Physics-Informed Multi-Task Learning for Battery State of Health Prediction with Uncertainty Quantification

Tianwen Zhu[1], Guangyu Wu[1], Zhiwei Cao[1], Ruihang Wang[1], Jimin Jia[1], Yong Luo[2], Yonggang Wen[1,*]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]School of Computer Science, Wuhan University, China
{tianwen001, zhiwei003}@e.ntu.edu.sg, {guangyu.wu, ruihang.wang, jimin.jia, ygwen}@ntu.edu.sg, luoyong@whu.edu.cn

Abstract

Existing battery State of Health (SOH) prediction approaches often struggle to provide both accurate predictions and reliable uncertainty estimates. This paper presents a novel Multi-Task Learning (MTL) framework that jointly tackles SOH prediction and provides a proxy metric for uncertainty through a unified architecture. The framework combines a Physics-Informed Neural Network (PINN) for SOH prediction with a deep autoencoding Gaussian mixture model for uncertainty modeling. Particularly, the energy score from the Gaussian mixture model serves as a proxy metric for uncertainty, where a higher score indicates potential prediction unreliability. Moreover, to enhance task-specific learning, we employ a multi-head attention mechanism that adaptively captures distinct feature relationships. Our experiments show improvements in prediction performance compared to the state-of-the-art baseline. A comprehensive evaluation on six XJTU battery benchmark datasets demonstrates that our framework achieves a prediction accuracy of 99.50% (MAPE: 0.0050) while providing reliable uncertainty quantification through the proxy metric.

Introduction

Accurate battery State of Health (SOH) prediction, which quantifies the remaining usable capacity of a battery, is crucial for modern battery management systems, particularly in safety-critical applications. Although deep learning approaches have advanced battery SOH prediction (Haifeng, Xuezhe, and Zechang 2009), achieving accurate predictions remains challenging due to complex microscopic electrochemical interactions across various operational conditions. Moreover, beyond mere accuracy improvements, quantifying prediction reliability has emerged as an essential requirement for practical deployment, especially in high-stakes applications where the cost of prediction failures is substantial.

The accurate and robust prediction of SOH faces two fundamental challenges. First, degradation of battery capacity involves complex electrochemical processes that are difficult to model explicitly. Moreover, various operational conditions, including dynamic current loads, temperature fluctuations, and various charging protocols, further complicate the modeling process (Shao et al. 2023). These complexities pose significant challenges to capture nonlinear degradation dynamics. Second, current works on SOH prediction lack mechanisms to quantify prediction uncertainty (Wen et al. 2023; Wei, Dong, and Chen 2018; Li and Tao 2020), particularly when encountering unfamiliar degradation patterns. This limitation makes it difficult to assess the reliability of their predictions in real-world applications. Although some probabilistic models have been proposed (Wang et al. 2019), they typically rely on strong assumptions about data distributions and prior knowledge, which may not hold under practical battery degradation scenarios.

Existing approaches for SOH prediction can be broadly categorized into physics-based and learning-based methods. Physics-based methods attempt to model battery degradation dynamics through physical laws, such as equivalent circuit modeling (Amir et al. 2022) and Shepherd model (Jung and Tullu 2023). Such methods explicitly model battery dynamics via electrochemical principles but rarely leverage data-driven frameworks. Learning-based approaches, on the other hand, can be further divided into deterministic and probabilistic methods. From the perspective of the bias-variance tradeoff (Hastie, Tibshirani, and Friedman 2017), these two branches of research lie at opposite ends of the spectrum. Deterministic methods, such as Support Vector Regression (SVR) (Wei, Dong, and Chen 2018) and Physics-informed Neural Network (PINN) (Wang et al. 2024b), directly model battery degradation patterns through neural networks to produce point estimates, but often exhibit high variance and are prone to overfitting. Probabilistic methods, including Bayesian neural networks (Jiang et al. 2021), Gaussian process regression (Liu et al. 2019), and Monte Carlo dropout (Wei et al. 2021), explicitly model uncertainties in the prediction process. However, certain implementations with overly strong priors or limited model complexity can lead to increased bias and potential underfitting (Lakshminarayanan, Pritzel, and Blundell 2017), particularly in battery applications where degradation data is often scarce.

To address this bias-variance tradeoff, we propose a novel Multi-Task Learning (MTL) framework that not only enhances SOH prediction accuracy but also accounts for the inherent uncertainty of battery degradation datasets and provides a proxy metric for uncertainty quantification. Our framework achieves this through two key tasks. Specifi-

*Corresponding author: Yonggang Wen.

cally, we introduce a novel soft-constraint PINN for the SOH prediction task. Traditional PINNs typically enforce physical consistency by incorporating explicitly formulated governing equations as penalty terms into the training objective, where deviations from known physical equations directly increase the training loss (Raissi, Perdikaris, and Karniadakis 2019). However, battery degradation involves inherently complex electrochemical processes lacking explicit symbolic equations (O'Kane et al. 2022). Our soft-constraint PINN overcomes this limitation by implicitly capturing physical consistency through automatic differentiation of latent representations, encoding physically meaningful derivatives as a weak prior. This enables effective modeling of battery dynamics without explicitly formulating governing equations, significantly support the predictions in battery applications. In parallel, we introduce Deep Autoencoding Gaussian Mixture Models (DAGMM) for the uncertainty task. DAGMM naturally encodes feature distributions into a compact latent representation and uses an energy-based score to assess deviations from training distribution and data corruption (Ma et al. 2021), thus explicitly capturing uncertainties arising from distributional shifts. By capturing these uncertainties, DAGMM provides an inductive bias that guides latent feature representation learning during training. Unlike conventional uncertainty metrics (e.g., expected calibration error) that assume independent and identically distributed settings and become unreliable under non-stationary or shifted distributions (Ovadia et al. 2019), our approach leverages an energy-based Negative Log-Likelihood (NLL) metric. This energy score acts as a shift-aware proxy for uncertainty, detecting distributional anomalies and avoid overconfident predictions in out-of-sample regimes. To the best of our knowledge, this is the first work to introduce an energy-based uncertainty proxy metric specifically tailored for battery SOH estimation algorithms under dataset shifts.

The main contributions can be summarized as:

- We develop a MTL framework integrating SOH prediction with an energy-based proxy metric for uncertainty quantification, where the uncertainty task regularizes feature representations during training. This work proposes a general surrogate modeling paradigm for stochastic physical systems, highlighting the potential of integrating deep learning with uncertainty quantification.
- We propose a novel soft-constraint PINN that implicitly encodes physical consistency through automatic differentiation, modeling battery degradation without explicit equations.
- Our experiments on the XJTU battery datasets demonstrate that our approach achieves state-of-the-art prediction accuracies (99.50% for single protocols, 97.65% for heterogeneous protocols) and an effective proxy uncertainty metric to identify possible unreliable predictions.

Model Architecture

In this section, we propose a unified MTL framework integrating SOH prediction with uncertainty assessment. Our framework consists of four key components shown in Fig. 1: (1) A shared encoder that extracts common representations from battery datasets, (2) A multi-head self-attention mechanism to refine the latent features, (3) Two parallel tasks to perform SOH prediction and uncertainty assessment, and (4) A joint learning mechanism to balance two tasks.

Shared Encoder

The accurate prediction of SOH and the uncertainty quantification rely heavily on the extracted features from battery datasets. In line with XJTU datasets (Wang et al. 2024b), we extract statistical features (mean, variance, kurtosis, etc.) from Constant Current (CC) and Constant Voltage (CV) phases during battery charging, and design a shared encoder to transform these features into a latent space.

Then, we incorporate a multi-head self-attention mechanism to capture interdependencies among the input features rather than directly identifying the importance of individual features relative to the SOH prediction. Its primary role is to enhance representation learning by encoding complex nonlinear relationships among the diverse features.

Task 1: SOH Prediction

We present a novel soft-constraint PINN-based module to efficiently predict battery SOH as Task 1 in MTL. Inspired by gradient-based physics learning (Wang et al. 2024b; Meng et al. 2020), our soft-constraint PINN approach differs by not relying on explicit governing equations. As shown in Fig. 2, our soft-constraint PINN architecture consists of two key components: a predictor network and a physics network. The predictor network takes the attention-weighted battery features $z_1, ..., z_n$ and cycle number t as inputs. The output SOH prediction S and the latent feature representation z, are then fed into the physics network, which implicitly captures physical constraints through automatic differentiation rather than explicit equations.

The physics network is a three-layer Multilayer Perception (MLP) that leverages automatic differentiation to model the battery degradation dynamics in $\mathcal{H}(z, S, S_{\text{t}}, S_{\text{z}})$. S, S_{t} and S_{z} denote the current SOH and its gradients with respect to time and features, respectively. Unlike conventional PINNs explicitly employing accurate governing equations, our soft-constraint PINN leverages these derivative constraints as implicit physical regularization. This design choice leverages a weak physical prior derived from general degradation behaviors, making it particularly suited for battery degradation scenarios lacking explicit governing equations.

While we compute S_{t} to capture how SOH changes over time, $\mathcal{H}(\cdot)$ independently predicts this degradation rate based on the aforementioned physical information it receives. We utilize $\mathcal{L}_{\text{physics}} = (S_{\text{t}} - \mathcal{H}(z_{\text{p}}, S, S_{\text{t}}, S_{\text{z}}))^2$ to enforce consistency between the observed and physically expected degradation rates. The physics-informed loss is formulated as:

$$\mathcal{L}_{\text{soh}} = \mathcal{L}_{\text{mse}} + \alpha \mathcal{L}_{\text{physics}} + \beta \mathcal{L}_{\text{reduced}}. \tag{1}$$

This loss function consists of three parts. First, $\mathcal{L}_{\text{mse}} = \frac{1}{n}\sum_{i=1}^{n}(S_i - \hat{S}_i)^2$ refers to the mean squared error between true S and predicted $\hat{S}$. Then, we introduce a physics

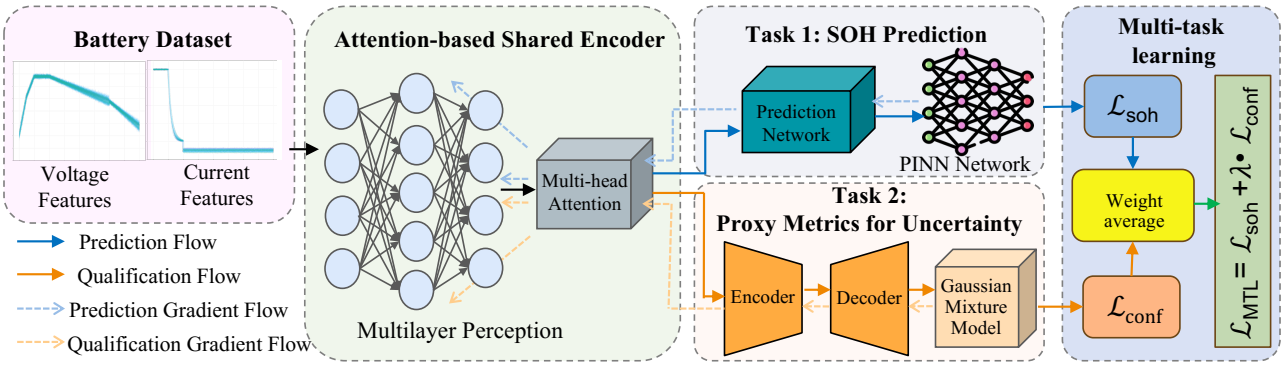

Figure 1: Overview of our MTL framework. The framework consists of four main components: (1) A shared encoder with multi-head attention mechanisms for feature extraction, (2) a PINN-based task for SOH prediction, (3) a DAGMM-based task for the proxy metric for uncertainty, and (4) a MTL module that integrates both tasks. The blue and orange arrows indicate the information flows, while the dashed arrows show the corresponding gradient flows during training.

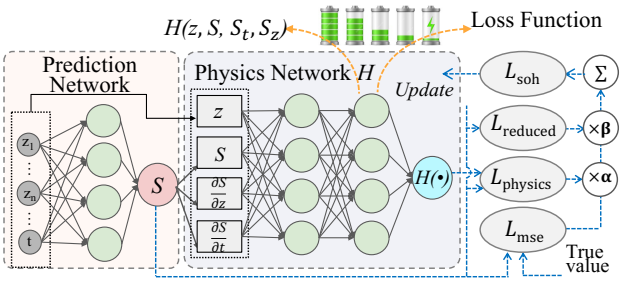

Figure 2: Architecture of the module for Task 1. A predictor network that maps battery features to SOH values, and a physics-guided network that enforces physical consistency.

loss term $\mathcal{L}_{\text{physics}}$. This physics-informed constraint helps our model maintain reasonable predictions even when encountering unfamiliar degradation patterns. Although the physics network's outputs are explicitly used only within the loss function, this design explicitly guides the latent feature space towards better representation of physical consistency and underlying degradation behaviors to improve generalization. Since S inevitably decreases during usage, we apply a monotonicity constraint $\mathcal{L}_{\text{reduced}} = \sum_{i=1}^{n-1} \max(0, \hat{S}_{i+1} - \hat{S}_i)$ that penalizes any predicted increases in SOH.

Task 2: The Proxy Metric for Uncertainty

In Task 2, we leverage DAGMM to quantify prediction uncertainty, which combines a deep autoencoder and a Gaussian Mixture Model (GMM). Traditional uncertainty quantification includes Calibration Error (CE)-based metrics, such as Expected Calibration Error (ECE) are typically non-differentiable and serve only as post-hoc evaluation tools (Karandikar et al. 2021). Consequently, neural network parameters cannot be directly optimized for uncertainty using these methods. Meanwhile, physics-informed probabilistic methods, such as Bayesian PINNs (Yang, Meng, and Karniadakis 2021), Bayesian Neural ODEs (Dandekar et al. 2020), and Gaussian processes with physical priors (Pförtner et al. 2022), require fully or partially known PDEs or ODEs, which are typically unavailable due to unobservable degradation processes in batteries (Li et al. 2019).

In real-world battery operations, environmental, manufacturing, and usage-induced variability often causes test data to diverge significantly from the training distribution (von Bülow, Heinrich, and Paxton 2024). Therefore, instead of directly analyzing the uncertainty in the high-dimensional sample space, we utilize the energy score (Zong et al. 2018) in Eq. 2 as a proxy metric operating on the condensed latent representations $\boldsymbol{z}$ learned by the model, naturally providing a differentiable, structured regularization during training. Specifically, this energy score evaluates how well a latent representation fits the learned GMM distribution, explicitly capturing uncertainties arising from distributional shifts.

When DAGMM processes new battery data, the energy score provides a quantitative measure of distribution shift by evaluating the NLL of the latent vectors under the trained GMM part. Specifically, if a battery exhibits unusual voltage patterns or degradation behaviors that differ from the training distribution, its latent representation will be mapped to low-density regions in the GMM, resulting in a higher energy score. This score, therefore, serves as a metric for identifying potentially unreliable predictions where the model encounters unfamiliar degradation patterns.

Instead of relying on known differential equations, our approach imposes implicit physics-informed constraints in the latent space. The energy score examines the distributional characteristics in the learned latent space, where essential features of battery degradation patterns are preserved through the autoencoder in DAGMM. As shown in Fig. 3, we implement two main components to construct the energy-based proxy metric: an autoencoder network for feature compression and GMM for energy scoring. Firstly, we design the encoder $h(x; \theta_e)$ to progressively compress the features through multiple fully connected layers to z_c, while the decoder $g(\boldsymbol{z}_c; \theta_d)$ reconstructs the original features. Such a compression-reconstruction process enables

the network to learn the most essential battery degradation patterns, forming the basis for uncertainty assessment.

Secondly, we compute reconstruction error z_{r} and cosine similarity z_{e} between original and reconstructed features. Then, we combine $[z_{\mathrm{c}}, z_{\mathrm{r}}, z_{\mathrm{e}}]$ into a joint representation vector $\boldsymbol{z}$ capturing both battery state and the model's capability to reconstruct observed patterns.

Thirdly, we model the density distribution of $\boldsymbol{z}$ using the GMM part since battery degradation patterns naturally form clusters under different operating conditions. GMM first performs membership prediction and outputs γ_{ik}, representing the probability of sample i belonging to the k-th Gaussian component. These membership predictions are then used to compute the GMM parameters: mixture weights $\omega_k = \sum_{i=1}^{N} \gamma_{ik}$, component means $\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \gamma_{ik} \boldsymbol{z}_i}{\sum_{i=1}^{N} \gamma_{ik}}$, and covariance matrices $\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} \gamma_{ik} (\boldsymbol{z}_i - \boldsymbol{\mu}_k)(\boldsymbol{z}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \gamma_{ik}}$, where N is the sample batch size, ω_k represents mixture weights, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are mean and covariance matrices for the k-th component. Finally, the energy score is:

$$E(\boldsymbol{z}) = -\log\left(\sum_k \frac{\omega_k}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \cdot \exp\left(-\tfrac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_k)\right)\right). \tag{2}$$

This energy score serves as a proxy metric for uncertainty. Rather than directly estimating uncertainty, the energy score efficiently captures epistemic uncertainty when encountering out-of-distribution samples, echoing spectral regularization ideas for robustness in rank aggregation (Ma et al. 2022). In battery applications, environmental variability such as temperature fluctuations and varying charging rates introduces data distribution shifts, which can be effectively modeled by Gaussian noise (Zhang et al. 2024). Therefore, we primarily adopt additive Gaussian noise to simulate battery environmental variability in this paper, and we have the following proposition.

Proposition 1 *Given the encoder network $h(\boldsymbol{x}; \theta_e)$ is Lipschitz continuous, adding Gaussian noise to input features increases the expected energy score evaluated by the GMM. Furthermore, the expected energy score increases monotonically with the standard deviation of additive Gaussian noise.*

Remark 1 *By Proposition 1, higher energy scores effectively indicate increased uncertainty under noisy or shifted conditions. To our best knowledge, the proposed framework is the first having these properties among existing battery estimation algorithms with proofs in Appendix A, indicating the energy score as an effective uncertainty metric.*

To optimize our uncertainty proxy metric, we consider a joint loss function: $\mathcal{L}_{\mathrm{conf}} = \mathcal{L}_{\mathrm{rec}} + \lambda_1 \mathcal{L}_{\mathrm{energy}} + \lambda_2 \mathcal{L}_{\mathrm{cov}}$. The reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ measures the mean square error between input and reconstructed features, ensuring our encoder captures essential characteristics while filtering out noise. Energy-based loss $\mathcal{L}_{\mathrm{energy}} = \frac{1}{N}\sum_{i=1}^{N} E(\boldsymbol{z}_i)$ promotes effective uncertainty detection by penalizing deviations from normal patterns. Covariance regularization term $\mathcal{L}_{\mathrm{cov}}$ maintains numerical stability of the GMM components.

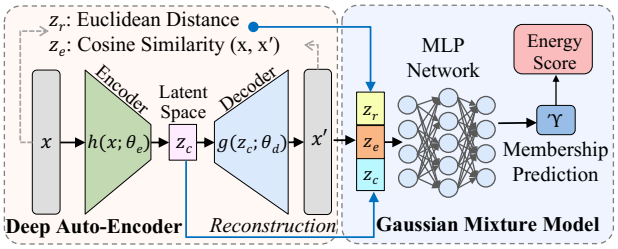

Figure 3: Architecture of the module for Task 2. A deep autoencoder for feature compression and reconstruction, and a joint representation to generate energy scores.

This loss formulation enables our model to learn discriminative representations for uncertainty quantification under the reliable energy-based proxy metric.

MTL Framework and Task Interaction

We design our joint optimization objective to combine $\mathcal{L}_{\mathrm{soh}}$ for Task 1 and $\mathcal{L}_{\mathrm{conf}}$ for Task 2:

$$\mathcal{L}_{\mathrm{MTL}} = \mathcal{L}_{\mathrm{soh}} + \lambda \mathcal{L}_{\mathrm{conf}}. \tag{3}$$

The shared encoder maps raw battery data into a unified latent space, providing a compact and structured representation of the high-dimensional input. This latent representation is simultaneously utilized for two subsequent tasks. Following Kontolati et al. (2024), which demonstrated that enforcing physical consistency in latent spaces enhances the modeling of complex physical systems, we constrain latent representations via PINN to satisfy physical priors. Concurrently, we develop the DAGMM network to encourage the latent representations to follow a more concentrated distribution, ensuring that data samples from the same category are closer together. The parameters of the PINN and DAGMM networks, along with those of the shared encoder, are jointly optimized through Eq. 3. The reason behind this joint optimization is that DAGMM explicitly regularizes the latent feature space toward a structured multimodal representation, complementing PINN's focus on physical consistency. As detailed justified in Appendix B, the energy-based metric provided by DAGMM monotonically increases under data shifts, thereby offering a robust uncertainty signal and ensuring a more informed feature representation. During backpropagation, the shared encoder receives gradient updates from both tasks, leading to an enriched and physically meaningful latent representation. As a result, the learned encoder not only effectively separates data from different categories within the latent space but also ensures that the latent representations inherit essential physical priors.

The effectiveness of our MTL framework can be analyzed through an (a, b)-model framework (Baxter 1997), where a represents the dimension of the minimal hypothesis space sufficient for learning all tasks, and b denotes the dimension of the learner's representation space. For an (a, b)-model with n tasks, let m denote the number of examples required per task. The sample complexity is bounded by $m = O(a + \frac{b}{n})$. Following Baxter (1997), our task satisfies

the (a, b)-model by: (1) The degradation patterns share underlying physical mechanisms (a common hypothesis space a), (2) Our MTL provides the representation space b, (3) The training dataset size m provides sufficient samples. In our MTL framework with two jointly trained tasks, we achieve loss reduction through the (a, b)-model analysis. Following Baxter (1997), we have the cumulative loss for joint learning $C_{joint} \doteq \frac{\log m}{2}\left(a + \frac{b}{2}\right) + o(\log m)$, and for independent learning is $C_{indep} \doteq \frac{\log m}{2}(a + b) + o(\log m)$. Therefore, $C_{joint} - C_{indep} = -\frac{b \log m}{4} + o(\log m)$, demonstrating our MTL framework can achieve better performance through joint training.

Our MTL framework balances SOH prediction and uncertainty quantification through adaptive loss weighting and gradient balancing. The weighting factor λ is tuned to prioritize prediction accuracy while maintaining effective uncertainty assessment, as accurate SOH prediction serves as the foundation for meaningful uncertainty evaluation. We apply gradient normalization to avoid domination of one gradient stream, a design that aligns with the elegant fixed-point perspective on equilibrium optimization proposed in spectral rank aggregation (Ma et al. 2024).

Experiments and Evaluation

We evaluate our MTL framework against five baseline models: Long Short-Term Memory (LSTM), MLP, Convolutional Neural Network (CNN), Transformer, and PINN with detailed configurations in Appendix E. The PINN baseline (Wang et al. 2024b) is currently recognized as a state-of-the-art (SOTA) approach for battery SOH estimation. The LSTM, MLP, and CNN are widely adopted SOH estimation baselines (Wang et al. 2024a), and we further include a Transformer-based baseline. We perform the following four sets of experiments: (1) A performance study of prediction accuracy, (2) Cross-dataset generalization tests, (3) Validation of the energy-based uncertainty proxy metric, and (4) Ablation studies examining the contributions of the attention mechanism, uncertainty task, and MTL structure.

Datasets and Evaluation Metrics

To further demonstrate our framework's applicability in realistic and highly stochastic scenarios, we conduct experiments on six XJTU battery datasets (Wang et al. 2024b) representing diverse charging/discharging protocols beyond simple accelerated aging. In addition to fixed-rate protocols (named 2C, 3C) and variable-rate discharging with varying depths (named R2.5, R3), the datasets comprise random walk discharges (named RW) with randomized currents and durations, and satellite simulations (named Satellite) that reflect non-uniform operational cycles typical of satellite batteries. The detailed charging/discharging profiles of these six datasets are illustrated in Appendix D. We report four widely adopted error metrics: Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) as detailed in Appendix E. In the main experiments, we particularly focus on MAPE and RMSE as they provide complementary insights.

Evaluation of SOH Prediction

Evaluation Setting. We evaluate model performance under single-dataset training (specific protocols) and joint-dataset training (generalization across diverse conditions).

Evaluation Results. In single-dataset training, we compare the prediction accurancy of our MTL model against LSTM, CNN, Transformer, and PINN baselines in Tab. 1. Our approach consistently improves performance, reducing MAPE and RMSE by approximately 57.2% and 60.5% compared to LSTM, MLP, CNN, and Transformer baselines. It also shows notable improvements over the current SOTA baseline, PINN, with reductions of 40.7% in MAPE and 48.4% in RMSE. In particular, even for the RW dataset, which is challenging due to its stochastic charging/discharging protocols that significantly deviate from standard fixed protocols, our model maintains strong performance (MAPE: 0.80%, RMSE: 0.98%) along with PINN (MAPE: 1.35%, RMSE: 1.90%), while CNN suffers significant performance degradation (MAPE: 3.50%, RMSE: 4.53%). As clearly illustrated in Appendix E, our approach provides distinctly predictions to reflect diverse battery degradation patterns in six datasets. For example, dataset 3C exhibits a rapid and consistent monotonic SOH decline characteristic, while dataset R3 clearly demonstrates staged degradation trends with intermittent slow-decline phases and abrupt capacity drops. The predicted SOH curves accurately track the actual degradation patterns across all datasets.

In all-dataset evaluation, our MTL model achieves MAPE of 1.09% and RMSE of 1.17%, improving upon PINN's 1.79% and 2.37% by 39.1% and 50.6%, as shown in Tab 3. This demonstrate that our MTL framework effectively leverages the diverse degradation patterns from multiple datasets to enhance its prediction capability. This improvement can be attributed to the effective knowledge sharing in our MTL framework helps identify universal degradation patterns across different charging/discharging protocols.

Cross-dataset Evaluation

Evaluation Setting. We further assess generalization via cross-dataset evaluation, creating 15 training-testing combinations from the six datasets, focusing on MAPE and RMSE. Details about these combinations in Tab 2 are provided in Appendix E. This setting examines how well our model can adapt to and perform on previously unseen charging/discharging protocols, especially when transferring from simple to more complex protocols.

Evaluation Results. Tab. 2 presents the performance for all 15 cross-dataset combinations. When trained on fixed charging protocols, our MTL model achieves MAPE of 0.0266 compared to PINN's 0.0357, showing a 25.5% improvement. This advantage becomes more pronounced when combining simple and complex charging protocols. For instance, when trained on 3C (fixed-rate protocol) and Satellite (complex protocol), ours MTL model maintains strong performance (MAPE: 0.0262) while PINN's performance degrades (MAPE: 0.037759), representing a 30.6% improvement. This enhanced generalization can be attributed to the

Training Dataset	Ours		LSTM		MLP		CNN		Transformer		PINN	
	MAPE↓	RMSE↓	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
2C	**0.67±0.03**	**0.74±0.17**	1.30±0.40	3.62±1.18	2.60±0.22	2.77±0.20	2.70±1.04	3.30±1.06	0.95±0.15	1.11±0.14	1.80±0.07	2.32±0.19
3C	**1.54±0.06**	**1.65±0.13**	1.79±0.04	2.12±0.94	2.75±0.12	3.04±0.11	2.98±0.49	3.52±0.46	1.80±0.93	2.12±1.01	1.84±0.04	2.45±0.02
R2.5	**1.02±0.05**	**1.15±0.08**	2.30±0.27	2.68±0.36	2.11±0.20	2.37±0.39	1.77±0.35	2.12±0.46	1.24±0.16	1.46±0.19	1.61±0.11	2.27±0.17
R3	**0.50±0.09**	**0.63±0.21**	1.41±0.54	2.28±0.64	2.00±0.07	2.35±0.09	1.50±0.42	1.89±0.57	1.01±0.31	1.24±0.27	1.56±0.06	2.14±0.07
RW	**0.80±0.08**	**0.98±0.10**	1.87±0.99	2.39±1.07	1.83±0.26	2.17±0.25	3.50±0.25	4.53±0.33	3.15±0.08	3.62±0.08	1.35±0.21	1.90±0.42
Satellite	**0.67±0.02**	**0.74±0.05**	2.21±0.65	2.79±0.73	2.04±0.05	2.42±0.07	1.49±0.50	1.94±0.65	1.75±0.86	2.09±1.03	1.47±0.04	2.00±0.21

Table 1: SOH prediction accuracy of single dataset evaluation (in %). Bold: best, underline: second-best.

No.	Ours		LSTM		MLP		CNN		Transformer		PINN	
	MAPE↓	RMSE↓	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	**2.66±0.14**	**2.71±0.25**	3.84±0.08	6.35±0.43	4.06±0.24	9.91±0.60	3.89±0.49	4.54±0.91	3.63±0.41	4.27±0.53	3.57±0.16	4.24±0.28
2	**2.63±0.03**	**2.62±0.08**	3.59±0.42	4.64±1.35	3.68±0.64	9.02±0.76	3.57±0.67	4.23±1.43	2.99±0.56	3.49±0.65	3.25±0.08	3.88±0.21
3	**2.70±0.04**	**2.64±0.10**	3.87±1.05	5.76±1.65	3.54±0.10	8.52±0.46	3.79±0.23	4.44±0.78	3.00±0.65	3.50±0.79	3.46±0.47	4.10±1.15
4	**3.01±0.12**	**2.97±0.77**	3.44±0.48	5.41±0.95	4.79±0.14	6.75±0.63	3.90±0.56	4.52±0.89	4.03±0.19	4.61±0.14	3.69±1.23	4.27±1.90
5	**2.63±0.33**	**2.62±0.20**	3.85±1.37	4.59±1.30	3.31±0.65	4.72±0.91	3.87±0.17	4.58±0.49	3.54±0.69	4.24±0.79	3.49±0.40	4.31±1.06
6	**2.94±0.10**	**2.85±0.51**	4.38±0.84	6.30±2.44	3.36±0.57	4.64±1.04	3.81±0.70	4.47±1.17	3.14±1.09	3.62±1.13	3.46±0.79	4.17±1.31
7	**2.70±0.06**	**2.66±0.31**	4.06±0.35	4.68±0.51	3.83±0.72	7.83±1.64	4.04±0.48	4.68±0.60	3.75±0.59	4.34±0.72	3.63±0.10	4.36±0.29
8	**2.90±0.41**	**2.84±0.91**	5.54±1.01	5.50±1.42	3.41±0.45	6.22±0.92	6.76±0.19	6.91±0.34	4.26±0.24	4.82±0.17	3.80±0.48	4.46±0.76
9	**2.62±0.36**	**2.55±0.64**	4.13±0.70	4.89±0.94	3.43±0.80	7.18±2.05	4.20±0.12	4.97±0.27	4.28±0.06	4.87±0.02	3.77±0.66	4.62±0.69
10	**2.48±0.14**	**2.47±0.15**	3.24±0.50	3.82±0.72	3.71±0.15	4.38±0.50	3.62±0.08	4.43±0.15	2.75±0.77	3.30±0.88	3.35±0.09	4.02±0.21
11	**2.35±0.10**	**2.32±0.24**	3.79±0.33	4.43±0.63	3.80±0.10	4.43±0.21	4.12±0.41	5.55±0.69	2.95±1.06	3.47±1.19	3.51±0.12	4.15±1.07
12	**3.10±0.55**	**2.97±0.57**	3.88±0.35	5.45±0.58	3.82±0.08	4.55±0.23	3.85±0.26	5.61±0.69	3.54±0.57	4.26±0.61	3.41±0.53	4.28±0.55
13	**2.95±0.30**	**2.89±1.60**	4.00±0.71	4.62±1.34	3.99±0.33	4.62±0.87	4.01±0.13	4.62±0.26	3.60±0.78	4.21±0.90	3.70±0.34	4.33±1.56
14	**3.21±0.41**	**3.11±0.86**	3.49±0.80	4.23±1.13	4.10±0.38	4.81±1.29	3.59±0.15	4.41±0.34	3.66±0.93	4.36±0.96	3.68±0.46	4.52±0.97
15	**2.51±0.69**	**2.47±0.74**	6.19±0.65	9.89±1.72	3.62±0.27	9.89±0.41	4.08±0.11	4.77±0.24	4.12±0.10	4.80±0.03	3.71±0.06	4.52±0.34

Table 2: SOH prediction accuracy of cross-dataset evaluation (in %). Bold: best, underline: second-best.

complementary roles of both tasks in our MTL framework. Task 1 captures fundamental degradation patterns, and Task 2 models distributional regularities, jointly ensuring robust predictions for unseen charging conditions.

The knowledge sharing mechanism enables our model to learn more robust feature representations, which proves particularly beneficial when generalizing to unseen complex patterns like RW and Satellite charging protocols. When RW dataset is included in training, the performance gap between our MTL model and PINN is relatively smaller (25.6% improvement) compared to when training only on fixed-rate protocols like 2C and 3C (38.6% improvement). This suggests that the inherent variability in RW protocols provides rich degradation information that even baseline models can partially capture, while MTL's advantage becomes more pronounced when learning from more structured charging protocols and generalizing to diverse conditions.

Evaluation of the Proxy Metric for Uncertainty

Evaluation Setting. We evaluate the effectiveness of our energy-based metric using noise injection experiments with controlled Gaussian and non-Gaussian anomalies (simulating sudden jumps and sporadic outliers), generating 11 test datasets with progressively increased noise (from 0.01 to 0.5). This setup evaluates how prediction accuracy degrades with increasing noise, and whether the energy score correlates well with prediction uncertainty. A high energy score effectively signals unfamiliar or anomalous conditions, thus preventing potential SOH prediction errors.

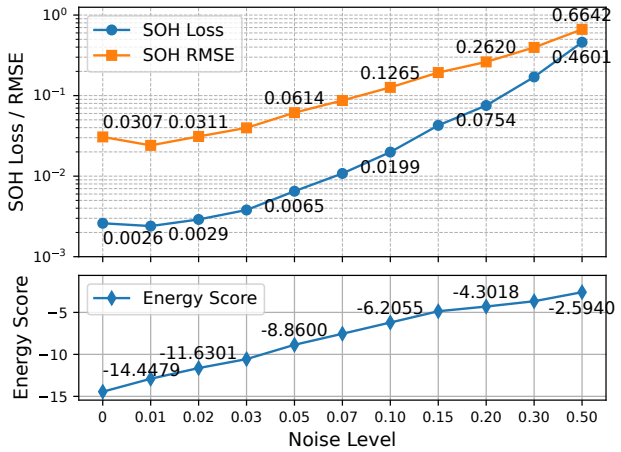

Figure 4: Impact of noise injection on prediction error and the proxy metric for uncertainty.

Evaluation Results. The effectiveness of our proxy metric is demonstrated through the positive correlation between noise levels and the energy score. As shown in Fig. 4, as the noise level increases from 0 to 0.5, the SOH prediction loss and RMSE show a consistent exponential growth pattern, with SOH loss increasing from 0.0026 to 0.4601 and RMSE rising from 0.0307 to 0.6642. Simultaneously, the energy score exhibits a strong positive correlation, progressively increasing from -14.4479 to -2.5940. The energy score proves

Model	MAE↓	MAPE↓	MSE↓	RMSE↓
Ours	**1.00**	**1.09**	**0.02**	**1.17**
LSTM	1.80	1.95	0.07	2.45
MLP	3.14	3.42	0.17	4.01
CNN	3.45	3.76	0.20	4.44
Transformer	2.02	2.21	0.08	2.67
PINN	1.67	1.79	0.06	2.37

Table 3: Prediction accuracy (%) averaged over 10 runs.

Model	MAE↓	MAPE↓	MSE↓	RMSE↓
Ours	**1.00**	**1.09**	**0.02**	**1.17**
Ours (1-atten)	1.47	1.59	0.03	1.71
Ours (w/o atten)	1.57	1.69	0.04	1.80
Ours (w/o Task2)	1.63	1.75	0.05	2.22
Ours (w/o MTL)	1.72	1.84	0.09	2.96

Table 4: Ablation study (%) averaged over 10 runs.

to be an effective proxy metric for potentially unreliable predictions. When the energy score exceeds a pre-defined threshold (e.g., -8.0), it signals potential data anomalies that could lead to unreliable SOH predictions. This proxy uncertainty metric is particularly valuable for electric vehicle operations, which enables proactive measures such as adjusting charging strategies to avoid high-risk operational conditions, scheduling maintenance before significant capacity loss occurs, or even recommending battery replacement when SOH predictions become consistently unreliable. The timely alerts enable operators to initiate preventive maintenance or retrain models before actual system failures occur.

Ablation Study

We conduct ablation studies to evaluate three components of our framework: the multi-head attention mechanism, the uncertainty task, and the MTL structure. Through comparisons with simplified variants, we aim to validate our design choices and their individual contributions.

Ablation of Attention Mechanism. To validate the effectiveness of the multi-head attention mechanism in our MTL framework, we compare our complete model with two variants: Ours (1-atten) using a single attention head, and Ours (w/o atten) without the attention mechanism. The ablation results in Tab. 4 demonstrate the effectiveness of our multi-head attention design. Reducing to a single attention head leads to 45.9% higher MAPE and 46.2% higher RMSE, while completely removing attention further degrades performance with 55.0% higher MAPE and 53.8% higher RMSE. Each of the 17 attention heads independently encodes inter-feature dependencies, with visualizations in Appendix E. This diversity in captured inter-feature relationships is inherently limited when using a single head or no attention, resulting in compromised predictive performance.

Ablation of Uncertainty Task (Task 2). To isolate the benefits of the uncertainty quantification task (Task 2) within our MTL framework, we evaluate a variant (Ours w/o Task2) that maintains the same encoder-PINN architecture but removes the uncertainty quantification task entirely. By comparing convergence speeds under identical training settings, we evaluate the impact of the MTL framework on training efficiency. Ablation results presented in Appendix E demonstrate that Task 2 substantially accelerates convergence, with our MTL model converging around 45-th epoch compared to over 100 epochs for PINN baseline. Notably, despite having nearly three times more parameters, typically implying a more challenging optimization landscape, the MTL model converges faster, indicating implicit regularization and more effective gradient utilization. Regarding inference complexity, Appendix E indicates that our MTL model incurs an increased inference latency (0.246 ms/sample) compared to the PINN baseline (0.109 ms/sample), primarily due to its dual-network structure and GMM inference module. Nevertheless, this absolute latency remains very low, and the achieved inference throughput (approximately 4062 samples/s) is still sufficiently high, acceptable for practical deployment scenarios. Thus, the additional computational overhead is modest and justified, given the significant improvements in prediction accuracy and uncertainty quantification. The performance comparison in Tab. 4 further validates the effectiveness of Task 2. Our complete MTL model achieves an additional 37.7% reduction in MAPE, confirming the benefits from the collaborative learning both tasks.

Ablation of MTL using Single-Head Probabilistic Regression. To further validate the advantage of formulating prediction accuracy and uncertainty quantification as two distinct tasks within our MTL framework, we implemented a single-task variant, ours (w/o MTL), using a probabilistic regression head trained with an NLL loss. Unlike our MTL structure, this single-head variant jointly predicts both SOH and its associated uncertainty from a unified output. Results presented in Tab. 4 clearly show that ours (w/o MTL) variant consistently underperforms our proposed MTL approach, exhibiting increases of approximately 35.5% in MAPE.

Conclusion

This paper proposes a novel MTL framework that jointly optimizes battery SOH prediction and a proxy uncertainty metric. The framework combines a PINN for SOH estimation with a DAGMM for uncertainty modeling through a multi-head attention mechanism. Extensive experiments show that the proposed approach significantly outperforms baseline methods, achieving lower MAPE under both single charging protocols and heterogeneous operating conditions, while simultaneously providing a meaningful proxy for uncertainty. While this unified framework represents just one possible approach, we hope it motivates broader research into surrogate modeling strategies for inherently noisy physical systems. This is particularly relevant for mission-critical systems, where both prediction accuracy and uncertainty quantification are crucial to ensure safe and reliable operation.

Acknowledgments

This work was supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 of RG104/23, in part by Singapore Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic of M23L9b0052, and in part by Leoch Battery Pte. Ltd.

References

Amir, S.; Gulzar, M.; Tarar, M. O.; Naqvi, I. H.; Zaffar, N. A.; and Pecht, M. G. 2022. Dynamic equivalent circuit model to estimate state-of-health of lithium-ion batteries. *IEEE Access*, 10: 18279–18288.

Baxter, J. 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28: 7–39.

Dandekar, R.; Chung, K.; Dixit, V.; Tarek, M.; Garcia-Valadez, A.; Vemula, K. V.; and Rackauckas, C. 2020. Bayesian neural ordinary differential equations. *arXiv preprint arXiv:2012.07244*.

Haifeng, D.; Xuezhe, W.; and Zechang, S. 2009. A new SOH prediction concept for the power lithium-ion battery used on HEVs. In *2009 IEEE vehicle power and propulsion conference*, 1649–1653. IEEE.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2017. The elements of statistical learning: data mining, inference, and prediction.

Jiang, B.; Gent, W. E.; Mohr, F.; Das, S.; Berliner, M. D.; Forsuelo, M.; Zhao, H.; Attia, P. M.; Grover, A.; Herring, P. K.; et al. 2021. Bayesian learning for rapid prediction of lithium-ion battery-cycling protocols. *Joule*, 5(12): 3187–3203.

Jung, S.; and Tullu, A. 2023. Characteristics Evaluation of 14 Battery Equivalent Circuit Models. *IEEE Access*.

Karandikar, A.; Cain, N.; Tran, D.; Lakshminarayanan, B.; Shlens, J.; Mozer, M. C.; and Roelofs, B. 2021. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34: 29768–29779.

Kontolati, K.; Goswami, S.; Em Karniadakis, G.; and Shields, M. D. 2024. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. *Nature Communications*, 15(1): 5101.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Li, X.; Colclasure, A. M.; Finegan, D. P.; Ren, D.; Shi, Y.; Feng, X.; Cao, L.; Yang, Y.; and Smith, K. 2019. Degradation mechanisms of high capacity 18650 cells containing Si-graphite anode and nickel-rich NMC cathode. *Electrochimica Acta*, 297: 1109–1120.

Li, Y.; and Tao, J. 2020. CNN and transfer learning based online SOH estimation for lithium-ion battery. In *2020 Chinese Control And Decision Conference (CCDC)*, 5489–5494. IEEE.

Liu, K.; Hu, X.; Wei, Z.; Li, Y.; and Jiang, Y. 2019. Modified Gaussian process regression models for cyclic capacity prediction of lithium-ion batteries. *IEEE Transactions on Transportation Electrification*, 5(4): 1225–1236.

Ma, K.; Xu, Q.; Zeng, J.; Cao, X.; and Huang, Q. 2021. Poisoning attack against estimating from pairwise comparisons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6393–6408.

Ma, K.; Xu, Q.; Zeng, J.; Li, G.; Cao, X.; and Huang, Q. 2022. A tale of hodgerank and spectral method: Target attack against rank aggregation is the fixed point of adversarial game. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4090–4108.

Ma, K.; Xu, Q.; Zeng, J.; Liu, W.; Cao, X.; Sun, Y.; and Huang, Q. 2024. Sequential manipulation against rank aggregation: theory and algorithm. *IEEE transactions on pattern analysis and machine intelligence*, 46(12): 9353–9370.

Meng, X.; Li, Z.; Zhang, D.; and Karniadakis, G. E. 2020. PPINN: Parareal physics-informed neural network for time-dependent PDEs. *Computer Methods in Applied Mechanics and Engineering*, 370: 113250.

O'Kane, S. E.; Ai, W.; Madabattula, G.; Alonso-Alvarez, D.; Timms, R.; Sulzer, V.; Edge, J. S.; Wu, B.; Offer, G. J.; and Marinescu, M. 2022. Lithium-ion battery degradation: how to model it. *Physical Chemistry Chemical Physics*, 24(13): 7909–7922.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Pförtner, M.; Steinwart, I.; Hennig, P.; and Wenger, J. 2022. Physics-informed Gaussian process regression generalizes linear PDE solvers. *arXiv preprint arXiv:2212.12474*.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378: 686–707.

Shao, J.; Li, J.; Yuan, W.; Dai, C.; Wang, Z.; Zhao, M.; and Pecht, M. 2023. A novel method of discharge capacity prediction based on simplified electrochemical model-aging mechanism for lithium-ion batteries. *Journal of Energy Storage*, 61: 106788.

von Bülow, F.; Heinrich, F.; and Paxton, W. A. 2024. The future of battery data and the state of health of lithium-ion batteries in automotive applications. *Communications Engineering*, 3(1): 173.

Wang, B.; Lu, J.; Yan, Z.; Luo, H.; Li, T.; Zheng, Y.; and Zhang, G. 2019. Deep Uncertainty Quantification: A Machine Learning Approach for Weather Forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2087–2095. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

Wang, F.; Zhai, Z.; Liu, B.; Zheng, S.; Zhao, Z.; and Chen, X. 2024a. Open access dataset, code library and benchmarking deep learning approaches for state-of-health estimation of lithium-ion batteries. *Journal of Energy Storage*, 77: 109884.

Wang, F.; Zhai, Z.; Zhao, Z.; Di, Y.; and Chen, X. 2024b. Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. *Nature Communications*, 15(1): 4332.

Wei, J.; Dong, G.; and Chen, Z. 2018. Remaining useful life prediction and state of health diagnosis for lithium-ion batteries using particle filter and support vector regression. *IEEE Transactions on Industrial Electronics*, 65(7): 5634–5643.

Wei, M.; Gu, H.; Ye, M.; Wang, Q.; Xu, X.; and Wu, C. 2021. Remaining useful life prediction of lithium-ion batteries based on Monte Carlo Dropout and gated recurrent unit. *Energy Reports*, 7: 2862–2871.

Wen, P.; Ye, Z.-S.; Li, Y.; Chen, S.; Xie, P.; and Zhao, S. 2023. Physics-informed neural networks for prognostics and health management of lithium-ion batteries. *IEEE Transactions on Intelligent Vehicles*.

Yang, L.; Meng, X.; and Karniadakis, G. E. 2021. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *Journal of Computational Physics*, 425: 109913.

Zhang, C.; Zhang, Y.; Li, Z.; Zhang, Z.; Nazir, M. S.; and Peng, T. 2024. Enhancing state of charge and state of energy estimation in Lithium-ion batteries based on a TimesNet model with Gaussian data augmentation and error correction. *Applied Energy*, 359: 122669.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*.