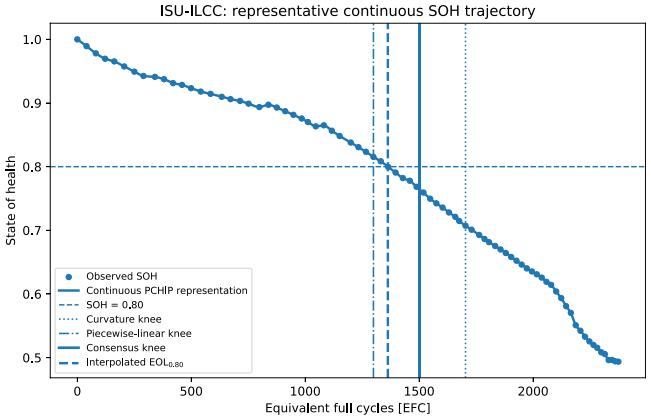


Fig. 8. Representative continuous SOH-EFC trajectory from the ISU–ILCC validation domain. Observed degradation measurements are shown together with a continuous shape-preserving visual representation, the curvature-based and piecewise-linear knee estimates, their consensus location, and the linearly interpolated first crossing of the 80% SOH threshold. The shape-preserving curve is included as a visual guide; knee locations were calculated using the consensus detector described in Algorithm 1. The example illustrates the geometrical quantities used in the full-trajectory knee analysis and is not a prospective early-life prediction.

CI $[0.356, 0.635]$). These analyses distinguish a genuine pre-threshold association from correlation induced primarily by the extent of the recorded post-threshold trajectory.

The corresponding cell-level relationships are shown in Figs. A.9 and A.10.

Fig. 8 provides a trajectory-level view of the quantities summarized by the population knee-EOL analysis. The two component detectors need not return identical locations; their median was used as the consensus estimate when both were valid. The example also illustrates that the detected knee may occur close to or after the operational 80% SOH threshold when measurements continue beyond that threshold.

Among the 172 cells included in the full-trajectory knee-EOL analysis, 23 cells (13.4%) had a robust consensus knee located at or after the first crossing of the SOH = 0.80 threshold. Restricting the descriptive sample to cells for which the detected knee preceded operational EOL retained 149 cells (86.6%). Thus, a pre-EOL knee was observed for the majority of the analysed cells, but the presence of 23 late or post-threshold knees confirms that the strong pooled association should not be interpreted as a universally prospective warning signal.

Table 16
Sensitivity analyses of the association between consensus knee location and interpolated EOL at the 80% SOH threshold. Bootstrap confidence intervals used 10,000 cell-level resamples. The partial rank result controls for validation domain and maximum recorded EFC.

Analysis	n	Spearman ρ_S	95% CI	p
Full available trajectories	172	0.834	[0.738, 0.907]	1.12×10^{-45}
Knee preceding operational EOL	149	0.820	[0.714, 0.903]	1.81×10^{-37}
Partial rank: domain and maximum recorded EFC	149	0.687	–	7.01×10^{-22}
Trajectories truncated at first 80% SOH crossing	157	0.501	[0.356, 0.635]	2.34×10^{-11}

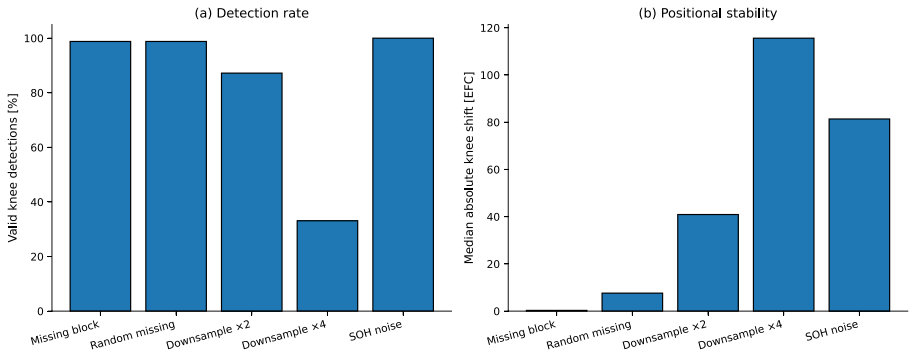


Fig. 9. Sensitivity of robust consensus knee estimation to missing observations, reduced sampling resolution, and additive SOH noise: (a) valid knee-detection rate and (b) median absolute displacement of the estimated knee relative to the clean-data result.

7.4. Knee-point sensitivity

Knee robustness depended strongly on the perturbation type. In the pooled analysis, removing one contiguous 20% block of observations retained a knee estimate in 98.8% of evaluations and produced a median absolute shift of only 0.31 EFC. Random removal of 20% of observations similarly retained detection in 98.8% of evaluations, with a median absolute shift of 7.62 EFC. In contrast, two-fold downsampling reduced the detection rate to 87.2% and produced a median shift of 40.91 EFC, while four-fold downsampling reduced detection to 33.1% and increased the median shift to 115.59 EFC. Adding SOH noise with standard deviation 0.005 retained numerical knee detection in all evaluations, but the median absolute shift was 81.37 EFC. These findings show that the detector is comparatively tolerant of missing observations but not invariant to severe loss of sampling resolution or derivative-amplifying noise.

Fig. 9 separates two aspects of robustness: whether the algorithm returns a valid knee and whether the returned knee remains close to its clean-data position. Additive SOH noise retained a numerical detection in all evaluations but produced a large median displacement. Four-fold downsampling reduced both detection frequency and positional stability. A high detection rate should therefore not be interpreted as evidence of an accurate knee estimate.

7.5. Continuous trajectory reconstruction

Fig. 10 shows only small numerical differences between the evaluated reconstruction models. The neural models obtained validation RMSE values of approximately 0.064–0.065, while the spline baseline obtained an RMSE of approximately 0.067. The result indicates that the normalized capacity trajectories are sufficiently smooth to be represented by several continuous model families. It does not establish a statistically significant advantage of a particular INR architecture.

7.5.1. Sparse voltage–capacity reconstruction audit

An additional per-curve sparse voltage–capacity reconstruction benchmark was conducted on 46 cleaned CALCE single-discharge trajectories. Every model was fitted to the same trajectory-specific training points and evaluated using an identical fixed set of clean held-out points. Table 17 reports the distribution of trajectory-level RMSE values. PCHIP achieved the lowest median held-out error, and paired Wilcoxon comparisons against PCHIP remained significant for every alternative model after Holm correction. These results indicate that densely sampled single-discharge curves are particularly favourable to shape-preserving interpolation. The benchmark is restricted to within-curve reconstruction and does not constitute evidence of cross-cell transfer, cross-domain generalization, or lifetime-prediction superiority of any neural representation.

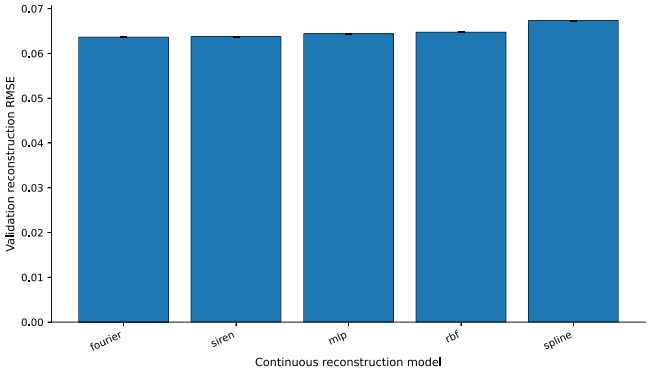
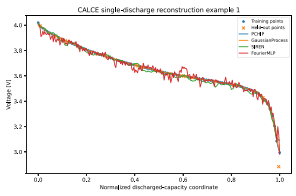


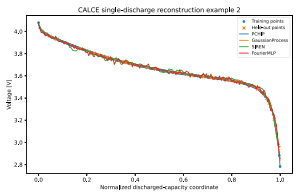
Fig. 10. Validation reconstruction RMSE across the evaluated continuous reconstruction models. Fourier-feature, SIREN, MLP, and RBF networks are coordinate-based neural representations, whereas the spline is included as a conventional continuous baseline.

Table 17
Held-out reconstruction error for the cleaned single-discharge CALCE voltage–capacity benchmark. Errors are reported in volts and summarize trajectory-level RMSE values.

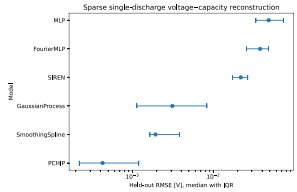
Model	n curves	Median RMSE	IQR	Mean RMSE
PCHIP	46	0.000427	0.000220–0.001199	0.004811
SmoothingSpline	46	0.001929	0.001648–0.003816	0.008706
GaussianProcess	46	0.003116	0.001138–0.008340	0.007567
SIREN	46	0.021822	0.017246–0.026585	0.026126
FourierMLP	46	0.037881	0.025667–0.047897	0.040931
MLP	46	0.048287	0.033634–0.073467	0.053680



(a) Representative curve 1.



(b) Representative curve 2.



(c) Aggregate held-out RMSE.

Fig. 11. Sparse voltage–capacity reconstruction for cleaned CALCE single-discharge curves. Panels (a) and (b) show trajectory-specific training points, identical clean held-out points, and continuous model fits. Panel (c) summarizes median held-out RMSE with the interquartile range across the retained curves. This is a per-curve reconstruction benchmark, not a cross-cell or cross-domain lifetime-prediction experiment.

Fig. 11 complements the aggregate reconstruction comparison with direct curve-level evidence. The shape-preserving baseline tracked densely sampled discharge curves most accurately, whereas neural models exhibited larger interpolation errors. This result motivates conservative interpretation of continuous neural representations: differentiability and flexible parameterization do not imply superiority over classical interpolants for every trajectory type.

7.6. Exploratory Random-Forest attribution and retrospective MC-dropout diagnostic

On the retrospective 244-cell descriptor cohort, the exploratory out-of-fold Random-Forest attribution identified the initial-segment slope and power-law fit RMSE as the two leading descriptors, with mean absolute SHAP contributions of 21.31 and 21.28 EFC, respectively. The next largest contributions were obtained for mean curvature (15.14 EFC), end-segment slope (14.96 EFC), minimum early SOH (14.88 EFC), and linear-fit RMSE (14.37 EFC).

The ranking indicates that the exploratory Random Forest relied on complementary summaries of early trajectory direction, curvature, SOH level, and fit quality. Because several descriptors were correlated functions of the same five-observation prefix, SHAP contributions should not be interpreted as independent or causal effects. Seven cells had already reached the operational EOL threshold within the input prefix; consequently, this attribution analysis is retrospective and is not part of the strict prospective benchmark based on 237 cells. The analysis used internal K-fold validation and does not demonstrate cross-domain or cross-protocol transferability. The complete SHAP summary is shown in Fig. A.5, while the leading numerical attributions are reported in Table A.1).

As summarized in Table Table 18, on the retrospective 244-cell descriptor cohort, the internal MC-dropout diagnostic yielded an RMSE of 243.65 EFC and an MAE of 182.82 EFC. Stochastic prediction dispersion was weakly associated with absolute error (Spearman $\rho_S = 0.166$, $p = 0.0093$), indicating that cells with greater stochastic dispersion tended to be somewhat more difficult for the auxiliary model to predict.

The median MC-dropout standard deviation was only 51.02 EFC, and the nominal Gaussian 90% and 95% bands covered only 32.8% and 39.8% of the held-out targets. MC-dropout therefore substantially underestimated total prediction uncertainty. Because this diagnostic used the retrospective 244-cell cohort, it was not used as an uncertainty estimate for the strict prospective cINR benchmark. Fold-external conformal intervals constructed for the prospectively eligible cINR cohorts remained the principal uncertainty representation. The corresponding out-of-fold diagnostic is presented in Fig. A.6.

Table 18
Retrospective internal four-fold MC-dropout diagnostic on the 244-cell five-observation descriptor cohort. Seven cells had already reached the operational 80% SOH threshold within the input prefix; therefore, this analysis is not part of the strict prospective early-life benchmark. Coverage was calculated using Gaussian mean ± zSD bands and is reported only as a diagnostic of stochastic model dispersion; these bands are not conformal prediction intervals.

n	RMSE	MAE	Median SD	$\rho_{SD,\lvert error\rvert}$	90% coverage	95% coverage
244	243.65	182.82	51.02	0.166	0.328	0.398

7.7. Exploratory materials-level context

After quality control and material-family classification, the ChemDataExtractor analysis retained 28,815 specific-capacity and voltage records. Balancing repeated measurements at the publication level yielded 10,420 DOI-material-family-property summaries.

The median reported gravimetric capacities were 145.75 $\mathrm{mAh\,g^{-1}}$ for LCO, 145.00 $\mathrm{mAh\,g^{-1}}$ for LFP, 171.40 $\mathrm{mAh\,g^{-1}}$ for NMC, 183.00 $\mathrm{mAh\,g^{-1}}$ for NCA, and 128.00 $\mathrm{mAh\,g^{-1}}$ for LMO. The corresponding median voltage values were 4.00, 3.45, 4.30, 3.915, and 4.00 V, respectively.

Substantial within-family dispersion remained, reflecting differences in composition, synthesis, electrode formulation, voltage window, current rate, and experimental reporting across the source literature. These results provide materials-level context only and should not be interpreted as direct estimates of practical cell capacity, state of health, or lifetime. The NCA voltage result should be interpreted particularly cautiously because it was based on only 44 DOI-level summaries.

The auxiliary ChemDataExtractor analysis was conducted at the literature-derived materials level and was not linked to individual CALCE, NASA, or ISU–ILCC cells. Repeated records from the same publication, material family, and property were aggregated before comparison so that publications containing long extracted series did not dominate the summaries. The resulting DOI-balanced distributions were characterized descriptively using medians and interquartile ranges.

Because a single publication could contribute records to more than one material family, and because synthesis, electrode formulation, current rate, voltage window, test protocol, and reporting conventions were not harmonized across the source literature, the between-family differences were not interpreted as independent causal contrasts. Accordingly, Table 19 provides materials-property context rather than evidence about cell SOH, knee position, EOL, or RUL (see Figs. 12 and 13).

8. Discussion

The revised results provide a more qualified assessment of continuous battery-ageing representations than the original analysis. The strongest positive result concerns the geometrical relationship between robust consensus knee location and the first crossing of the 80% SOH threshold. Across 172 cells, the pooled Spearman correlation was $\rho_S = 0.834$, with a 95% bootstrap confidence interval of $[0.738, 0.907]$ and a Holm-adjusted p-value of 3.36×10^{-45}. This supports knee location as an informative full-trajectory degradation descriptor. The domain-specific NASA-randomized estimate was numerically high ($\rho_S = 0.900$) but was based on only five cells and was not statistically significant

Table 19
DOI-balanced descriptive statistics for specific capacity and voltage records extracted from the ChemDataExtractor battery-materials database. Values from repeated records within the same publication, material family, and property were first aggregated by their median. The database is literature-mined and the results provide materials-level context rather than cell-level lifetime measurements.

Material family	$n_{\mathrm{DOI},Q}$	Median capacity [IQR] (mAh g^{-1})	$n_{\mathrm{DOI},V}$	Median voltage [IQR] (V)
LCO	1272	145.75 [134.00–188.50]	1210	4.000 [3.600–4.200]
LFP	2122	145.00 [124.56–163.45]	1575	3.450 [3.350–3.717]
NMC	632	171.40 [147.60–200.00]	568	4.300 [3.700–4.500]
NCA	94	183.00 [167.00–200.00]	44	3.915 [3.587–4.300]
LMO	1281	128.00 [107.00–190.00]	1622	4.000 [3.500–4.400]

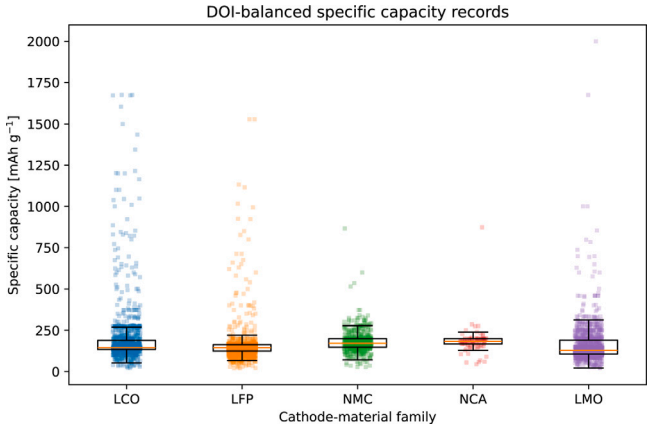


Fig. 12. DOI-balanced distributions of literature-reported gravimetric capacity for the five analysed cathode-material families. One median observation was retained for each DOI-material-family combination. The distributions provide materials-level context and are not cell-lifetime measurements.

under exact two-sided permutation inference ($p_{\mathrm{Holm}} = 0.0833$). It was therefore treated as a descriptive result and was not used as independent statistical confirmation of the pooled association.

The result does not imply that the detected knee always precedes the operational EOL threshold. For 23 cells, the robust consensus knee occurred at or after the first 80% SOH crossing because measurements continued beyond that threshold. Knee position should therefore not be interpreted automatically as an early-warning event. The dedicated sensitivity analyses in Table 16 further separated the full-trajectory association from the effects of post-threshold trajectory extent. They examined the subset in which knee preceded EOL, controlled for validation domain and maximum recorded EFC, and repeated knee detection after truncating trajectories at the interpolated 80% SOH crossing. These analyses strengthen the geometrical interpretation of the association while still not converting it into prospective evidence of advance warning. Moreover, the perturbation analysis showed that derivative-based knee estimates were not uniformly stable: missing observations were comparatively well tolerated, whereas aggressive downsampling and additive noise produced substantial EFC shifts. Continuous differentiability facilitates knee extraction but does not eliminate the numerical sensitivity of second-order descriptors.

The early-life prediction results distinguish representational value from predictive superiority. Raw cINR predictions were strongly dependent on architecture, seed, and observation window. Leakage-safe bias correction and protocol blending markedly improved the raw neural estimates and removed negative RUL predictions. The recovered outer-fold weights reported in Table 5 demonstrate that the bias-corrected cINR estimate retained a non-zero contribution to the final hybrid rather than being discarded completely in favour of the protocol prior (see Fig. A.8). Nevertheless, the paired comparisons in Table 6 did not establish a statistically supported advantage over ProtocolMedianEOL after multiplicity correction. The best primary result was obtained by the MLP hybrid at $N = 10$, with an MAE of 107.79 EFC, but its 7.79 EFC numerical improvement over ProtocolMedianEOL was not statistically significant. Accordingly, the evidence supports cINR as a potentially complementary component of a protocol-informed predictor, not as a uniformly superior standalone prognostic model.

The matched-cohort analysis further clarified the interpretation of the observation-window results. In the primary benchmark, increasing N changed both the amount of available trajectory information and the composition of the eligible cohort. Restricting the comparison to the same $n = 44$ cells removed the latter source of variation. Within this common cohort, the hybrid cINR-MLP MAE decreased from 201.929 EFC at $N = 5$ to 195.728 EFC at $N = 10$ and 112.823 EFC at $N = 20$. The paired comparisons did not identify a statistically supported improvement between $N = 5$ and $N = 10$, but both the $N = 5$ versus $N = 20$ and the $N = 10$ versus $N = 20$ comparisons favoured the longer observation window after Holm correction. Thus, access to the $N = 20$ prefix was associated with lower prediction error for the same cells, whereas the first increase from five to ten observations did not provide a statistically supported improvement.

The ProtocolMedianEOL reference exhibited a similar reduction in matched-cohort MAE, from 185.598 and 182.684 EFC at $N = 5$ and $N = 10$, respectively, to 125.021 EFC at $N = 20$. The matched-cohort improvement should therefore not be attributed uniquely to the cINR component. Window-specific cohort construction, model calibration, and the physical EFC extent represented by each landmark also remain relevant to the interpretation.

The seed-aggregation sensitivity analysis showed that replacing arithmetic-mean aggregation with median aggregation changed hybrid MAE by 0.431, 2.992, and −1.741 EFC for $N = 5$, 10, and 20, respectively. These changes were small relative to the overall prediction errors, indicating that the principal early-life conclusions were robust to the choice of seed-aggregation rule. Arithmetic-mean aggregation was retained as the prespecified primary summary.

The continuous formulation occupies a different methodological position from recent Transformer, recurrent, Gaussian-process, and physics-informed approaches. Transformer and recurrent models can exploit long temporal contexts and large homogeneous datasets, whereas Gaussian processes provide an explicit probabilistic function model and physics-informed networks can impose governing equations or conservation constraints [2,17,18,52]. The present framework instead prioritizes coordinate-continuous reconstruction, direct derivative extraction, and a common perturbation and EFC audit. Because the datasets, targets, validation splits, and lifetime units differ across published studies, the current results should not be interpreted as a numerical head-to-head superiority claim over those model families. Transferable long-term ageing-trajectory prediction incorporating internal resistance and capacity-regeneration information has also been investigated, illustrating the importance of explicitly accounting for cross-condition trajectory variability [53].

The curvature and plateau descriptors also admit a cautious electrochemical interpretation. Changes in voltage-curve slope or plateau extent may reflect altered electrode phase behaviour, loss of active material, impedance growth, or loss of cyclable lithium [46,54]. Accelerated capacity curvature can indicate that one or more degradation processes have entered a nonlinear regime. These descriptors

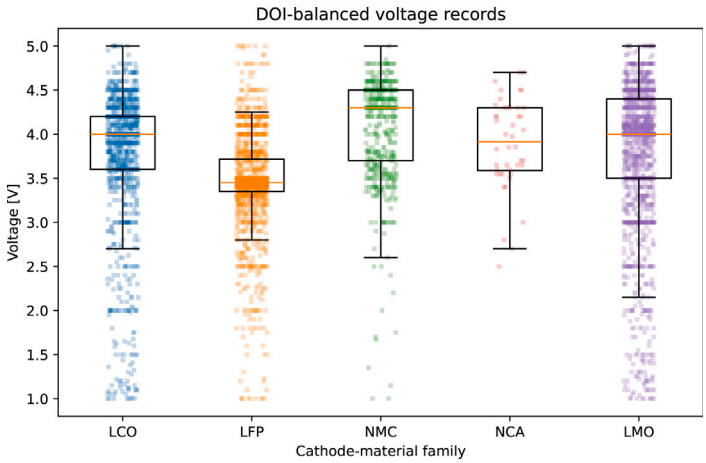


Fig. 13. DOI-balanced distributions of literature-reported voltage for the five analysed cathode-material families. Differences reflect both material chemistry and heterogeneous experimental and reporting conditions across the source literature.

are nevertheless phenomenological: a specific curvature peak cannot be uniquely assigned to a single electrochemical mechanism without complementary measurements such as differential voltage analysis, impedance spectroscopy, reference-electrode measurements, or post-mortem characterization.

The protocol-median results demonstrate that operating conditions contain substantial prior information about lifetime. Consequently, model comparisons that omit a protocol-informed baseline may overstate the incremental value of cell-specific neural adaptation. The current study addresses this issue by reporting raw cINR, protocol-only, and leakage-safe hybrid predictions separately and by testing paired errors rather than relying solely on ranking by mean MAE.

The auxiliary attribution and uncertainty diagnostics clarify two additional aspects of the early-life analysis. First, the Random-Forest SHAP ranking emphasized initial slope, curvature, SOH level, and parametric fit error, indicating that useful predictive information was distributed across several related summaries of the short degradation prefix. This attribution applies to the exploratory Random Forest and not directly to the cINR decoder.

MC-dropout dispersion contained only limited information about cell-specific prediction difficulty. Although its association with absolute error was statistically significant, nominal Gaussian bands showed severe undercoverage. This contrast is practically important: stochastic neural dispersion should not be treated automatically as a calibrated prediction interval. In the present study, fold-external conformal calibration provided substantially more reliable empirical coverage, although its broad widths reflected the underlying lifetime heterogeneity.

For battery-management deployment, several additional requirements remain. Recent in-situ and on-board SOH-estimation studies similarly emphasize the need for deployment-compatible electrical measurements and validation under conditions representative of electromobility applications [55]. Field data contain irregular charging, partial cycles, changing ambient temperature, varying depth of discharge, calendar ageing, sensor drift, maintenance events, and software-dependent state estimation. A deployable system would therefore require online coordinate updating, drift detection, uncertainty-triggered abstention, calibration under fleet shift, and validation against operational rather than laboratory EOL definitions. The broad conformal intervals observed here reinforce the need for risk-aware decision rules instead of a single deterministic lifetime estimate.

9. Limitations

Several limitations constrain the interpretation of this study. The strict protocol-held-out cINR benchmark was restricted to the ISU-ILCC cohort, which supplied the systematic protocol grouping and the number of cells required for the outer-fold evaluation. CALCE and NASA contributed to the EFC lifetime, censoring, knee, reliability, and descriptive cross-domain audits, but they were not included in the strict protocol-held-out cINR prediction benchmark. Their smaller effective sample sizes also prevent equally precise domain-specific conclusions.

The term "observation" does not have an identical physical interpretation across data sources. EFC reduces this ambiguity by expressing lifetime through normalized throughput, but it does not remove differences in chemistry, temperature, current profile, rest periods, calendar ageing, or source preprocessing. Physical cycle counts were not inferred when an explicit mapping was unavailable.

The protocol-held-out folds constitute internal cross-protocol validation rather than independent prospective validation. Similarly, fold-external conformal intervals quantify calibration within the available cohort and should not be interpreted as guarantees for a new fleet, chemistry, or battery-management system.

An exploratory analysis of 21 leakage-safe descriptors extracted from the first five degradation observations identified a two-cluster statistical structure among 244 cells, with a silhouette score of 0.543. Because this analysis was unsupervised and was not independently validated against electrochemical degradation modes, it is reported as descriptive evidence of early-trajectory heterogeneity rather than as a mechanistic classification.

The raw cINR predictors exhibited substantial seed- and architecture-dependent bias. Post-processing reduced this bias, but the resulting hybrid remained strongly influenced by the protocol prior and did not significantly outperform ProtocolMedianEOL. The relatively small $N = 20$ evaluation cohort ($n = 44$) also limited statistical power.

Knee estimates depend on smoothing, derivative calculation, threshold selection, sampling density, and the extent of the available post-threshold trajectory. Although the selected procedure combined curvature- and piecewise-linear criteria with a bootstrap-based stability assessment, the perturbation analysis demonstrated sensitivity to severe downsampling and measurement noise. Moreover, knee location was derived from the complete available trajectory and was not constrained to precede the first 80% SOH crossing; in 23 of 172 analysed cells, the robust knee occurred at or after that operational EOL threshold. The knee-EOL result should therefore be interpreted as a descriptive full-trajectory association rather than as prospective evidence of advance warning. Practical use would require online knee estimation together with an explicit stability or uncertainty measure.

The exploratory embedding, Random-Forest attribution, and MC-dropout diagnostic used the complete 244-cell five-observation descriptor cohort. Seven of these cells had already crossed the operational 80% SOH threshold within the input prefix and were therefore ineligible for the strict prospective $N = 5$ prediction benchmark. These three

auxiliary analyses are consequently retrospective and should not be compared directly with the prospectively eligible cINR and conformal results based on 237 cells.

The Random-Forest and MC-dropout models additionally used internal K-fold validation because the merged descriptor table did not retain a reliable protocol or validation-domain grouping variable. They therefore do not provide cross-protocol or cross-domain evidence. The SHAP ranking explains the behaviour of the exploratory Random Forest and should not be transferred directly to the cINR models, while the MC-dropout dispersion should not be interpreted as calibrated prospective uncertainty.

MC-dropout standard deviation was only weakly associated with absolute error and produced severe undercoverage of held-out EOL values. It therefore captured only part of the model-related uncertainty and did not provide calibrated prediction intervals. Aleatoric uncertainty associated with sensor noise, unrecorded operating conditions, and intrinsic cell variability was not modelled separately. Fold-external conformal intervals partly addressed total residual uncertainty but remained conditional on the present internal study design.

Finally, all datasets were obtained under laboratory or curated experimental conditions. The study did not validate the framework prospectively in an electric vehicle, stationary-storage installation, or field battery-management system. Claims concerning service-life extension, waste reduction, or sustainability therefore describe the potential application of improved prognostics rather than a measured life-cycle environmental benefit.

10. Green approach

This study adopts a green approach by extending lithium-ion battery service life through accurate early-life health prediction and reducing unnecessary battery replacement and waste, aligning primarily with the principles of Design for Energy Efficiency through computationally efficient battery management and Real-time Analysis for Pollution Prevention by enabling early degradation monitoring to prevent premature failure and resource loss [56]. In the present work, this sustainability contribution remains prospective: no formal life-cycle assessment or quantitative greenness assessment was performed, and the environmental benefit would need to be verified in an operational deployment [19].

11. Conclusion

This study developed and audited a continuous battery-ageing framework combining coordinate-based neural representations, EFC harmonization, knee extraction, early-life prediction, perturbation analysis, and conformal uncertainty quantification. Robust consensus knee location was strongly associated with the first crossing of the 80% SOH threshold ($n = 172$, Spearman $\rho_S = 0.834$, 95% bootstrap CI $[0.738, 0.907]$, Holm-adjusted $p = 3.36 \times 10^{-45}$). The corresponding NASA-randomized estimate was based on only five cells and was not statistically significant under exact permutation inference; it was therefore interpreted descriptively. This result supports degradation-transition geometry as an informative full-trajectory descriptor, but does not establish that knee detection necessarily provides prospective advance warning of EOL. The complementary pre-EOL, partial-rank, and EOL-truncated sensitivity analyses further distinguished the association from a relationship driven solely by the extent of the recorded post-threshold trajectory.

The early-life results were more nuanced. Raw cINR performance was architecture- and window-dependent and included substantial systematic bias. Leakage-safe calibration and blending reduced MLP MAE to 113.57, 107.79, and 112.82 EFC for observation windows $N = 5$, 10, and 20, respectively. These hybrid predictions significantly improved the raw MLP cINR estimates for $N = 5$ and $N = 10$, but their numerical improvements over ProtocolMedianEOL were not statistically significant after multiplicity correction. Therefore, the current evidence supports cINR as a complementary continuous representation within a protocol-informed predictor rather than as a universally superior standalone model.

The matched-cohort analysis showed statistically supported reductions in hybrid prediction error when the observation prefix was extended to $N = 20$, but not when it was extended from $N = 5$ to $N = 10$. Because the protocol-informed reference exhibited a similar window-dependent pattern, this result does not establish that the improvement arose uniquely from the cINR component. The seed-aggregation sensitivity analysis produced only small MAE differences between arithmetic-mean and median aggregation, indicating that the main early-life conclusions were robust to the aggregation rule used across the three seeded runs.

Fold-external conformal intervals achieved empirical coverage close to or above their nominal levels, although their large widths reflected substantial cell-level lifetime uncertainty. By contrast, the auxiliary MC-dropout diagnostic produced severe undercoverage and was not retained as the principal uncertainty representation. Exploratory Random-Forest attribution indicated that early slope, curvature, SOH level, and fit quality all contributed to prediction, but this ranking was descriptive and specific to the auxiliary Random-Forest model. Robustness analysis further showed that missing observations were less damaging than aggressive downsampling or noise, particularly for derivative-based knee estimation.

The main contribution is consequently a leakage-controlled framework for separating trajectory representation, protocol information, calibration, uncertainty, and robustness. Future work should evaluate the method prospectively on field battery-management data, explicitly model aleatoric uncertainty, incorporate operational covariate shift, and determine whether continuous trajectory descriptors provide reproducible incremental value across independent chemistries and fleets.

CRediT authorship contribution statement

Agnieszka Pregowska: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stefan Marynowicz:** Writing – review & editing, Writing – original draft, Software, Investigation.

Funding

This research received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Fig. A.1 indicates a two-cluster statistical structure in the retrospective five-observation descriptor space. The clusters were derived without using EOL, RUL, knee, censoring, analysis-time, or other full-trajectory variables and were not used for prediction-model fitting, hyperparameter selection, or evaluation.

Seven of the 244 cells had already crossed the operational 80% SOH threshold within the five-observation prefix. The clustering result is therefore not part of the strict prospective early-life benchmark and should be interpreted only as an exploratory description of descriptor heterogeneity rather than as an independently confirmed electrochemical ageing mechanism.

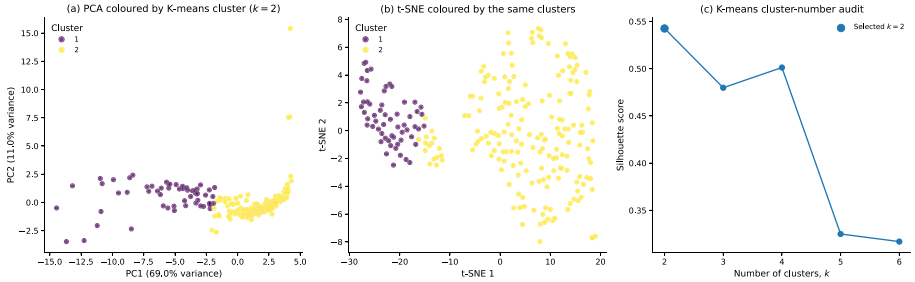


Fig. A.1. Exploratory embedding of 21 standardized leakage-safe early-life descriptors calculated from the first $N = 5$ available degradation observations for 244 cells: (a) PCA projection coloured by K-means cluster; (b) t-SNE projection coloured using the same cluster assignments; and (c) silhouette-based evaluation of the candidate numbers of clusters. K-means clustering was conducted in the complete standardized descriptor space rather than in either two-dimensional projection. EOL, RUL, knee, censoring, analysis-time, and other full-trajectory variables were excluded before clustering. The selected two-cluster solution achieved a silhouette score of 0.543.

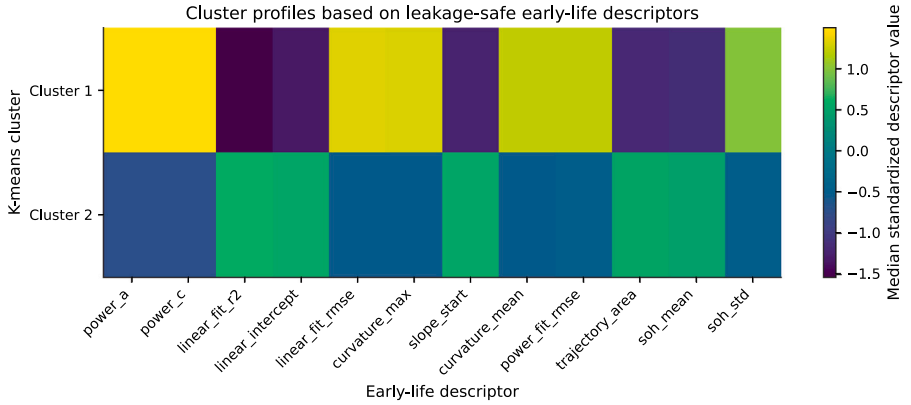


Fig. A.2. Standardized median profiles of the early-life descriptors showing the largest differences between the two K-means clusters. Values are expressed relative to the full-sample mean and standard deviation of each descriptor. The profiles characterize the statistical clustering solution and do not establish distinct physical or electrochemical degradation mechanisms.

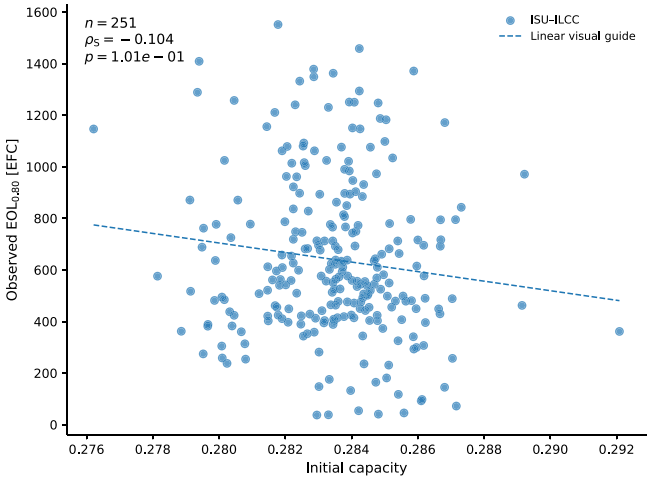


Fig. A.3. Exploratory association between the initial observed capacity q_0 and the first crossing of the 80% SOH threshold in the ISU-ILCC cohort. EOL is expressed in equivalent full cycles. The dashed line is included only as a visual guide. No statistically significant monotonic association was detected ($n = 251$, Spearman $\rho_S = -0.104$, $p = 0.101$). CALCE and NASA-randomized cells were not included because a consistently matched initial-capacity variable was unavailable for those domains.

Fig. A.2 identifies the early-life descriptors that contributed most strongly to the separation of the two exploratory clusters. Because several descriptors summarize related aspects of the same five-observation SOH prefix, the profiles should be interpreted jointly rather than as independent causal effects.

Fig. A.3 shows that initial observed capacity alone was not strongly associated with EOL within the ISU-ILCC cohort. The Spearman correlation was $\rho_S = -0.104$ ($n = 251$, $p = 0.101$), indicating no statistically supported monotonic relationship. This result suggests that between-cell lifetime variability cannot be explained by initial capacity alone and supports the use of trajectory shape, protocol information, and degradation-transition descriptors in the broader analysis. Because consistently matched q_0 values were unavailable for CALCE and NASA-randomized cells, this exploratory result should not be generalized across all evaluated datasets.

Fig. A.4 provides descriptive distributions of the quantities used in the full-trajectory knee analysis. These descriptors were not included in the strict early-life prediction benchmark. In particular, the knee-to-EOL distance uses information from the complete available trajectory and should not be interpreted as an independently available early-life predictor.

Fig. A.7 shows that the effects of perturbation were model dependent. Missing points and downsampling reduced the local support available for interpolation, whereas modest additive voltage noise had a smaller effect on several smooth-function models. These results concern within-curve reconstruction only and are reported separately from early-life EOL-prediction sensitivity.

Data availability

Data will be made available on request.

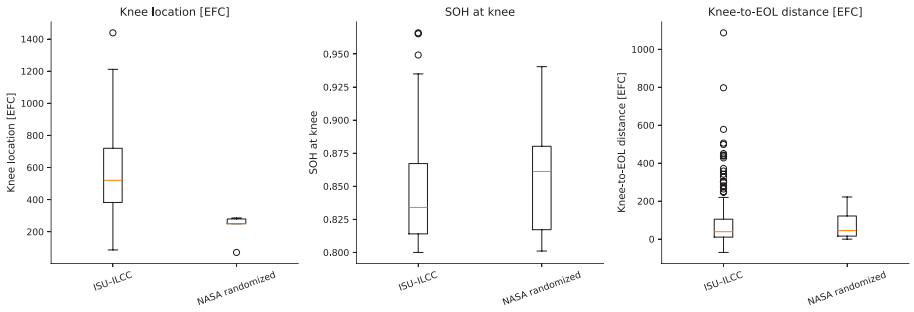


Fig. A.4. Distributions of robust full-trajectory knee descriptors for the ISU-ILCC and NASA-randomized cells: (a) consensus knee location on the EFC scale; (b) SOH at the detected knee; and (c) EFC distance between the detected knee and the first 80% SOH crossing. The NASA-randomized distributions are based on only five cells and should therefore be interpreted cautiously.

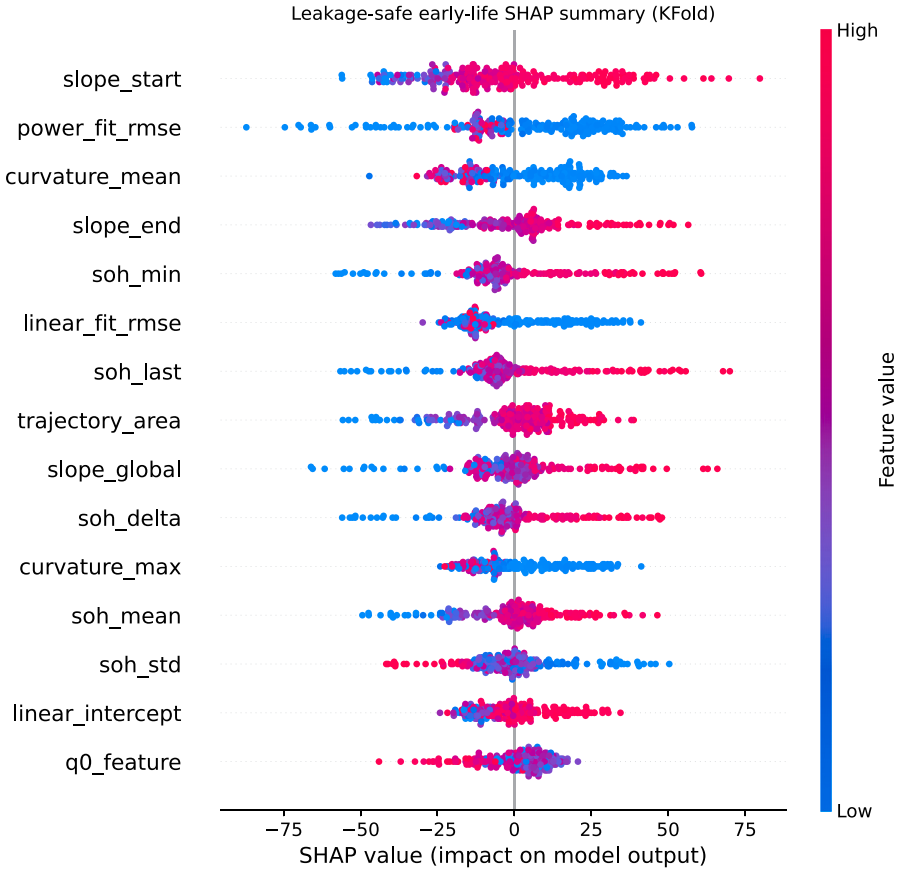


Fig. A.5. Out-of-fold TreeSHAP summary for the retrospective exploratory Random-Forest EOL model using 21 leakage-safe descriptors calculated from the first $N = 5$ available degradation observations in the 244-cell descriptor cohort. Seven cells had already reached the operational 80% SOH threshold within the input prefix; therefore, this attribution analysis is not part of the strict prospective early-life benchmark. Colour indicates the standardized descriptor value, whereas horizontal position represents the SHAP contribution to predicted EOL on the EFC scale. The figure explains the exploratory Random-Forest model and should not be interpreted as direct attribution of the cINR predictor.

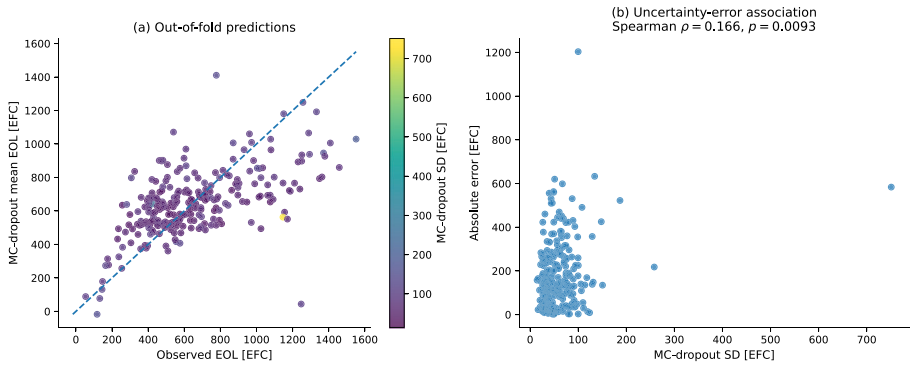
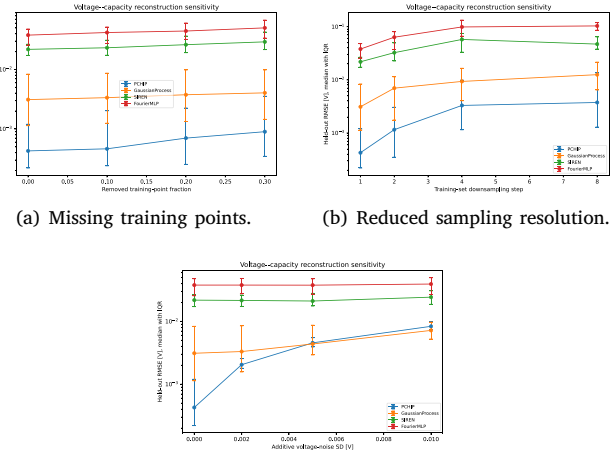


Fig. A.6. Retrospective internal out-of-fold MC-dropout diagnostic using 21 leakage-safe descriptors calculated from the first $N = 5$ available degradation observations in the 244-cell descriptor cohort: (a) observed EOL versus mean stochastic prediction, with colour representing MC-dropout standard deviation; and (b) association between MC-dropout standard deviation and absolute prediction error. Seven cells had already reached the operational EOL threshold within the input prefix; therefore, the analysis is not part of the strict prospective benchmark. The stochastic dispersion was weakly associated with error but severely under-calibrated and was not used as the principal uncertainty interval.



(a) Missing training points. (b) Reduced sampling resolution.

(c) Additive voltage noise.

Fig. A.7. Sensitivity of sparse single-discharge voltage–capacity reconstruction to missing training observations, reduced training-set resolution, and additive voltage noise. Points denote median held-out RMSE and bars denote the interquartile range. The logarithmic ordinate permits comparison of classical and neural reconstruction models over their different error scales.

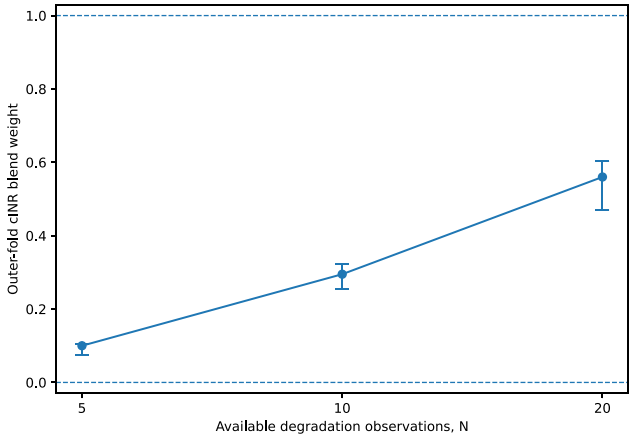


Fig. A.8. Outer-fold cINR contribution to the final MLP–protocol hybrid. Points show the median blending weight and error bars denote the interquartile range across outer folds. A weight of zero corresponds to the protocol prior alone, whereas a weight of one corresponds to the bias-corrected cINR prediction alone.

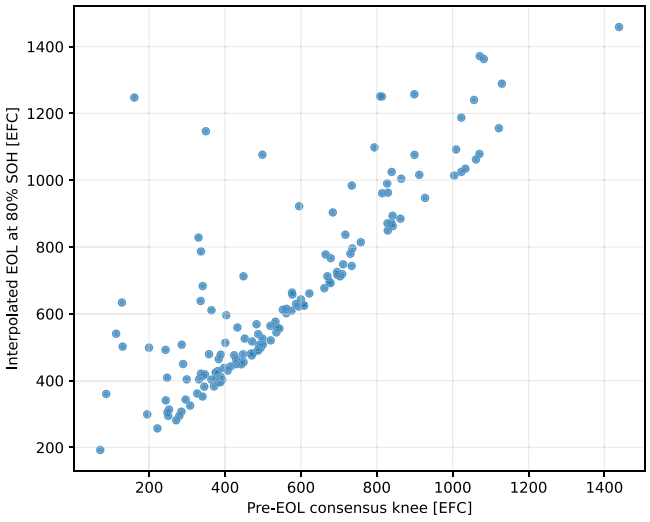


Fig. A.9. Association between consensus knee location and interpolated EOL after excluding cells whose full-trajectory knee occurred at or after the 80% SOH crossing.

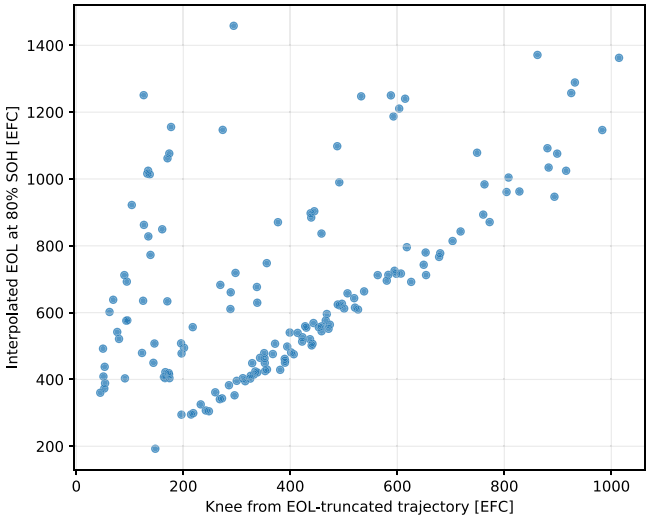


Fig. A.10. Association between EOL and knee location recalculated after truncating each trajectory at its linearly interpolated first 80% SOH crossing.

Table A.1
Leading out-of-fold SHAP attributions for the retrospective exploratory Random-Forest EOL model applied to the 244-cell five-observation descriptor cohort. Seven cells had already reached the operational EOL threshold within the input prefix; therefore, this analysis is not part of the strict prospective benchmark. Values are mean absolute SHAP contributions expressed in EFC.

Descriptor	Mean absolute SHAP [EFC]
Initial-segment slope	21.31
Power-law fit RMSE	21.28
Mean curvature	15.14
End-segment slope	14.96
Minimum early SOH	14.88
Linear-fit RMSE	14.37
Last early SOH observation	14.05
Trajectory area	12.90
Global slope	12.60
Early SOH change	12.34

References

[1] von Bulow F, Hahn Y, Meyes R, Meisen T. Transparent and interpretable state of health forecasting of lithium-ion batteries with deep learning and saliency maps. Int J Energy Res 2023;9922475. http://dx.doi.org/10.1155/2023/9922475.

[2] Hu J, Fu P, Wei Z, et al. Early prediction of lithium-ion battery degradation with a generative pre-trained transformer. Nat Commun 2026;17:126. http://dx.doi.org/10.1038/s41467-025-66819-0.

[3] Zhang Guangxu, Wei Xuezhe, Chen Siqi, Han Guangshuai, Zhu Jiangong, Dai Haifeng. ACS Appl Energy Mater 2022;5(5):6462–71. http://dx.doi.org/10.1021/acsaem.2c00957.

[4] Nazeeruddin MA, Li R, OKane SE, Marinescu M, Offer GJ. Lithium-ion battery degradation: Introducing the concept of reservoirs to design for lifetime. 2025, arXiv preprint https://doi.org/10.48550/arXiv.2512.15440.

[5] Chen L, Xu C, Bao X, et al. State-of-health estimation of Lithium-ion battery based on back-propagation neural network with adaptive hidden layer. Neural Comput Appl 2023;35:14169–82. http://dx.doi.org/10.1007/s00521-023-08471-7.

[6] Pregowska A, Osial M, Urbanska W. The application of artificial intelligence in the effective battery life cycle in the closed circular economy model- A perspective. Recycling 2022;7:81. http://dx.doi.org/10.3390/recycling7060081.

[7] Meng H, Li YF. A review on prognostics and health management (PHM) methods of lithium-ion batteries. Renew Sustain Energy Rev 116:109405. http://dx.doi.org/10.1016/j.rser.2019.109405.

[8] Berecibar M, Gandiaga I, Villarreal I, Omar N, Van Mierlo J, Van den Bossche P. Critical review of state of health estimation methods of Li-ion batteries for real applications. Renew Sustain Energy Rev 2016;56(C):572–87. http://dx.doi.org/10.1016/j.rser.2015.11.042, Elsevier.

[9] How DNT, Hannan MA, Hossain Lipu MS, Ker PJ. State of charge estimation for lithium-ion batteries using model-based and data-driven methods: A review. IEEE Access 2019;7:136116–36. http://dx.doi.org/10.1109/ACCESS.2019.2942213.

[10] Hu X, Li S, Peng H. A comparative study of equivalent circuit models for Li-ion batteries. J Power Sources 2012;198:359–67. http://dx.doi.org/10.1016/j.jpowsour.2011.10.013.

[11] Jokar A, Rajabloo B, Desilets M, Lacroix M. Review of simplified Pseudo-two-Dimensional models of lithium-ion batteries. J Power Sources 2016;327:44–55. http://dx.doi.org/10.1016/j.jpowsour.2016.07.036.

[12] Chen Z, Sun M, Shu X, Shen J, Xiao R. On-board state of health estimation for lithium-ion batteries based on random forest. In: 2018 IEEE international conference on industrial technology. ICIT, 2018, p. 175–1759. http://dx.doi.org/10.1109/ICIT.2018.8352448.

[13] Xiao Z, Jiang B, Zhu J, Wei X, Dai H. State of health estimation for lithium-ion batteries using an explainable XGBoost model with parameter optimization. Batteries 2024;10:394. http://dx.doi.org/10.3390/batteries10110394.

[14] Al-Rahamneh A, Izco I, Serrano-Hernandez A, Faulin J. Machine learning-based state-of-health estimation of battery management systems using experimental and simulation data. Mathematics 2025;13:2247. http://dx.doi.org/10.3390/math13142247.

[15] Lee G, Kim J, Lee C. State-of-health estimation of Li-ion batteries in the early phases of qualification tests: An interpretable machine learning approach. Expert Syst Appl 2022;197:C. http://dx.doi.org/10.1016/j.eswa.2022.116817.

[16] Wei Y, Wu D. Prediction of state of health and remaining useful life of lithium-ion battery using graph convolutional network with dual attention mechanisms. Reliab Eng Syst Saf 2023;230:108947. http://dx.doi.org/10.1016/j.ress.2022.108947.

[17] Thelen A, Huan X, Paulson N, et al. Probabilistic machine learning for battery health diagnostics and prognostics-review and perspectives. npj Mater Sustain 2:14. http://dx.doi.org/10.1038/s44296-024-00011-1.

[18] Dietrich F, Schilders W. Scientific machine learning. Math Semesterber 2025;72:89–115. http://dx.doi.org/10.1007/s00591-025-00399-4.

[19] Fegade SL. Ssessment of greenness of catalytic deoxygenation of crop oil for green diesel production. Clean Circ Bioeconomy 2024;8:100091. http://dx.doi.org/10.1016/j.clcb.2024.100091.

[20] Fegade SL. Green hydrocarbons and fuels from municipal polymer waste co-fed with natural gas using a batch catalytic slurry process. Green Technol Sustain 2024;2(3):100099. http://dx.doi.org/10.1016/j.grets.2024.100099.

[21] Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. DeepSDF: Learning continuous signed distance functions for shape representation. In: IEEE/CVF conference on computer vision and pattern recognition. CVPR, Long Beach, CA, USA; 2019, p. 165–74. http://dx.doi.org/10.1109/CVPR.2019.00025.

[22] Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. Occupancy networks: Learning 3D reconstruction in function space. In: IEEE/CVF conference on computer vision and pattern recognition. CVPR, Long Beach, CA, USA; 2019, p. 4455–65. http://dx.doi.org/10.1109/CVPR.2019.00459.

[23] Sitzmann V, Martel J, Bergman A, Lindell D, Wetzstein G. Implicit neural representations with periodic activation functions. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems. Curran Associates, Inc.; 2020, 33. http://dx.doi.org/10.48550/arXiv.2006.09661, 7462–7473.

[24] Tancik M, Srinivasan PP, Mildenhall B, Fridovich-Keil S, Raghavan N, Singhal U, Ramamoorthi R, Barron JT, Ng R. Fourier features let networks learn high frequency functions in low dimensional domains. In: Proceedings of the 34th international conference on neural information processing systems. NIPS'20, Red Hook, NY, USA: Curran Associates Inc.; 2020, 632. http://dx.doi.org/10.48550/arXiv.2006.10739, 7537–7547.

[25] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: representing scenes as neural radiance fields for view synthesis. Commun ACM 2021;65(1):99–106. http://dx.doi.org/10.1145/3503250.

[26] Pregowska A, Larecki W, Szczepanski J. Application of neural networks for determine the radiation pressure in two-moment radiation hydrodynamics in slab geometry. Comput Assist Methods Eng Sci 2025;1–24. http://dx.doi.org/10.24423/cames.2025.1960.

[27] Xu D, Ma S, Ji X, Han X, Lin J, Liu K. A shapley additive explanations-based data-driven approach for lithium-ion battery state of health estimation using ultrasound technology. Energy Technol 2025;13:2500673. http://dx.doi.org/10.1002/ente.202500673.

[28] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd international conference on international conference on machine learning 48. ICML'16, JMLR.org; 2016, p. 1050–9.

[29] Kendall A, Gal A. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st international conference on neural information processing systems. NIPS'17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 5580–90. http://dx.doi.org/10.48550/arXiv.1703.04977.

[30] Xuan QL, Adhisantoso YG, Munderloh M, Ostermann J. Uncertainty-aware remaining useful life prediction for predictive maintenance using deep learning. Procedia CIRP 2023;118. http://dx.doi.org/10.1016/j.procir.2023.06.021, 116-112.

[31] Kuleshov V, Fenner N, Ermon S. Accurate uncertainties for deep learning using calibrated regression. 2018, http://dx.doi.org/10.48550/arXiv.1807.00263, ArXiv, abs/1807.00263.

[32] Mouais T, Kittaneh OA, Majid MA. Choosing the best lifetime model for commercial lithium-ion batteries. J Energy Storage 2021;41:102827. http://dx.doi.org/10.1016/j.est.2021.102827.

[33] Madani SS, Shabeer Y, Allard F, Fowler M, Ziebert C, Wang Z, Panchal S, Chaoui H, Mekhilef S, Dou SX, et al. A comprehensive review on lithium-ion battery lifetime prediction and aging mechanism analysis. Batteries 2025;11:127. http://dx.doi.org/10.3390/batteries11040127.

[34] Lu Yu, Zhou Sida, Zhou Xinan, Yang Shichun, Liu Mingyan, Liu Xinhua, Ling Heping, Lian Yubo. A novel method of prediction for capacity and remaining useful life of lithium-ion battery based on multi-time scale Weibull accelerated failure time regression. J Energy Storage 2023;68:107589. http://dx.doi.org/10.1016/j.est.2023.107589.

[35] Chu Ch-H, Lee Ch-J, Yeh H-Y. Developing deep survival model for remaining useful life estimation based on convolutional and long short-term memory neural networks. Wirel Commun Mob Comput 2020;8814658. http://dx.doi.org/10.1155/2020/8814658.

[36] Hu W, Qian Q. Lithium-ion battery state of health and failure analysis with mixture weibull and equivalent circuit model. iScience 2024;27(6):109980. http://dx.doi.org/10.1016/j.isci.2024.109980.

[37] dos Reis G, Strange C, Yadav M, Li S. Lithium-ion battery data and where to find it. Energy AI 2021;5:100081. http://dx.doi.org/10.1016/j.egyai.2021.100081.

[38] https://data.nasa.gov/dataset/li-ion-battery-aging-datasets.

[39] https://calce.umd.edu/battery-data.

[40] https://iastate.figshare.com/articles/dataset/_b_ISU-ILCC_Battery_Aging_Dataset_b_/22582234.

[41] https://doi.org/10.25380/iastate.22582234.

[42] https://doi.org/10.6084/m9.figshare.11888115.

[43] Rashid M, Faraji-Niri M, Sansom J, Sheikh M, Widanage D, Marco J. Dataset for rapid state of health estimation of lithium batteries using EIS and machine learning: Training and validation. Data Brief 2023;48:109157. http://dx.doi.org/10.1016/j.dib.2023.109157.

[44] Severson KA, Attia PM, Jin N, et al. Data-driven prediction of battery cycle life before capacity degradation. Nat Energy 2019;4:383–91. http://dx.doi.org/10.1038/s41560-019-0356-8.

[45] Attia PM, Grover A, Jin N, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature 2020;578:397–402. http://dx.doi.org/10.1038/s41586-020-1994-5.

[46] Feng X, Weng C, He X, Wang L, Ren D, Lu L, Han X, Ouyang M. Incremental capacity analysis on commercial lithium-ion batteries using support vector regression: A parametric study. Energies 2018;11:2323. http://dx.doi.org/10.3390/en11092323.

[47] Song Z, Zhang H, Jia J. Data-driven state of health interval prediction for lithium-ion batteries. Electronics 2024;13:3991. http://dx.doi.org/10.3390/electronics13203991.

[48] Diao W, Saxena S, Han B, Pecht M. Algorithm to determine the knee point on capacity fade curves of lithium-ion cells. Energies 2019;12:2910. http://dx.doi.org/10.3390/en12152910.

[49] Attia PM, Bills A, Planella FB, Dechent P, Dos Reis G, Dubarry M, Gasper P, Gilchrist R, Greenbank S, Howey D, Liu O. “Knees” in lithium-ion battery aging trajectories. J Electrochem Soc 2022;169(6). http://dx.doi.org/10.1149/1945-7111/ac6d13.

[50] You H, Zhu J, Wang X, Jiang B, Wei X, Dai H. Nonlinear aging knee-point prediction for lithium–on batteries faced with different application scenarios. eTransportation 2023;18:100270. http://dx.doi.org/10.1016/j.etran.2023.100270.

[51] Jia X, Zhang C, Li Y, et al. Knee-point-conscious battery aging trajectory prediction of lithium-ion based on physics-guided machine learning. IEEE Trans Transp Electrif 2024;10(1):1056–69. http://dx.doi.org/10.1109/TTE.2023.3266386.

[52] Wei M, Gu H, Ye M, Wang Q, Xu X, Wu C. Remaining useful life prediction of lithium-ion batteries based on Monte Carlo Dropout and gated recurrent unit. Energy Rep 2021;7:2862–71. http://dx.doi.org/10.1016/j.egyr.2021.05.019.

[53] Huang Y, Zhang P, Lu J, Xiong R, Cai Z. A transferable long-term lithium-ion battery aging trajectory prediction model considering internal resistance and capacity regeneration phenomenon. Appl Energy 2024;360:122825. http://dx.doi.org/10.1016/j.apenergy.2024.122825.

[54] Lv Zhe, Si Huinan, Yang Zhe, Cui Jiawen, He Zhichao, Wang Lei, Li Zhe, Zhang Jianbo. Simplified mechanistic aging model for lithium ion batteries in large-scale applications. Materials 2025;18(6):1342. http://dx.doi.org/10.3390/ma18061342.

[55] Bustos JEG, Schiele BB, Baldo L, Masserano B, Jaramillo-Montoya F, Troncoso-Kurtovic D, Orchard ME, Perez A, Silva JF. In situ estimation of li-ion battery state of health using on-board electrical measurements for electromobility applications. Batteries 2025;11:451. http://dx.doi.org/10.3390/batteries11120451.

[56] Fegade SL. Red chemistry: Principles and applications. Next Sustain 2024;4:100048. http://dx.doi.org/10.1016/j.nxsust.2024.100048.