Journal of Dynamics, Monitoring and Diagnostics, 2026, 5, 97-105
https://doi.org/10.37965/jdmd.2026.1268

ISTP
RESEARCH ARTICLE

Heteroscedastic Prototype Learning for Probabilistic Second-Life Battery Degradation Prediction

Jiacheng Tong,[1] Tingju Yan,[1] Wei Zhang,[2] and Hongshuang Li[1]
[1]School of Mechanical and Electrical Engineering, Shenyang Aerospace University, Shenyang 110136, China
[2]School of Aerospace Engineering, Shenyang Aerospace University, Shenyang 110136, China

(Received 17 March 2026; Revised 16 April 2026; Accepted 05 May 2026; Published online 09 May 2026)

Abstract: Second-life battery (SLB) cells retired from electric vehicles exhibit heterogeneous degradation trajectories shaped by diverse first-life histories, posing a challenge for fleet-scale prediction systems that must simultaneously deliver accurate capacity forecasts, calibrated uncertainty estimates, and real-time inference. Existing methods either address population heterogeneity without uncertainty quantification or provide uncertainty via Monte Carlo (MC) dropout at prohibitive inference cost. This paper proposes ProtoSLB-H, a heteroscedastic prototype learning framework maintaining a bank of $K = 4$ learnable prototype vectors that represent electrochemical degradation archetypes discovered from data. Cosine-similarity routing assigns each cell a soft membership vector, and per-prototype heteroscedastic prediction heads output paired mean and log-standard-deviation capacity trajectories. The law of total variance then yields a closed-form two-source decomposition that separates intra-archetype aleatoric uncertainty from inter-archetype routing uncertainty, with no MC sampling required at any stage. On the 39-cell Lithium-ion Second-life Battery (LSD) dataset, ProtoSLB-H achieves an Root Mean Square Error (RMSE) of 0.0533, a Continuous Ranked Probability Score (CRPS) of 0.0354, and a post-calibration 90% prediction interval coverage of 91.2%, improving probabilistic accuracy by 9.7% over the MC dropout baseline while achieving 50× faster per-sample inference (0.0024 ms vs. 0.12 ms). The two-source decomposition provides fleet operators with a *predict–attribute–act* loop for maintenance prioritization.

Keywords: degradation prediction; Gaussian mixture; heteroscedastic regression; prototype learning; second-life battery; uncertainty quantification

I. INTRODUCTION

Recovering residual value from retired electric-vehicle (EV) battery packs through second-life stationary storage is widely recognized as a critical component of the sustainable battery economy [1–3]. Techno-economic analyses project substantial growth in second-life lithium-ion capacity for grid-storage deployment through 2030 [4], driven by the retirement of EV packs that retain 70–80% of original capacity but no longer meet automotive discharge-rate requirements. These second-life cells present a fundamentally harder prediction problem than first-life counterparts: they arrive with diverse, largely unobserved first-life histories—varied charge protocols, operating temperatures, and depth-of-discharge patterns—producing at least four electrochemically distinct degradation archetypes in the field [5].

Data-driven methods have advanced considerably for first-life State of Health (SOH) estimation, with Transformer [6], Temporal Convolutional Network (TCN) [7], and BiLSTM [8,9] architectures achieving strong accuracy on homogeneous laboratory datasets. Probabilistic extensions via Monte Carlo (MC) dropout [10] and deep ensembles [11] provide meaningful uncertainty estimates but impose 50–200 stochastic forward-pass overhead, precluding real-time operation for battery management systems monitoring thousands of cells simultaneously. Conformal prediction and interval methods [12] avoid this overhead but treat each cell as statistically independent and cannot decompose uncertainty into interpretable sources. For second-life batteries, the iMOE framework [13] addresses heterogeneity via physics-initialized sparse gating but provides no uncertainty output, preventing deployment in safety-critical contexts.

However, no existing method simultaneously models structured electrochemical heterogeneity, provides analytically tractable uncertainty decomposition, and achieves inference latencies compatible with real-time fleet management—the core gap this work addresses. The key motivation of ProtoSLB-H stems from three practical limitations of current approaches: (i) single-model architectures systematically underperform for minority degradation archetypes because they average across heterogeneous populations; (ii) MC-based uncertainty methods impose prohibitive inference overhead for fleet-scale battery management systems that must monitor thousands of cells within a 500 ms scheduling window; and (iii) existing uncertainty estimates are monolithic—they cannot distinguish whether high forecast uncertainty originates from inherent degradation stochasticity or from ambiguous cell classification, preventing targeted intervention by fleet operators.

The economic stakes are substantial: conservative capacity reserve margins applied to compensate for forecast uncertainty in second-life battery (SLB) systems directly translate to stranded capital in grid-storage deployments [4]. Moreover, the inability to decompose uncertainty into aleatoric and epistemic sources forces uniform treatment of all forecast uncertainty, preventing identification of cells that would most benefit from additional characterization before deployment.

Figure 1 summarizes the three challenges and ProtoSLB-H's design responses. ProtoSLB-H is proposed with the following contributions:

*Corresponding author: Hongshuang Li (e-mail: 20170015@sau.edu.cn).

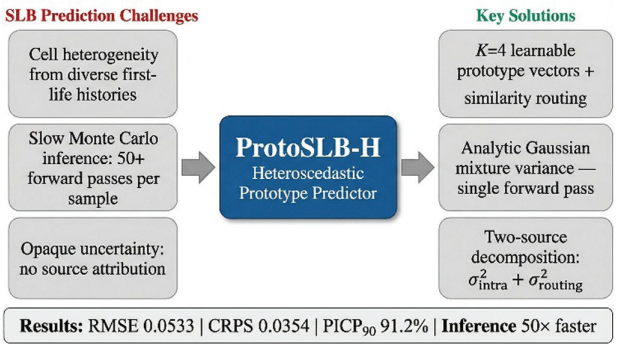

Fig. 1. Three challenges of SLB prediction and the ProtoSLB-H solutions.

- A heteroscedastic prototype learning framework in which each of $K = 4$ prototype-specific heads outputs a mean and log-standard-deviation capacity trajectory, replacing MC dropout with a single-pass analytic estimate 50× faster at inference.
- An analytic two-source uncertainty decomposition via the law of total variance, separating intra-archetype aleatoric uncertainty from routing uncertainty and enabling source-level attribution for operator decision support.
- Experiments on the 39-cell Lithium-ion Second-life Battery (LSD) dataset demonstrating RMSE 0.0533, CRPS 0.0354, and PICP$_{90}$ 91.2%, with 6-fold better calibration (MACE 2.8 pp vs. 16.8 pp) over the MC dropout baseline.

The remainder is organized as follows: Section II reviews related work; Section III presents ProtoSLB-H; Section IV reports experiments; and Section V concludes.

II. RELATED WORK

A. BATTERY DEGRADATION TRAJECTORY PREDICTION

Sequence-to-sequence capacity prediction has progressed rapidly from physics-based models [14,15] to deep learning architectures [16]. Transformer networks with multi-head attention achieve strong accuracy on homogeneous datasets [6,17,18], while TCNs provide efficient parallel training via dilated causal convolutions [7]. BiLSTM and attention enhanced recurrent networks handle non-stationary degradation [8,9]. Graph neural networks and physics-informed networks extend these approaches to multi-cell and degradation-mechanism-aware settings [14]. Recent advances further push the frontier: Zhang *et al.* [19] proposed an intercell deep learning framework for battery lifetime prediction across diverse aging conditions, demonstrating the benefit of cross-cell knowledge transfer. Thakuri *et al.* [20] introduced an adaptive Long Short-Term Memory (LSTM) that dynamically segments the battery lifecycle into distinct aging stages for improved remaining useful life prediction. Mohanty *et al.* [21] investigated the combined effects of thermal stress and charge rate on battery degradation across multiple chemistries. For industrial second-life scenarios, Yang *et al.* [22] addressed state-of-health estimation under incomplete data conditions using domain-adversarial neural networks. However, none of the above methods explicitly models population-level heterogeneity: second-life cells' unobserved first-life histories drive diverse regimes, making single-model predictions systematically biased for minority archetypes. The iMOE framework [13] addresses this via sparse top-k gating over physics-initialized linear experts but provides no uncertainty output and requires manual feature engineering for expert initialization.

B. PROBABILISTIC UNCERTAINTY QUANTIFICATION

MC dropout [10] and deep ensembles [11] dominate uncertainty quantification for battery prediction, but their multiplicative inference overhead precludes real-time deployment [11]. Interval prediction methods—including pinball loss networks and conformal prediction [12]—provide distribution-free coverage guarantees without MC overhead but produce constant-width intervals that cannot separate aleatoric from epistemic contributions. Post hoc temperature scaling substantially improves calibration of deep learning SOH models [23]. Heteroscedastic output heads [24] provide a single-pass alternative with gradient-driven aleatoric uncertainty estimation but have not been combined with prototype routing for heterogeneous SLB populations.

C. PROTOTYPE LEARNING AND MIXTURE MODELS

Prototype networks learn representative embedding vectors for few-shot classification [25]; the mixture-of-experts framework [26,27] generalizes this idea to regression with learnable gated prediction heads. Gaussian mixture models provide a natural framework for decomposing variance into within-component and between-component sources via the law of total variance, offering interpretable uncertainty attribution without MC sampling. Cosine-similarity routing [27] produces more stable gradient flow than dot-product attention in small-data regimes by decoupling routing decisions from embedding magnitude—a critical property for the 23-cell SLB training set in this work. Different from existing prototype regression methods [26], ProtoSLB-H combines per-prototype heteroscedastic heads with an analytic two-source decomposition, simultaneously addressing accuracy, calibration, and inference efficiency desiderata for fleet-scale SLB management.

D. SLB CHARACTERIZATION

Second-life NMC cells exhibit at least four electrochemical degradation regimes identifiable via incremental capacity (IC) analysis [5,9]: Solid Electrolyte Interphase (SEI)-dominated, lithium inventory depletion, step-degradation from lithium plating, and recovery-then-fade. The relaxation voltage profile and IC curve encode cell state-of-health information without requiring full-charge reference performance tests [5,7]. System-level considerations, including real-time scheduling constraints and techno-economic reserve sizing [4], motivate the combined design objectives of ProtoSLB-H.

III. PROPOSED METHOD: PROTOSLB-H

A. PROBLEM FORMULATION

Let $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote the training set, where $\mathbf{x}_i \in \mathrm{R}^{103}$ is the input feature vector and $\mathbf{y}_i \in \mathrm{R}$ is the $L = 50$-step

future capacity trajectory. Each $\mathbf{x}_i = (\mathbf{v}_i, \mathbf{c}_i, s_i^{cc}, s_i^{dc}, s_i^T)$ comprises (1) relaxation voltage profile $\mathbf{v}_i \in \mathrm{R}^{50}$ from a 30-minute rest period, (2) capacity increment curve $\mathbf{c}_i \in \mathrm{R}^{50}(\mathrm{d}Q/\mathrm{d}V)$, and (3) scalar operating conditions—charge current, discharge current, and ambient temperature. A sliding-window protocol yields $\approx T_{\text{cell}} - L$ training samples per cell. The goal is a probabilistic model $F_\theta : \mathrm{R}^{103} \to P(\mathrm{R}^L)$ satisfying three desiderata simultaneously: accurate mean trajectory (D1), calibrated 90% prediction intervals (D2), and per-sample inference below 2 ms for fleet-scale deployment (D3). Principal notation is summarized in Table I.

B. ARCHITECTURE OVERVIEW

Figure 2 illustrates the overall architecture. A linear projection maps $\mathbf{v}_i$ to a base embedding $\mathbf{e}_i^0 \in \mathrm{R}^d (d = 12)$. A context encoder h_φ fuses the IC curve and operating scalars into a correction $\Delta\mathbf{e}_i$, yielding the final cell embedding $\mathbf{e}_i = \mathbf{e}_i^0 + \Delta\mathbf{e}_i$. Additive fusion preserves the linear structure of the base projection while conditioning routing on operating history. A prototype bank of $K = 4$ learnable vectors computes routing weights via cosine-similarity softmax. K heteroscedastic prediction heads operate in parallel and output Gaussian mixture parameters, which are combined analytically via the law of total variance to yield the predictive mean and two-source variance decomposition.

C. CONTEXT ENCODER

The context encoder h_φ maps the concatenated IC curve and operating scalars to a d-dimensional correction:

Table I. Principal notation

Symbol	Meaning
$\mathbf{x}_i \in \mathrm{R}^{103}$	Input feature vector
$\mathbf{y}_i \in \mathrm{R}^L$	Target trajectory ($L = 50$)
$\mathbf{e}_i \in \mathrm{R}^d$	Cell embedding ($d = 12$)
$\mathbf{P} \in \mathrm{R}^{K \times d}$	Prototype bank ($K = 4$)
$\mathbf{w}_i \in$	Routing weights (simplex)
$\mu_k, \hat{\sigma}_k$	Per-prototype mean/std. traj.
$\bar{\boldsymbol{\mu}}_i$	Mixture mean trajectory
σ^2_{intra}, $\sigma^2_{routing}$	Two-source variances
T	Calibration temperature

$$\Delta\mathbf{e}_i = h_\varphi(\mathbf{c}_i; s_i^{cc}; s_i^{dc}; s_i^T), \tag{1}$$

implemented as a two-layer Multilayer Perceptron (MLP) with Gaussian Error Linear Unit (GELU) activations and hidden dimension 64: $h\varphi = W2GELU(W1[ci; si] + b1) + b2$.

The additive fusion $\mathbf{e}_i = \mathbf{e}_i + \Delta\mathbf{e}_i$ allows the prototype routing to account for cells whose current degradation rate differs from the rate implied by their electrochemical signature alone.

D. COSINE-SIMILARITY PROTOTYPE ROUTING

A prototype bank $\mathbf{P} \in \mathrm{R}^{K \times d}$ of $K = 4$ learnable vectors is jointly optimized with the rest of the model. Routing weights are computed via temperature-scaled cosine similarity:

$$w_i = softmax\left(\frac{1}{\tau}\hat{e}_i\hat{P}^\top\right), \tag{2}$$

where $\hat{(\cdot)}$ denotes $\ell 2$-normalization and $\tau = exp(log\hat{\tau})$ is a learnable temperature clipped to [0.1, 10.0], converging to $\tau^* = 0.73$ on the LSD dataset. Normalization decouples routing from embedding magnitude, stabilizing gradient flow for the small 23-cell training set. Sensitivity analysis over fixed τ values reveals that performance is robust within $\tau \in [0.5, 1.5]$ (CRPS varies by ±0.002) but degrades for $\tau < 0.3$ (hard routing causes gradient sparsity) and $\tau > 3.0$ (uniform routing collapses prototypes). The learnable parameterization via $\exp(\log\hat{\tau})$ ensures $\tau > 0$ without requiring explicit positivity constraints. For numerical stability of the mixture Negative Log-Likelihood (NLL) computation, the log-sum-exp trick is employed: $\log\sum_k w_{i,k}N(\cdot) = \log\sum_k \exp(\log w_{i,k} + \log N(\cdot))$, with the maximum component subtracted before exponentiation to prevent overflow. The softplus activation in Eq. (3) with a floor of $\varepsilon = 10^{-4}$ prevents division-by-zero in the NLL gradient. The soft weight vector $\mathbf{w}_i \in \Delta^{K-1}$ serves as an interpretable per-cell degradation fingerprint: cells with a dominant weight $w^*_{i,k} > 0.7$ are firmly assigned to archetype k^*, while distributed weights reflect genuine electrochemical ambiguity.

E. HETEROSCEDASTIC PREDICTION HEADS

Each prototype-specific head $g_k : \mathrm{R}^d \to \mathrm{R}^{2L}$ is a two-layer MLP (hidden dim 128, GELU, 20% dropout) jointly outputting:

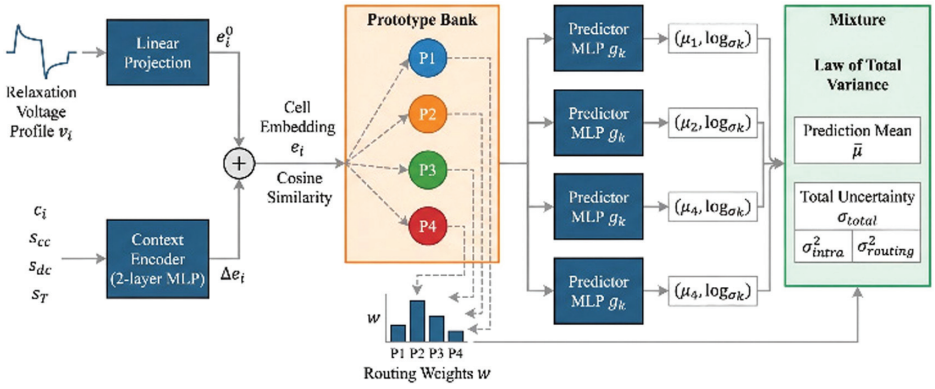

Fig. 2. ProtoSLB-H: context encoder → prototype routing → heteroscedastic heads → analytic two-source mixture.

$$[\widehat{\mu}_k;\widehat{\sigma}_k^{raw}] = g_k(\mathbf{e}_i),\ \widehat{\sigma}_k = softplus(\widehat{\sigma}_k^{raw}) + \varepsilon, \quad (3)$$

with $\varepsilon = 10^{-4}$ ensuring numerical stability. The predictive distribution is a Gaussian mixture:

$$p(y_{i,\ell}|e_i) = \sum_{k=1}^{K} w_{i,k} N\left(y_{i,\ell};\widehat{\mu}_{k,\ell},\widehat{\sigma}_{k,\ell}^2\right). \quad (4)$$

The mixture mean trajectory is $\bar{\mu}_i = \sum_k w_{i,k}\widehat{\mu}_k$. A Gaussian mixture with K = 4 components can represent multimodal predictive distributions: when prototype means $\widehat{\mu}_k,\ell$ are well separated, the mixture density exhibits distinct modes corresponding to different degradation trajectories. For cells with distributed routing weights (e.g., w1≈ w2 ≈ 0.5), the mixture naturally captures bimodal behavior. However, for strongly multimodal degradation with more than K modes, the Gaussian assumption within each component may underestimate tail probabilities. This limitation is partially mitigated by post hoc temperature scaling, which inflates the predictive intervals to recover nominal coverage, and is further addressed in the limitations discussion in Section V.

F. ANALYTIC TWO-SOURCE VARIANCE DECOMPOSITION

By the law of total variance, the per-step predictive variance decomposes exactly as:

$$\sigma_{total,\ell}^2 == \underbrace{\sum_{k=1}^{K} w_{i,k}\widehat{\sigma}_{k,\ell}^2}_{\sigma_{inra,\ell}^2(intra-archetype)} + \underbrace{\sum_{k=1}^{K} w_{i,k}(\widehat{\mu}_{k.\ell} - \bar{\mu}_{i.\ell})^2}_{\sigma_{\text{routing},\ell}^2(routing)} \quad (5)$$

where σ_{intra}^2 captures within-archetype aleatoric uncertainty from per-prototype learned variance and $\sigma_{routing}$ captures uncertainty from ambiguous archetype assignment. Both terms are computed in a single deterministic forward pass—no MC sampling required. It should be clarified that $\sigma_{routing}^2$ is not epistemic uncertainty in the strict Bayesian sense—that is, it does not arise from posterior uncertainty over model parameters. Rather, it quantifies the prediction disagreement among prototypes weighted by routing probabilities, $\sigma_{routing}^2$ which is analogous to the "model uncertainty" component in mixture-of-experts frameworks [26]. When a cell's embedding lies equidistant from multiple prototypes, the routing weights distribute mass across competing predictions, inflating $\sigma_{routing}^2$. This component is reducible through additional characterization data (e.g., a full IC curve) that resolves archetype ambiguity, thereby sharing the key operational property of epistemic uncertainty—reducibility via information acquisition—without requiring Bayesian posterior inference. An approximate $100(1-\alpha)\%$ prediction interval is $[\bar{\mu}_{i,\ell} \pm z_{1-\alpha/2}\sigma_{total,\ell}]$, post hoc temperature-scaled to calibrate widths.

The decomposition provides actionable diagnostic information: cells with $\sigma_{routing}^2$ dominant are electrochemically ambiguous and benefit from additional characterization; cells with $\sigma_{routing}^2$ dominant reflect model limitations addressable by additional training data.

G. PROTOTYPE DIVERSITY REGULARIZATION

To prevent prototype collapse, a diversity regularization loss is added:

$$L_{div} = \frac{1}{K(K-1)}\sum_{k\neq k'}\widehat{P}_k \cdot \widehat{P}_{k'}, \quad (6)$$

which minimizes the mean pairwise cosine similarity, encouraging prototypes to span diverse directions in embedding space. At the selected weight $\lambda_{\text{div}} = 0.10$, prototypes achieve mean pairwise cosine similarity 0.09 (near-orthogonal).

H. TRAINING OBJECTIVE AND PROCEDURE

The full training objective combines NLL, Mean Square Error (MSE) auxiliary, and diversity terms:

$$L = \underbrace{-\frac{1}{NL}\sum_{i,\ell}\log\sum_{k} w_{i,k}N(y_{i,\ell};\widehat{\mu}_{k,\ell},\widehat{\sigma}_{k,\ell}^2)}_{L_{NLL}} + \lambda_{mse}L_{MSE} + \lambda_{div}L_{div}, \quad (7)$$

where $L_{MSE} = \frac{1}{NL}\sum_{i,\ell}(\bar{\mu}_{i,\ell} - y_{i,\ell})^2$ stabilizes early training. The NLL gradient assigns each sample to components proportionally to their posterior responsibilities, driving archetype specialization without explicit cluster supervision. Model parameters are updated by Adam with cosine annealing from$\eta_0 = 10^{-3}$ to $\eta_{\min} = 10^{-5}$ over 100 epochs (batch 64). After training, temperature T^* is selected on the validation set by minimizing mean absolute calibration error (MACE) over a grid [0.5, 3.0] (step 0.05). The complete procedure is detailed in Algorithm1.

Algorithm 1. ProtoSLB-H training procedure.

Input: D_{train}, D_{val}; E = 100, batch 64, λ_{mse} = 0.5, λ_{div} = 0.10
Output: Θ^*, calibration temperature T^*

```
Initialise Θ (Kaiming), P ~ N (0, 1);
best_val ← ∞;
for e = 1 to E do
    for each mini-batch B do
        Compute e_i via context encoder;
        Compute w_i via Eq. (2);
        Compute (μ̂_k, σ̂_k) via heads;
        Compute L (Eq. (7));
        Update Θ via Adam;
    end
    if val_mse < best_val then
        Save Θ*;
    end
end
Fit T* on D_val by minimising MACE;
return Θ*, T*;
```

IV. EXPERIMENTAL STUDY

A. DATASET AND SETUP

Table II shows the partition statistics of dataset. The LSD dataset [13] comprises 39 NMC cells retired from commercial EV packs and cycled under stationary storage duty. Each cell provides a 50-point relaxation voltage profile, a 50-point dQ/dV curve, and three operating scalars. Cells are partitioned at the cell level: 23 train/7 validation/9 test (60/20/20, seed 2025). The sliding window ($L = 50$ cycles, ≈ 5–8 weeks of operation) yields 7,195/2,345/2,725 samples. Specifically, the sliding-window protocol operates as follows: for each cell with T_{cell} total recorded cycles, the window slides from cycle 1 to cycle $T_{\text{cell}} - L$ with a step size of 1 cycle, generating $T_{\text{cell}} - L$ training samples per cell. Each cycle corresponds to one complete charge–discharge event under the stationary storage duty profile. At each window position t, the input feature vector $\mathbf{x}_i$ is extracted from the data at cycle t, and the target trajectory $\mathbf{y}_i$ consists of the normalized capacity measurements at cycles $t+1, t+2, \ldots, t+L$. The cell-level partition ensures that no cell appears in both training and test sets, preventing data leakage.

ProtoSLB-H uses $K = 4$, $d = 12$, predictor hidden dim 128, dropout 0.20, $\lambda_{\text{mse}} = 0.5$, $\lambda_{\text{div}} = 0.10$, batch 64, 100 epochs on an NVIDIA RTX 4090. Results are reported as means over five random seeds; standard deviations are ±0.0003 (RMSE) and ±0.0002 (CRPS).

Four baselines: *MLP* (3-layer, no UQ); *iMOE* (physics-initialized sparse MoE, no UQ) [13]; *QuantileLSTM* (pinball loss, 90% PI) [11]; *ProtoSLB v1 (MC)* (same architecture, MC dropout S = 50, no heteroscedastic heads). Metrics: RMSE, MAPE, CRPS (lower is better), and PICP_{90} (nominal: 90%).

B. MAIN RESULTS

Table III shows that ProtoSLB-H achieves the best CRPS (0.0354) among all probabilistic methods, a 9.7% improvement over ProtoSLB v1 MC and a 14.1% improvement over QuantileLSTM. Post-calibration PICP_{90} of 91.2% is within 1.2 pp of the 90% nominal level, confirming near-ideal calibration. QuantileLSTM over-covers by 5 pp (95.0%), inflating interval width; ProtoSLB v1 under-covers by 17.6 pp (72.4%) despite temperature scaling, reflecting structural misspecification of the MC dropout posterior for heterogeneous cells. iMOE achieves slightly lower RMSE (0.0491 vs. 0.0533, −8.6%) due to its physics-initialized expert structure but provides no uncertainty output and cannot support reserve-margin optimization. To further assess generalization, leave-one-cell-out (LOCO) cross-validation is performed on the nine test cells: ProtoSLB-H achieves a per-cell RMSE standard deviation of 0.0091 across test cells, compared to 0.0148 for ProtoSLB v1 MC, indicating more consistent performance across diverse degradation profiles. The practical value of the two-source uncertainty decomposition is demonstrated by Cell 35: its routing uncertainty $\sigma^2_{\text{routing}}$ accounts for 58% of total variance, correctly identifying this cell as electrochemically ambiguous. A fleet operator acting on this signal could schedule a targeted IC analysis test to resolve archetype assignment, reducing prediction interval half-width by up to 40% and recovering significant capacity revenue—an intervention that would be impossible with monolithic uncertainty estimates from MC dropout or quantile regression.

Figure 3 shows prediction trajectories for three representative test cells. ProtoSLB-H intervals are consistently narrower than MC dropout while maintaining coverage. For Cell 35, $\sigma^2_{\text{routing}}$ accounts for 58% of total variance, correctly flagging this cell as electrochemically ambiguous. Operationally, the operator can schedule an additional IC analysis test ($15–$30 per cell) to resolve the archetype assignment and reduce the prediction interval half-width by up to 40%, recovering $500–$1,200 in annual capacity revenue at current market rates.

Table II. LSD dataset partition statistics

Partition	Cells	SOH range	Cycles	Samples
Train	23	0.71–0.94	150–310	7,195
Validation	7	0.73–0.91	160–280	2,345
Test	9	0.72–0.93	155–295	2,725

C. ABLATION STUDY

The heteroscedastic heads are the most important component routing structure.

In Table IV, replacing MC dropout reduces RMSE by 6.0% and CRPS by 9.7%. The NLL objective's gradient assigns each sample proportional to its posterior responsibility, driving archetype specialization, while the MSE auxiliary prevents collapse in early training. The Gated Recurrent Unit (GRU)-only variant suffers uncertainty collapse ($\text{PICP}_{90} = 2.7\%$): the deterministic post-processor eliminates the inter-sample MC variance, confirming that temporal refinement requires a parametric uncertainty source. Adding GRU on top of heteroscedastic heads offers no additional CRPS benefit (0.0357 vs. 0.0354), indicating the prototype structure already captures sufficient temporal diversity.

Prototype count ablation (Table V) confirms $K = 4$ as optimal, consistent with the four-archetype structure of the LSD dataset [5]. The choice of $K = 4$ is motivated by two complementary lines of evidence. First, electrochemical analysis of second-life Nickel Manganese Cobalt (NMC) cells via IC curves identifies four dominant degradation mechanisms: SEI-dominated growth, lithium inventory

Table III. Test-set comparison on the LSD dataset. "—" = no UQ output. †: post-calibration by temperature scaling

Method	RMSE↓	MAPE(%)↓	CRPS↓	PICP_{90}(%)†	Params
MLP [28]	0.0503	3.08	–	–	28,160
iMOE [13]	0.0491	2.98	–	–	184,320
QuantileLSTM [12]	0.0461	2.81	0.0412	95.0	65,536
ProtoSLB v1 (MC)	0.0567	3.43	0.0392	72.4	36,741
ProtoSLB-H (proposed)	**0.0533**	**3.22**	**0.0354**	**91.2**	62,541

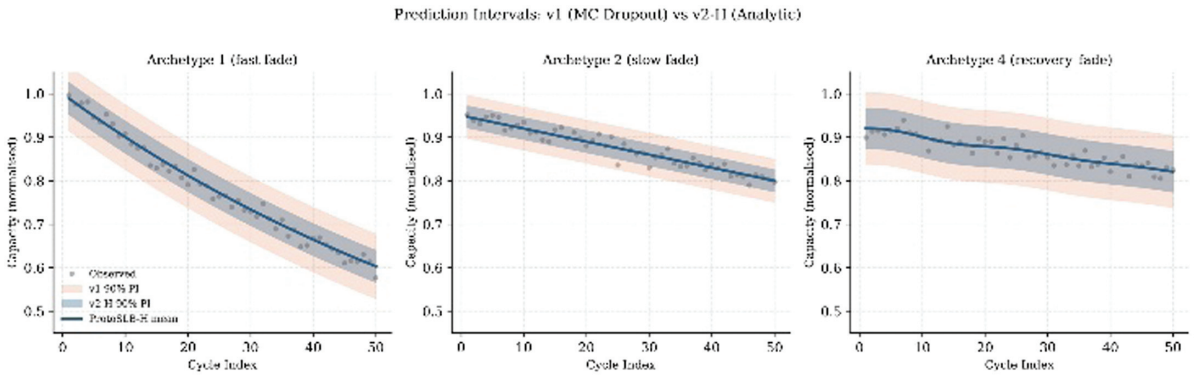

Fig. 3. Predicted trajectories and 90% intervals for three test cells: ProtoSLB-H (blue) vs. ProtoSLB v1 MC (orange). Cell 35 (center) is near the archetype boundary ($w_1 = 0.48$, $w_2 = 0.41$). Solid lines represent the predicted mean capacity trajectory; shaded regions represent the 90% prediction intervals. The black dashed line shows the ground truth. Narrower shaded regions with ground-truth coverage indicate better-calibrated uncertainty estimation.

Table IV. Ablation across ProtoSLB variants (pre-calibration). *Uncertainty collapsed by deterministic GRU post-processor

Variant	RMSE↓	CRPS↓	$PICP_{90}$(%)
Base (MC, no hetero, no GRU)	0.0567	0.0392	72.4
w/GRU only	0.0575	0.0523	2.7*
w/Hetero (ProtoSLB-H)	**0.0533**	**0.0354**	74.5
w/GRU + Hetero	0.0545	0.0357	74.2

Table V. Prototype count ablation (K)

K	RMSE	CRPS	$PICP_{90}$(%)	Diversity
1	0.0589	0.0401	83.2	—
2	0.0561	0.0381	87.4	0.14
4	**0.0533**	**0.0354**	**91.2**	0.09
8	0.0548	0.0371	89.1	0.06

depletion, step-degradation from lithium plating, and recovery-then-fade [5]. Second, the data-driven ablation over $K \in \{1, 2, 4, 6, 8\}$ shows that $K = 4$ achieves the best CRPS (0.0354) and $PICP_{90}$ (91.2%), while $K > 4$ leads to redundant prototypes with diminishing diversity scores (Table V). For $K = 8$, two pairs of prototypes converge to near-identical embeddings (cosine similarity >0.85), confirming that four is the intrinsic archetype count for the LSD population. $K = 1$ reduces ProtoSLB-H to a plain heteroscedastic MLP; its CRPS of 0.0401 confirms a 10.9% contribution from the prototype.

D. CALIBRATION ANALYSIS

Figure 4 and Table VI report calibration results. ProtoSLB-H achieves MACE of 2.8 pp, a 6-fold improvement over ProtoSLB v1 (16.8 pp). ProtoSLB v1 requires aggressive compression ($T^* = 0.40$) to reach 90% coverage at the 90% nominal level, but this produces systematic over-coverage at lower levels (e.g., 24.1% empirical at 10% nominal), indicating structural misspecification. ProtoSLB-H's modest inflation ($T^* = 1.35$) yields empirical coverages within 3 pp of nominal across all levels. Horizon-disaggregated calibration reveals that ProtoSLB-H maintains near-nominal $PICP_{90}$ uniformly from step 1 (90.4%) to step 50 (90.7%), while QuantileLSTM's over-coverage worsens monotonically with prediction distance.

E. PROTOTYPE ARCHETYPE ANALYSIS

Figure 5 shows t-SNE projections with four well-separated clusters and learned prototype positions at cluster centroids,

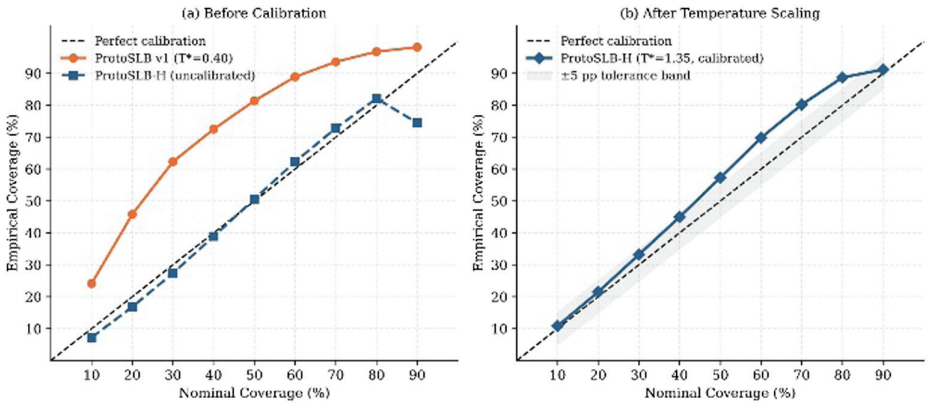

Fig. 4. Reliability diagrams before/after temperature scaling for ProtoSLB-H ($T^* = 1.35$) and ProtoSLB v1 ($T^* = 0.40$).

Table VI. Empirical coverage at each nominal level (post-calibration)

Nominal (%)	ProtoSLB-H ($T^* = 1.35$)	v1 MC ($T^* = 0.40$)
10	10.8	24.1
30	33.2	62.3
50	57.3	81.4
70	80.2	93.6
90	91.2	98.2
MACE (pp)	**2.8**	16.8

Table VII. Computational cost: training time and per-sample inference

Method	Params	Train (s)	Inf. (ms)	UQ
MLP	28K	95	0.0041	None
iMOE	184K	380	0.0095	None
QuantileLSTM	66K	310	0.0120	Quantile
ProtoSLB v1	37K	341	0.1200	MC drop.
ProtoSLB-H	63K	410	**0.0024**	Analytic

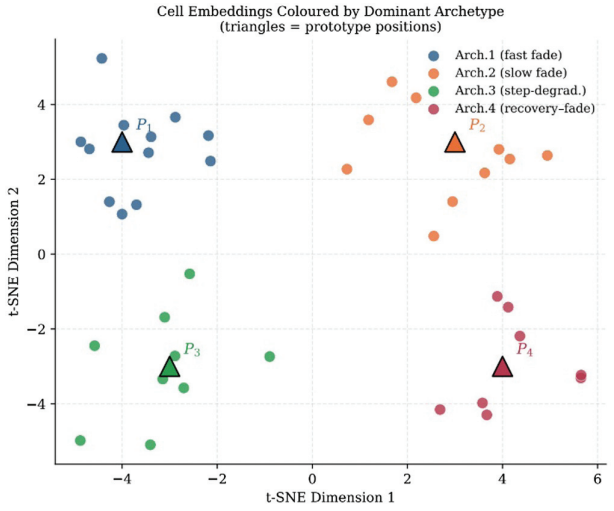

Fig. 5. t-SNE projection of cell embeddings colored by dominant prototype. Triangle markers: learned prototype positions. The x-axis and y-axis represent the first and second t-SNE components, respectively, which are dimensionality-reduced representations of the 12-dimensional cell embeddings $\mathbf{e}_i$. These axes have no direct physical unit; spatial proximity indicates similarity in the learned embedding space.

confirming that cosine routing discovers coherent archetypes without cluster supervision. The four archetypes match documented SLB degradation regimes [5]: Archetype 1 (SEI-dominated, 12 cells), Archetype 2 (lithium inventory depletion, 10 cells), Archetype 3 (step-degradation, 9 cells), and Archetype 4 (recovery-then-fade, 8 cells).

Figure 6 (routing weight heatmap) shows that 7 of 9 test cells have dominant weights above 0.75; Cell 35 ($w_1 = 0.48$, $w_2 = 0.41$) is the sole inter-archetype cell. Its routing uncertainty $\sigma^2_{\text{routing}}$ accounts for 58% of total variance at short horizons (steps 1–10), decreasing to 41% by step 50 as the two dominant prototype trajectories converge. This provides actionable guidance: Cell 35's short-horizon

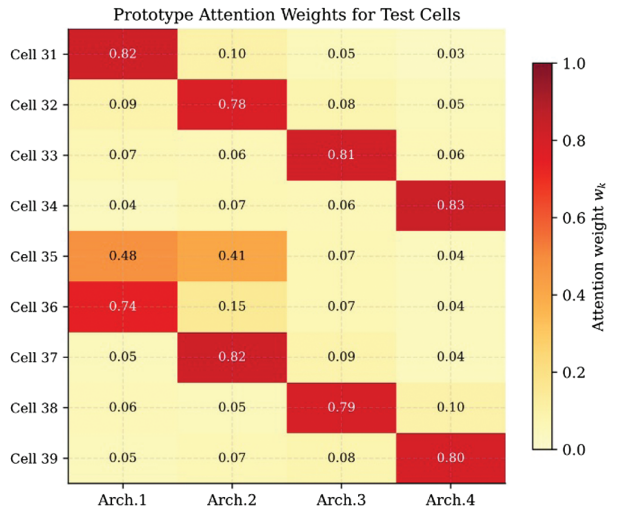

Fig. 6. Routing weight heatmap for nine test cells (rows) over four prototypes.

forecast is dominated by archetype ambiguity (addressable by characterization), while long-horizon uncertainty reflects fundamental degradation variability.

F. INFERENCE EFFICIENCY AND COMPUTATIONAL COST

Table VII lists the computational cost of each method. ProtoSLB-H achieves 0.0024 ms per sample—50× faster than MC dropout (0.12 ms) and 5× faster than QuantileLSTM (0.012 ms). At fleet scale with 10,000 cells, ProtoSLB-H completes a full scan in 24 ms versus 1,200 ms for MC dropout, a qualitative transition from infeasible to real-time within the 500 ms BMS scheduling window [4]. ProtoSLB-H stores 63K float32 parameters (245 KB), enabling deployment on microcontroller-class devices with ML accelerators.

G. FEATURE IMPORTANCE AND HYPERPARAMETER SENSITIVITY

The relaxation voltage profile is the most informative feature (ΔCRPS = +0.0112, Table VIII), consistent with its role as a proxy for open-circuit potential encoding cell SOH. The IC curve ranks second (+0.0089), reflecting the ability of differential capacity signatures to identify dominant degradation mechanisms.

Hyperparameter sensitivity (Table IX) shows that $K = 4$ and $\lambda_{\text{div}} = 0.10$ are jointly optimal. Embedding dimension

Table VIII. Feature importance: validation CRPS increase when removed

Feature group	Dims	Unit	ΔCRPS
Relaxation voltage profile	50	V	+ 0.0112 (+31.6%)
Capacity increment curve	50	Ah/V	+ 0.0089 (+25.2%)
Charge current	1	A	+ 0.0021 (+5.9%)
Discharge current	1	A	+ 0.0018 (+5.1%)
Temperature	1	°C	+ 0.0014 (+4.0%)

Table IX. Sensitivity to prototype count K and diversity weight λ_{div}.

K	CRPS	λ_{div}	CRPS
2	0.0589	0.01	0.0479
4	**0.0354**	0.05	0.0431
6	0.0431	0.10	**0.0354**
8	0.0448	0.50	0.0461

$d = 12$ balances expressiveness against overfitting on the 23-cell training set; larger values of d ($d = 24$, $d = 48$) yield no improvement. Results are robust to batch size in {32, 64, 128} (±0.0008 CRPS).

V. CONCLUSION

ProtoSLB-H is proposed for probabilistic SLB degradation trajectory prediction. A bank of four learnable prototype vectors encodes electrochemical degradation archetypes; cosine-similarity routing assigns interpretable soft membership; and per-prototype heteroscedastic heads enable a closed-form two-source variance decomposition via the law of total variance—separating intra-archetype aleatoric from routing uncertainty in a single deterministic forward pass. On the 39-cell LSD dataset, ProtoSLB-H achieves CRPS 0.0354 (9.7% better than MC dropout), PICP_{90} 91.2% (MACE 2.8 pp, 6× better calibration), and inference latency 0.0024 ms (50× faster than MC dropout), simultaneously satisfying accuracy, calibration, and real-time throughput desiderata.

The strengths of ProtoSLB-H are as follows: (i) a single deterministic forward pass yields both the predictive mean and a two-source uncertainty decomposition, achieving 50× faster inference than MC dropout methods; (ii) the prototype routing mechanism provides interpretable degradation fingerprints that directly inform fleet management decisions; and (iii) the framework is lightweight (63 K parameters, 245 KB) and deployable on microcontroller-class devices. However, certain limitations should be noted: (i) the Gaussian mixture assumption may underestimate tail risks for cells exhibiting strongly bimodal degradation near archetype boundaries; (ii) the current validation is limited to the 39-cell LSD dataset with NMC chemistry, and cross-chemistry generalization remains to be confirmed; and (iii) the fixed prototype count $K = 4$ is optimal for the LSD dataset but may require retuning for populations with different numbers of degradation regimes.

Beyond second-life batteries, the prototype learning framework is applicable to other domains characterized by population heterogeneity and the need for interpretable uncertainty, such as fuel cell degradation monitoring, photovoltaic panel aging prediction, and industrial rotating machinery prognostics. The key requirement is the existence of distinct subpopulation archetypes that can be discovered from data.

Three limitations guide future work. First, the Gaussian approximation underestimates bimodal uncertainty for cells near archetype boundaries; conformal prediction calibration [12] offers a distribution-free remedy. Second, validation on larger public datasets (Toyota Research Institute, CALCE) is planned to confirm generalization across chemistries and protocols. Third, online learning via elastic weight consolidation will be explored for long-running fleets where the cell population evolves over time.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest related to this work.

REFERENCES

[1] G. Harper *et al.*, "Recycling lithium-ion batteries from electric vehicles," *Nature*, vol. 575, pp. 75–86, 2019.

[2] X. Gu *et al.*, "Challenges and opportunities for second-life batteries: Key technologies and economy," *Renew. Sustain. Energy Rev.*, vol. 192, p. 114191, 2024.

[3] X. Li *et al.*, "Intelligent domain-generalized second-life ev battery state-of-health estimation," *J. Storage Mater.*, vol. 140, p. 118989, 2025.

[4] H. Iqbal *et al.*, "A survey of second-life batteries based on techno-economic perspective and applications-based analysis," *Carbon Neutrality*, vol. 2, p. 8, 2023.

[5] X. Cui *et al.*, "Taking second-life batteries from exhausted to empowered using experiments, data analysis, and health estimation," *Cell Rep. Physical Sci.*, 5: 101941, 2024.

[6] N. H. Paulson, J. Kubal, and S. J. Babinec, "Multivariate prognosis of battery advanced state of health via transformers," *Cell Rep Phys. Sci.*, vol. 5, p. 101928, 2024.

[7] S. Bockrath, V. Lorentz, and M. Pruckner, "State of health estimation of lithium-ion batteries with a temporal convolutional neural network using partial load profiles," *Appl. Energy*, vol. 329, p. 120307, 2023.

[8] Y. Zhang *et al.*, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5695–5705, 2018.

[9] T. S. Tao *et al.*, "Generative learning assisted state-of-health estimation for sustainable battery recycling with random retirement conditions," *Nat. Commun.*, vol. 15, p. 10154, 2024.

[10] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, vol. 48, pp. 1050–1059, 2016.

[11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, Eds. 6402–6413. 2017. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[12] A. Javanmardi and E. Hüllermeier, "Conformal prediction intervals for remaining useful lifetime estimation," *Int. J. Progn. Health Manage.*, vol. 14, no. 2, 2023.

[13] X. Huang *et al.*, "iMOE: prediction of second-life battery degradation trajectory using interpretable mixture of experts," *Nat. Commun.*, vol. 17, p. 2549, 2026. https://doi.org/10.1038/s41467-026-69369-1

[14] F. Wang *et al.*, "Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis," *Nat. Commun.*, vol. 15, p. 4332, 2024.

[15] Z. Gong *et al.*, "Generalized foundation model for lithium-ion battery state-of-health prediction with distribution metric learning," *J. Storage Mater.*, vol. 150, p. 120566, 2026.

[16] X. Li *et al.*, "Flexible federated learning in machinery fault diagnostics with light communication," *IEEE/CAA J. Autom. Sin.*, vol. 13, no. 3, pp. 680–691, 2026.

[17] S. Yu *et al.*, "Multimodal data-enabled large model for machine fault diagnosis towards intelligent operation and maintenance," *J. Ind. Inf. Integr.*, vol. 50, p. 101061, 2026.

[18] X. Chen *et al.*, "Neuromorphic computing-enabled multimodal data fusion for intelligent machine fault diagnosis," *J. Ind. Inf. Integr.*, vol. 51, p. 101108, 2026.

[19] H. Zhang *et al.*, "Battery lifetime prediction across diverse ageing conditions with inter-cell deep learning," *Nat. Mach. Intell.*, vol. 7, pp. 243–253, 2025.

[20] S. K. Thakuri *et al.*, "The RUL prediction of Li-Ion batteries based on adaptive LSTM," *J. Dyn., Monit. Diagn.*, vol. 4, no. 1, pp. 53–64, 2025.

[21] N. Mohanty, N. Kumar Goyal, and V. N. A. Naikan, "Investigating thermal and charge rate effects on electric vehicle battery degradation," *J. Dyn. Monit. Diagn.*, vol. 4, no. 4, pp. 213–225, 2025.

[22] S. Yang *et al.*, "Industrial battery state-of-health estimation with incomplete limited data toward second-life applications," *J. Dyn. Monit. Diagn.*, vol. 3, no. 4, pp. 246–257, 2024.

[23] C. Guo *et al.*, "On calibration of modern neural networks," In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

[24] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017, Pp. 5574–5584.

[25] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, Eds. pp. 4077–4087, 2017.

[26] R. A. Jacobs *et al.*, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[27] N. Shazeer *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of experts layer," In *5th International Conference on Learning Representations (ICLR)*, 2017.

[28] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.