Early Prediction of the Failure Probability Distribution for Energy Storage Technologies Driven by Domain-Knowledge-Informed Machine Learning

Maher B. Alghalayini[1,2*†], Stephen J. Harris[1†] and Marcus M. Noack[2*†]

[1]Energy Storage and Distributed Resources Division, Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, 94720, California, USA.

[2]Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, 94720, California, USA.

*Corresponding author(s). E-mail(s): MAlghalayini@lbl.gov; MarcusNoack@lbl.gov;

Contributing authors: SJHarris@lbl.gov;

†These authors contributed equally to this work.

Abstract

There is a growing focus on sustainable energy sources and storage systems. The challenge with such emerging systems is their need to be warrantied for around 15 years with just a year of early testing. This requires accurate data extrapolation and estimation of the failure distribution. Physics-based approaches can be overwhelmed by the complexity of degradation, and pure data-driven approaches are inherently unable to extrapolate beyond the testing data. Here, we propose a framework for a hybrid approach for technology-agnostic customizations of a Gaussian process for stochastic and domain-knowledge-informed failure distribution predictions. We equip the Gaussian process with customized non-stationary kernels, heteroscedastic noise models, and prior mean functions to allow for accurate extrapolation with high accuracy. Furthermore, we minimize testing time with a novel experiment-stopping criterion, which can significantly reduce the required data. Our framework could revolutionize energy-storage testing, enabling the rapid development of new technologies.

Keywords: Machine Learning, Energy Storage, Early Lifetime Prediction, Failure Probability Distribution, Gaussian Processes

1 Introduction

As the climate warms at an accelerating rate, there has been a global shift from fossil fuels to more sustainable energy sources, such as solar and wind. While these renewable sources hold immense potential, their intermittent nature leads to fluctuations in energy generation, impeding their seamless integration into the energy grid [1]. Energy storage systems have emerged as indispensable solutions to store and release energy as needed. Because utilities will require warranties of perhaps more than approximately 15 years, it is essential to quantify the durability of newly developed energy storage technologies quickly and statistically.

Much research was done to predict the performance of energy storage systems over their lifetimes using physics-based modeling techniques and pure data-driven approaches [2, 3, 4]. While testing to failure under conventional operating conditions is another option, this would be prohibitively expensive and unrealistic for advancing new technologies. The physics-based approach uses models such as equivalent circuits [5, 6, 7], electrochemical [8, 9, 10], and empirical aging models [11, 12, 13] that rely on domain knowledge and expertise in battery degradation processes. Data-driven approaches use machine learning models such as Long Short-Term Memory Networks [14, 15, 16], Deep Neural Networks (DNN) [17, 18, 19, 20], and Gaussian processes (GP) [4, 21, 22, 3] that rely solely on a large amount of experimental data to predict battery failure [23, 24, 25, 26, 27, 28].

The above-referenced recent work predicted the expected failure date without providing information about the failure probability distribution. Quantifying failure distributions, which inherently encompasses the expected failures, is essential for establishing a warranty and evaluating second-life possibilities [29] since warranty costs depend on outliers that fail much earlier than the expected life. Battery degradation may vary significantly among nominally identical commercial batteries, even when operated under similar conditions [30, 31, 29, 32, 33, 34]. Therefore, it is crucial not only to quantify the average degradation path but also to estimate variability. Furthermore, this must be done efficiently to minimize the number of batteries tested to failure. Using the estimated variability, the probability of cells failing at each cycle number is computed to generate the failure distribution as a function of the cycle number. With this distribution, one can identify at what cycle number some fraction, for instance, the first 5%, of batteries are expected to fail, thus estimating the batteries' reliability. We aim to create a framework to predict battery failure distributions early, allowing us to evaluate whether a 20-year warranty is commercially feasible while requiring no more than one year of testing. This objective requires at least two building blocks: first, the demonstration of a valid accelerated testing protocol [35, 36] such that significant degradation occurs during the first year of testing [37]; and second, the ability to quickly assess the failure distribution for a given technology using those testing data. In this paper, we tackle the second issue.

One way to extrapolate and accurately estimate battery failure distributions is to implement machine learning tools, such as GP modeling. The GP is the most widely adopted class of stochastic processes and offers a robust and versatile framework for stochastic function approximation through Gaussian process regression. GPs have recently received attention from the battery community for failure prediction [4, 21, 24, 29]. The advantages of using GPs over other machine learning models include not requiring dense data — a large number of data points compared to the domain size — to give accurate predictions. In addition, they use Bayesian statistical methods to offer a robust approach to precisely quantify both aleatoric uncertainty, linked to inherent data variability, and epistemic uncertainty, associated with the lack of training data at prediction points. GPs are characterized by a normal prior probability distribution over latent function (the underlying, data-generating, unknown model) values, whose properties are controlled by the GP's covariance function (the kernel), the prior mean function, and the noise in the data. Domain experts can design all three functions to make GPs domain-knowledge aware.

The kernel, the prior mean, and the noise functions encode information about the predictions of the underlying latent function. The kernel function serves as the covariance operator and, therefore, quantifies the relationship between the data points and controls epistemic uncertainty quantification. In the vast majority of studies, the covariance is computed using stationary kernels, e.g., the Matérn kernel class, which depend only on the distance between the points in the input domain [38, 39]. The prior mean function encodes the users' prior knowledge and expectation of the general trend of the data before observing them. In most applications, the prior mean function is set to zero or a constant across the input domain. The noise model allows for quantifying variability, therefore, aleatoric uncertainty in the data and is usually assumed to be normally, independently, and identically distributed (i.i.d.) — also known as constant or homoscedastic — around the latent function values.

These standard GPs — stationary kernels, constant prior mean, and constant noise — can accurately predict the latent function values and the uncertainties in the proximity of existing data points, but lack domain-knowledge awareness and extrapolation capabilities. This limits the use of GPs for lifetime prediction of new battery technologies, especially when predicting degradation at operating conditions that have not been tested. Extrapolation capabilities can be provided to the GP by formulating flexible, unbiased, and physics-adhering prior mean functions. Moreover, since GPs get their uncertainty quantification capability from the kernel functions and noise models, using stationary kernels and i.i.d. noise is not generally advisable when accurate uncertainty quantification is required [40], which is the case for the predictions of failure distributions and decision-making regarding when an experimental campaign can be concluded — ideally early after only a handful of tested batteries.

Researchers prioritizing precise predictions of the latent function and its uncertainties while implementing standard GPs may conduct extensive experimentation across numerous batteries without a clear criterion for when to stop experiments on one battery and start experiments on another. This approach becomes particularly critical in scenarios lacking domain-knowledge awareness, limited extrapolation capabilities, and challenges in accurately quantifying uncertainties. Often, this extensive experimentation occurs without assessing whether additional testing provides useful new information. Generally, more data results in better predictions; however, the more data we collect, the more redundant information we acquired. Presently, researchers

typically rely on intuition or resource depletion to decide when to stop experiments. This generally results in inefficiencies and inaccurate predictions.

In this work, we aim to tackle domain awareness, extrapolation capability, accurate uncertainty quantification, and early stopping by incorporating domain knowledge into a GP model to identify the minimum necessary testing effort while achieving accurate predictions. Fortunately, one of the main advantages of GPs is their customizability. This customization includes advanced GP modeling with extended capabilities that account for domain knowledge by carefully choosing the prior mean, kernel, and noise functions. For instance, the battery literature indicates that battery degradation often occurs in two steps: a slow degradation rate followed by a knee and a faster degradation rate [30, 31, 41, 28]. Here, we demonstrate that a prior mean function that can model this (or any other) behavior significantly improves the extrapolation capabilities of the GP. Moreover, we show that using non-stationary kernels based on DNNs may allow for an accurate epistemic uncertainty estimation by warping the input domain. The battery literature also shows significant variability in the battery degradation paths that increases with cycling [23, 42]. We show that a flexible noise function can model this increase, significantly improving the aleatoric uncertainty estimation and allowing for an accurate prediction of the failure probability distribution of batteries. Finally, we propose a metric based on distribution entropy [43] to identify a stopping criterion for battery testing. Fig. 1 shows a general overview of the proposed framework and how it is tailored to batteries. This work is only the first step toward rapid validation of storage technology, where the aim is to predict cycling behavior; future work will build on the proposed methodology and add learning from cycling data of other batteries or technologies. We consider other novel approaches that have been used to estimate lifetimes in the Discussion section below.

The remainder of this paper is organized as follows. In Section 2, we explain our GP customizations step-by-step and demonstrate their effects using representative examples. We also discuss a novel experiment-stopping criterion driven by the Gaussian process posterior distribution. In Section 3, we apply the framework to two experimental battery datasets published in the literature. In Section 4, we discuss the results of our work and conclude.

2 Methods

We propose various extensions of the standard GP framework (equation 1) to provide it with properties desirable to model and analyze battery-testing data. The goal is to equip the GP with domain knowledge to approximate better the battery degradation latent function and quantity uncertainty. We also propose a novel stopping criterion for cycling experiments. Extending the GP's capabilities consists of defining three main building blocks. (1) A prior mean function that follows the expert-expected trend of the energy or capacity as a function of the cycle number while keeping it as unconstrained and, therefore, unbiased as possible. (2) A noise model tailored to the variability trends expected in the experimental data. (3) A kernel function that can reliably approximate uncertainties; generally, this means the kernel should be non-stationary [44]. In what follows, we discuss preliminaries, some machine learning terms used in this paper, the choice of each of the three functions, and the proposed stopping criterion in detail.

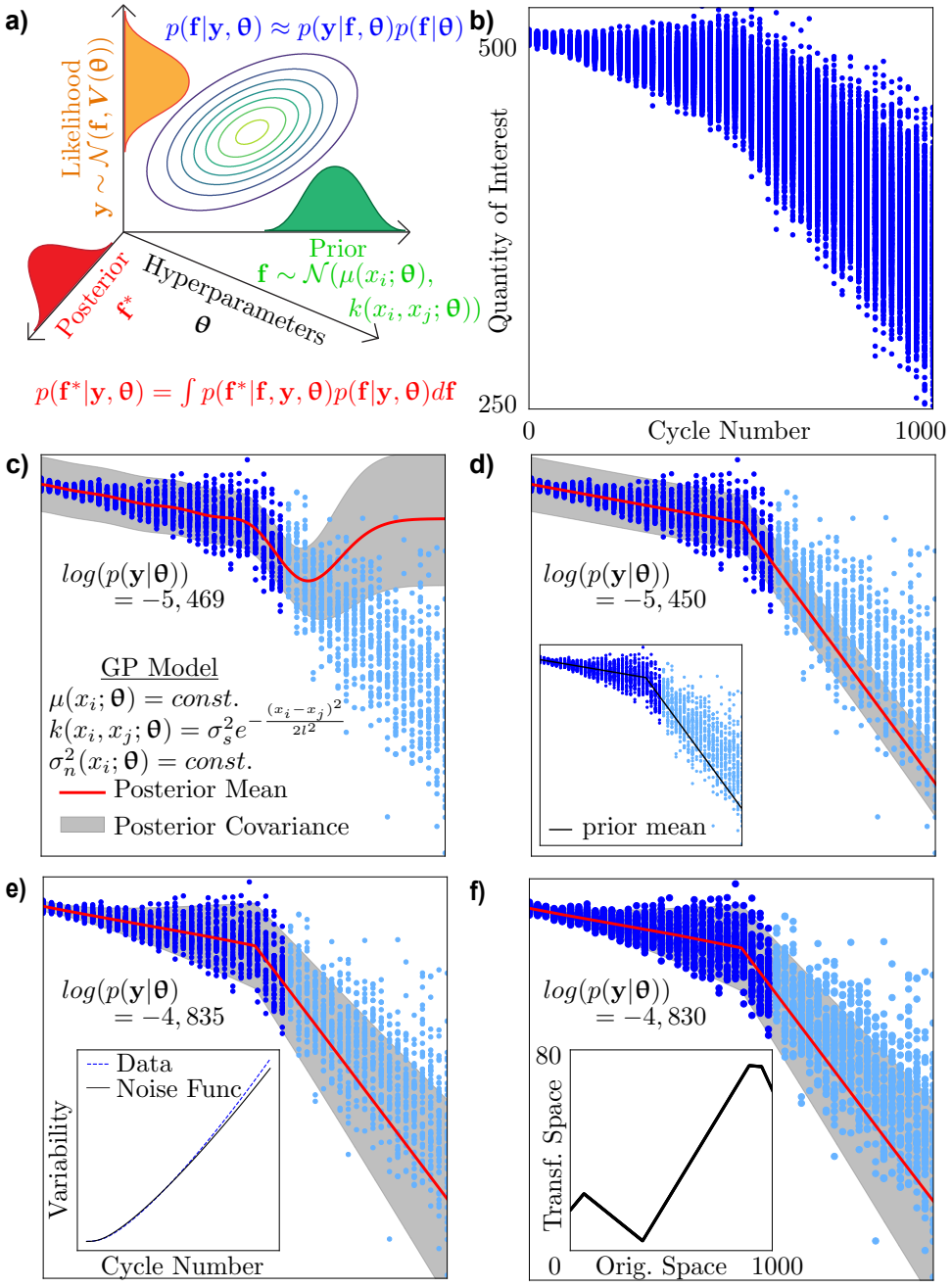

Fig. 1 Incorporating domain knowledge improves the GP's extrapolation and uncertainty quantification capabilities. (a) The different components of a GP model, where a prior normal distribution $\mathcal{N}(\mu(x_i;\theta), k(x_i,x_j;\theta))$ is defined over latent function values $\mathbf{f}$, and a likelihood $\mathcal{N}(\mathbf{f}, \mathbf{V}(\theta))$, with $\mathbf{V}(\theta) = \sigma_n^2(x_i;\theta)$ being a diagonal matrix and σ_n^2 the i.i.d. noise variance, over collected data $\mathbf{y}$. Both are used via Bayes' theorem to calculate the posterior probability density function over $\mathbf{f}^*$ as a function of the hyperparameters θ. (b) Drawn synthetic data from which a subset is selected to fit multiple variations of GP models. (c) A standard GP model — constant prior mean, noise, and a stationary kernel — poorly fits early battery data (dark blue) and does not allow extrapolation toward unseen regions (light blue). Tailoring a GP improves its predictions, where the log marginal likelihood, $\log(p(\mathbf{y}|\theta))$, increases when (d) the prior mean is defined as a 2-element piecewise linear function, shown in the inset, (e) the noise model is defined as a power-law, shown in the inset, and (f) a non-stationary deep kernel model is used as the covariance function — its space-warping ability is demonstrated in the inset.

2.1 Preliminaries

We employ GP modeling with various extensions to approximate a Quantity of Interest (QoI), battery energy or charge capacity, as a function of the cycle number $x \in \mathcal{X} \subset \mathbb{R}$. We assume an unknown data-generating latent function $f(x)$ and noisy function evaluations $y(x) = f(x) + \epsilon(x)$. We employ a GP to find a probabilistic representation of the latent function. In this section, we use synthetic data constructed based on the simulated testing of 25 identical batteries over 1000 cycles (see Fig. 2(a)) to illustrate our approach's approximation and uncertainty quantification capabilities. More specifically, fixed prior mean and noise models — referred to as the ground truth mean and noise in Fig. 2(a,b) — are chosen, then data are drawn from them using a multivariate normal sampling approach. The dataset exhibits the characteristic decreasing trend in battery energy with cycling. Batteries are assumed to have similar initial energy levels with minimal variability. This variability increases with cycling, aligning with the patterns observed in the literature [45, 30]. Although we consider this data in the context of battery degradation, it is important to note that our GP modeling approach is not limited to this specific application. It can be seamlessly customized to analyze and predict different QoIs with respect to any input variable of interest.

2.2 A Bird's Eye Perspective on Gaussian Processes

Gaussian processes are general-purpose function approximators that allow one to estimate a predicted function value and its uncertainty at an unobserved point in a multidimensional input space $\mathcal{X}$ based on observations at a set of given data points. Data are defined in this scope as a set of input-output pairs $\boldsymbol{D} = \{x_i, y_i\}\ \forall i \in \{1, 2, 3...\}$. The GP's basic principle is simple: Every known (observed) and unknown (of interest) function value (an underlying model) is thought of as a random variable. This could be any performance measure (QoI, e.g., discharge capacity) of a battery as a function of the battery cycle number. Then, a normal joint probability density function is defined over a finite set of function values. Basic statistical methods (marginalization and conditioning) let us calculate probability density functions for the model function value at unknown locations. To define a joint probability density $p(\mathbf{f}, \mathbf{f}^*)$ over known ($\mathbf{f}$) and unknown ($\mathbf{f}^*$) function values, a way to approximate covariances between the observed and unobserved function values is needed. This is called the kernel trick. A kernel is a function of two input locations $k(x_i, x_j; \boldsymbol{\theta})$ with some added properties (symmetry and positive semi-definiteness) that returns a scalar representing the estimated covariance between the function values at x_i and x_j. The most widely used stationary kernel is the squared exponential (SE) kernel [38]. A GP model with a constant prior mean μ, SE kernel function k_{SE}, and a normal i.i.d. noise ϵ is defined as

$$y(x) = f(x) + \epsilon(x), \tag{1a}$$

$$\boldsymbol{f} = f(x_i),\ \boldsymbol{y} = y(x_i)\ \forall i \in \{1, 2, 3, ...\}$$

$$\boldsymbol{f} \sim \mathcal{GP}(\mu(x_i; \boldsymbol{\theta}) = c_1, k_{SE}(x_i, x_j; \boldsymbol{\theta})), \tag{1b}$$

$$k_{SE}(x_i, x_j; \boldsymbol{\theta}) = \sigma_s^2 \exp\big[-\frac{||x_i - x_j||^2}{2l^2}\big], \tag{1c}$$

$$\epsilon(x, \boldsymbol{\theta}) \sim \mathcal{N}(0, \sigma_n^2(x, \boldsymbol{\theta}) = c_2) \tag{1d}$$

$$\Rightarrow \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{f}, diag(\sigma_n^2(x_i, \boldsymbol{\theta}))) \tag{1e}$$

Through this framework, we allow the prior mean, noise, and kernel functions to be exchangeable and to depend on arbitrary hyperparameters $\boldsymbol{\theta}$, endowing the GP with the capability to be customized for extra flexibility and domain awareness. This paper is focused on taking advantage of this extra flexibility to optimally predict failure distributions. For standard GPs, the hyperparameters $\boldsymbol{\theta} = \{c_1, \sigma_s^2, l, c_2\}$ result from constant prior mean and noise functions and a stationary kernel. The GP's hyperparameters are trained either via maximum log marginal likelihood estimation (MLE) — by solving $\text{argmax}_{\boldsymbol{\theta}}\, p(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y} = y_i,\ \forall i \in \{1, 2, 3, ...\}$ — or by using Markov Chain Monte Carlo sampling [44]. After learning the values of $\boldsymbol{\theta}$, we condition the marginalized prior on observed data $\mathbf{y}$ via Bayes' theorem to estimate the posterior distribution $p(\mathbf{f}^*|\mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\kappa}_p)$, where $\boldsymbol{\mu}_p$ and $\boldsymbol{\kappa}_p$ are the posterior mean and covariance, respectively, which constitute the GP predictions.

2.3 The Prior Mean Function

Domain (expert) knowledge is integrated into GPs by tailoring the prior mean function $\mu : \mathcal{X} \to \mathbb{R}$ to better fit the data trend and allow for a reliable extrapolation beyond the tested domain. The literature shows that a battery's energy or charge capacity (QoIs) generally degrades with cycling [1, 30]. The degradation rate may increase with cycling due to the onset of additional battery failure mechanisms. To include such knowledge in the GP model without constraining it to any particular shape, choosing a function that can describe this increased degradation rate is crucial. At the same time, the function needs the flexibility to correct the expert experimenter if needed, thereby enhancing the model, especially in cases where incorrect assumptions have been implemented, or to revert to a non-informative function if supported by data.

The power-law is one obvious candidate for a prior mean function that can model this trend. Such a prior mean can be written as

$$\mu(x; \boldsymbol{\theta}) = ax^p + b, \tag{2}$$

where x is the cycle number. This prior mean function depends on the hyperparameters $\{a, p, b\} \subset \boldsymbol{\theta}$. The slope a, which would be typically negative to show a decreasing trend, quantifies the model's degradation rate, the power p controls the non-linearity of the function, and the intercept b indicates the initial battery QoI before cycling. An example showing the use of the power-law function for the developed synthetic data after obtaining the values of the hyperparameters via MLE is shown in Fig. 2(a).

The literature shows that the onset of a new failure mechanism during cycling can cause a change in the rate of degradation and results in what is known as a "knee" — which means a point where the degradation curve is non-differentiable [46, 47, 48]. Several failure mechanisms could cause a "knee" in the degradation curve during cycling. These include lithium plating [49], electrolyte depletion [50], loss of active material [51], and mechanical deformation [52]. Researchers are often interested in identifying the cycle number at which this knee occurs. Unfortunately, the power-law model in equation 2 does not contain such a function. However, it can be generalized to the 2-element piecewise polynomial function

$$\mu(x; \boldsymbol{\theta}) = \begin{cases} a_1 x^{p_1} + b_1 & \text{if } x \leq x_0, \\ a_2 x^{p_2} + b_2 & \text{if } x > x_0, \end{cases} \tag{3a}$$

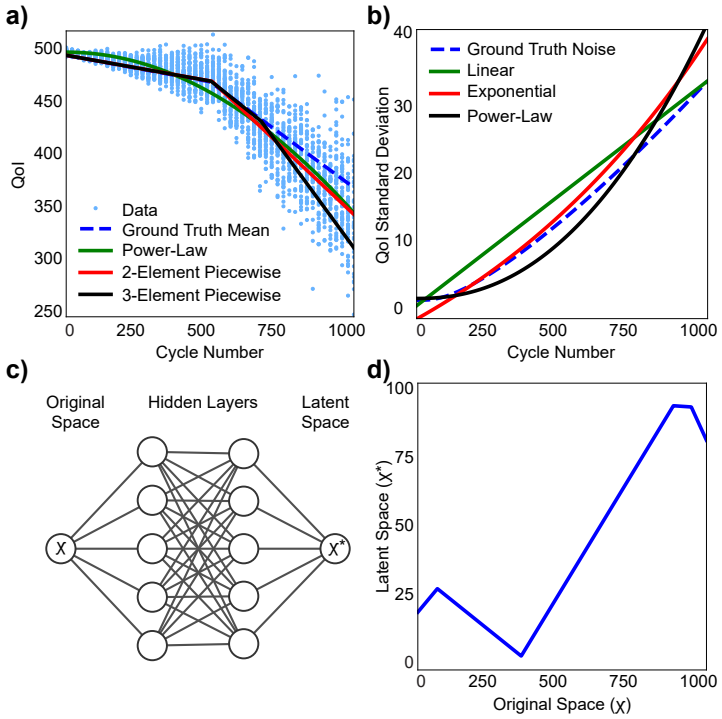

Fig. 2 GP customization performed to extend its capabilities to extrapolate and quantify uncertainty accurately. (a) The performance of the three prior mean functions compared to the ground truth mean function used to sample the synthetic data. (b) The performance of the three noise functions compared to the ground truth noise function. (c) The architecture of the DNN that transforms the original input space $\mathcal{X}$ to the latent space $\mathcal{X}^*$ as shown in (d).

$$b_2 = a_1 x_0^{p_1} - a_2 x_0^{p_2} + b_1, \tag{3b}$$

with hyperparameters $\{x_0, a_1, p_1, b_1, a_2, p_2\}$, and b_2 is computed by equation 3b to ensure that the two pieces of the function meet. The hyperparameter x_0 represents the location of the knee and indicates the onset of the second failure mechanism. Similar to the power-law above, the slopes a_1 and a_2 quantify the rate of degradation of each of the corresponding elements, the powers p_1 and p_2 control the linearity of the two elements, and b_1 indicate the initial QoI when $x = 0$. The values of p_1 and p_2 can be chosen such that the function becomes linear across the domain if either the experimenter believes that the failure mechanisms cause a steady battery degradation or such a decline results from the data as the most likely scenario. It is fundamental to the methodology to add the expert's knowledge in ways that allow the algorithm to ignore it if necessary to avoid bias. Fig. 2(a) shows an example of using the 2-element piecewise linear function.

Generally, multiple failure mechanisms can occur during battery life, and the mean function has to be sufficiently flexible to model that. To account for multiple failures, the user can generalize the 2-element function to an n-element piecewise polynomial function and allow the GP model to identify the number of failures most likely to occur. This is accomplished by checking the location of the knees resulting from the training. If these knees are within the domain, the GP predicts multiple failures will likely occur during cycling. Equation 4 demonstrates the transformation of the 2-element

piecewise polynomial function into a 3-element piecewise polynomial function,

$$\mu(x;\boldsymbol{\theta}) = \begin{cases} a_1 x^{p_1} + b_1 & \text{if } x \leq x_0, \\ a_2 x^{p_2} + b_2 & \text{if } x_0 < x \leq x_1, \\ a_3 x^{p_3} + b_3 & \text{if } x > x_1, \end{cases} \tag{4a}$$

$$b_2 = a_1 x_0^{p_1} - a_2 x_0^{p_2} + b_1, \tag{4b}$$

$$b_3 = a_2 x_1^{p_2} - a_3 x_1^{p_3} + b_2, \tag{4c}$$

with hyperparameters $\{x_0, x_1, a_1, p_1, b_1, a_2, p_2, a_3, p_3\}$, and $\{b_2, b_3\}$ are calculated using equation 4b-c. Fig. 2(a) shows an example of using this prior mean function. Similarly, the prior mean can be easily generalized to an n-element piecewise polynomial function.

2.4 The Noise Model

The noise function $\sigma_n^2 : \mathcal{X} \to \mathbb{R}$ quantifies the heteroscedastic aleatoric uncertainty in the QoI without having access to observed measurement variability [53]. In this context, distinguishing between two types of data variation is crucial: reducible measurement errors and inherent uncontrollable processes within batteries. For instance, when conducting experiments with different apparatuses, maintaining constant conditions can introduce data variation that inaccurately reflects battery degradation. In contrast, using the same apparatus under constant conditions highlights measurement variation attributed to uncontrollable events within batteries, such as the onset variation of second failure mechanisms. This variation more accurately represents battery degradation variation [22]. Here, we are interested in the latter. The core idea is to interpret those variations as measurement noise and to quantify it. In addition to improving uncertainty quantification, choosing a good noise model also improves the GP predictions, as it is a main ingredient in the MLE that allows performing Bayesian inference. Through the noise model, we focus on quantifying the aleatoric uncertainty with the least data possible to calculate the probability of failure distributions for warranty purposes. For this, it is crucial to integrate domain knowledge. For the synthetic example in this section, the literature shows that battery energy variability increases with cycling [30, 45]. To a large extent, this increase is due to the knee occurring at different cycles for different cells.

The linear function can capture a steady increase in variability and, therefore, predict a steady increase in aleatoric uncertainty. This noise function can be written as

$$\sigma_n^2(x;\boldsymbol{\theta}) = mx + n, \tag{5}$$

with hyperparameters $\{m, n\}$, that control the quantification of uncertainty in the predictions. Fig. 2(b) shows the trained linear noise function on the synthetic data compared to the actual variability of the data. However, as Fig. 2(b) shows, and the literature concludes, the variability in the battery energy (or capacity) may increase at a varying rate [29, 45]. It often starts with a slow increase at low cycle numbers, but the rate increases significantly when getting closer to the predicted knee. The linear function cannot model this varying rate of increase.

The exponential function allows for modeling the variability at an increasing rate and can be written as

$$\sigma_n^2(x;\boldsymbol{\theta}) = m\exp(x) + n, \tag{6}$$

with hyperparameters $\{m, n\}$. Fig. 2(b) also shows the trained exponential function. One potential limitation of this function is that it may be difficult to control the curvature of the function while ensuring that the y-intercept stays positive. The trained exponential function in Fig. 2(b) shows a negative y-intercept with an overestimation of the noise throughout the input domain. Having negative noise values leads to instabilities in the GP predictions, as the noise can never be negative. The power-law function can be used to help mitigate the issue of negative noise. This function, written as

$$\sigma_n^2(x;\boldsymbol{\theta}) = mx^p + n, \tag{7}$$

has the same trend as the exponential function, i.e., increasing at different rates based on the cycle number. The benefit of using this function is that its hyperparameters $\{m, p, n\}$ make it more flexible in controlling the rate of increase, the shape of the function, and the y-intercept. Fig. 2(b) also shows the trained power-law function on the data variability of the running example in this section.

2.5 The Kernel Function

The covariance function, or kernel, denoted as $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, is arguably the most important building block of a GP, and carefully choosing it is crucial for accurate prediction and epistemic uncertainty quantification. Accurate uncertainties (epistemic and aleatoric) are also crucial to defining valid experiment-stopping criteria. Generally, two classes of kernel functions exist: stationary kernels that depend only on the distance between points and non-stationary kernels that depend explicitly on the point's location in the input domain.

The vast majority of studies using GPs employ stationary kernels [38]. For most stationary kernels, the relationship is simple: the closer the data points, the more correlated they are. Therefore, for this class of kernels, the farther the prediction point from the tested points, the more uncertain the predictions become, which, while correct, is not the only aspect of data that should control uncertainty. We use the SE kernel presented in equation 1c and repeated here

$$k_{SE}(x_i, x_j;\boldsymbol{\theta}) = \sigma_s^2 \exp\big[-\frac{||x_i - x_j||^2}{2l^2}\big]. \tag{8}$$

Equation 8 is controlled by two constant scalar hyperparameters, σ_s^2 (the signal variance) and l (the length scale), whose values can be inferred from the data by MLE, and applied to the whole input domain $\mathcal{X}$. Additionally, and since equation 8 is a function of the Euclidean distance, i.e., the norm, $||\cdot||^2$, between the points in the input domain, it makes the kernel stationary. Stationary kernels are widely used in the machine learning community for their straightforward specification and the limited number of associated hyperparameters, simplifying the training. However, their dependence on only the distance between the points makes them prone to poor prediction performance and inaccurate uncertainty quantification, and as a result, poor quantification of the failure probability density function [40, 44].

Non-stationary kernels have emerged as a solution to improve predictions and uncertainty quantification, where they depend on the location of the points, not just the

distance between each other, i.e., $k(x_i, x_j) \neq k(|x_i - x_j|)$. This gives them flexibility and greater expressiveness in covariance calculations. However, this increased flexibility comes at higher costs, diminishing their popularity. Among other challenges, proving the positive semi-definite characteristic of a newly formulated non-stationary kernel is difficult. Additionally, these kernels are generally associated with significantly more hyperparameters to be estimated. This imposes a lower bound on the size of the datasets to which these kernels can be applied, along with increased computational requirements. Fortunately, significant work has been done to develop nonstationary kernels for GPs [54, 55, 56]. Among others, three methods were developed: Parametric non-stationarity, deep GPs, and deep kernels. In the first method, non-stationarity is generally accounted for in the signal variance via the form $k(x_i, x_j) = \sum_{d=1}^{N} g_d(x_i) g_d(x_j) k_{stat}(|x_i - x_j|)$, where $g_d(x)$ is any parametric function over the input domain, N a positive integer, and k_{stat} any stationary kernel. This principle has been extended to parametric length scales. In deep GPs, non-stationarity is achieved by stacking stationary GPs in multiple layers such that the output of one GP is the input of the other, similar to a DNN structure. Herein, we implement the third method, deep kernels, that uses DNN in the kernel [57, 58]. In such an approach, a DNN function ϕ is used to warp the input space $\mathcal{X}$ non-linearly to a latent space $\mathcal{X}^*$, with a potentially different number of dimensions. Then, a stationary kernel, here the SE, calculates the covariance between the points in the latent space associated with those of interest in the input domain. The deep kernel we implement is defined as

$$k(x_i, x_j; \boldsymbol{\theta}) = \sigma_s^2 \exp\Big[- \frac{||\phi(x_i) - \phi(x_j)||^2}{2l^2} \Big]. \tag{9}$$

This DNN, ϕ, comprises two hidden layers with five nodes per layer, with each node employing the rectified linear unit (ReLU) activation function. The DNN takes elements of the one-dimensional input space $\mathcal{X}$ and outputs elements of a one-dimensional warped space $\mathcal{X}^*$. Fig. 2(c) shows the architecture of the DNN used here. Using this DNN introduces a set of 46 additional hyperparameters. Fig. 2(d) shows the transformation of $\mathcal{X}$ to $\mathcal{X}^*$ using the set of synthetic data developed for this section. For more details about non-stationary kernels, interested readers are referred to [44].

2.6 Early-Stopping Criteria for Experimentation

After integrating domain knowledge into the GP model to minimize the number of experiments required, it is important to develop a stopping criterion that determines when further testing of a certain battery is no longer valuable. The aim is to identify, as early as possible, when a certain battery does not provide additional information about the failure probability. This minimizes the required testing resources while not risking the quality of the predictions.

Researchers often lack a reliable stopping criterion for their testing. Experiments typically conclude based on the experimenter's intuition or resource depletion. This approach leads to inefficiencies in the experimental process outcome. Premature stopping of experiments results in inaccurate predictions, consequently affecting the accuracy of the estimated battery failure distribution. On the other hand, scientists may prolong their testing beyond what is required without checking the effect of additional testing on the improvement of the predictions until resources are exhausted.

To further improve our early failure prediction framework, we develop a stopping criterion to identify when additional testing of a certain battery no longer provides sufficient new information to warrant continued experimentation. To achieve this, we compare the predictions of two different GP models: GP Model 1, which uses all the collected data until that point in time to make its predictions, and GP Model 2, which only uses the data of the current battery to make its predictions using the same trained hyperparameters of Model 1. If the predictions of Model 2 differ sufficiently from those of Model 1, this battery is expected to provide additional information that GP Model 1 does not have. Therefore, one would continue testing this battery. If the predictions from Model 2 do not vary significantly from those of Model 1, then it is expected that the current battery does not provide useful information that improves the predictions of Model 1. In this case, stopping the testing of the current battery and starting another would be more beneficial.

Distribution entropy provides the basis for a statistical metric to measure the amount of expected change in the GP predictions when new information is added [43]. More specifically, we compare the relative entropy of the posterior GP distributions $p(\mathbf{f}^*|\mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\kappa}_p)$ — the predicted distributions — over the same grid in the input domain $x \in \mathcal{X}$ with n data points from both Models 1 and 2. This relative entropy is defined by the Kullback-Leibler (KL) divergence [59] between the two models and is considered the expected information gain from the battery being tested. To illustrate this description, consider two datasets, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, such that $\boldsymbol{D}_2 \subset \boldsymbol{D}_1$ and a set of hyperparameters $\boldsymbol{\theta}$. The KL divergence between the posterior distributions of Models 1 and 2 is defined as

$$\mathrm{KL}\big(\, p(\mathbf{f}_2^*|\mathbf{y}_2, \boldsymbol{\theta}) \parallel p(\mathbf{f}_1^*|\mathbf{y}_1, \boldsymbol{\theta}) \,\big) = \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\kappa}_{p_1}^{-1}\boldsymbol{\kappa}_{p_2}) + (\boldsymbol{\mu}_{p_1} - \boldsymbol{\mu}_{p_2})^\top \boldsymbol{\kappa}_{p_1}^{-1}(\boldsymbol{\mu}_{p_1} - \boldsymbol{\mu}_{p_2}) - n + \log\frac{|\boldsymbol{\kappa}_{p_1}|}{|\boldsymbol{\kappa}_{p_2}|}\right). \tag{10a}$$

When the value of the KL divergence falls below a predetermined threshold, we conclude testing for the current battery. The threshold is computed as a fraction of the average gain — relative entropy, or KL divergence — calculated from previous experiments. To ensure a meaningful comparison between Models 1 and 2 and assuming no prior experiments were performed, the implementation of the stopping criterion begins when at least two batteries have been fully tested.

To showcase the performance of this stopping criterion, we apply it to the synthetic data with different fraction levels. Fig. 3(a) shows the expected information gain, represented by the blue markers, as a function of the number of experiments assuming batteries are successively tested to failure. It also marks which experiments correspond to which batteries, with the vertical dotted lines separating the data from successive batteries. The results imply that the expected information gain within each battery generally decreases with testing, signifying that testing some batteries becomes less informative past a certain point, for instance, batteries 7, 11, 13, and 14. Other batteries remain informative until failure, for instance, batteries 4, 8, 9, and 10. It is beneficial to continue testing those batteries until they fail. We use a fraction of the average of previous information gained to differentiate between informative and non-informative batteries. The green and black dashed lines in Fig. 3(a) correspond to one-tenth and half of the average gain, respectively. When the expected information gain drops below the dashed lines, it is recommended to stop testing the corresponding battery.

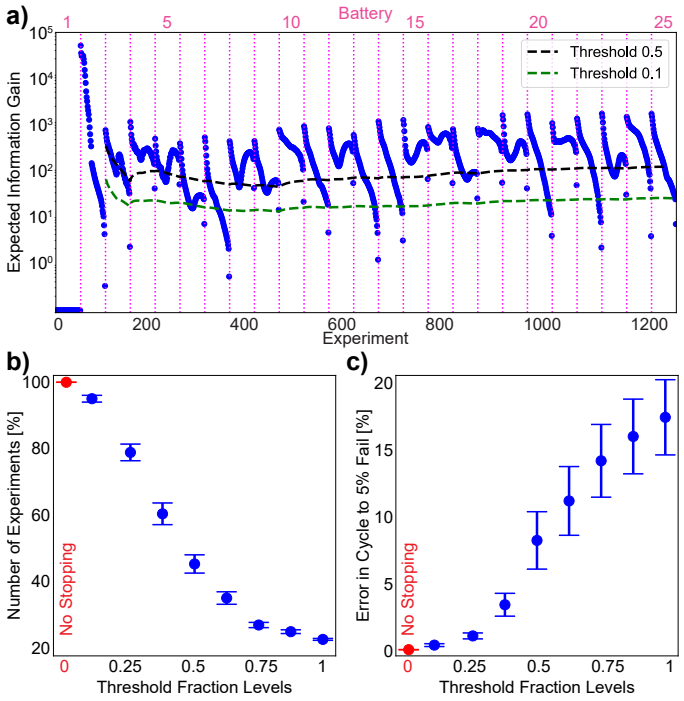

Fig. 3 Performance of the stopping criteria. (a) The expected information gain from each experiment of each battery is shown by the blue markers. The green and black dashed lines show the stopping threshold when using 0.1 and 0.5 of the average previous information gain, respectively. (b) Mean and its standard error of the number of experiments of 30 different permutations of batteries when the stopping criterion is implemented with different fraction levels of the threshold. (c) Similarly, the mean and its standard error of estimation error of the cycle number at which 5% of batteries fail as computed from failure probability. The zero fraction level corresponds to the No Stopping case.

We further studied the performance of the stopping criterion by considering various threshold fraction levels. We sequentially tested the 25 synthetic batteries, employing 30 sequence permutations. A battery is stopped when it reaches 1000 cycles — given by the available synthetic data for each battery — or its expected information gain falls below the threshold. For a visual reference, Fig. 3(a) represents one sequence permutation, and when the blue markers — the expected information gain — of a specific battery drop below the dashed line, testing of that battery is stopped. Fig. 3(b) shows the mean and its standard error of the 30 permutations of the number of experiments performed for each threshold fraction level, plotted as a percentage. Similarly, Fig. 3(c) shows the mean and its standard error of the inaccuracy in estimating the cycle number at which 5% of batteries fail calculated from the failure probability distribution as a function of the fraction levels. The zero fraction level refers to the case when the stopping criterion is not used, and therefore, 100% of the experiments are performed with 0% error in estimating the correct cycle number for 5% failure. The results show that as we increase the fraction level, fewer experiments are performed, and higher estimation error is incurred, as the probability of failure becomes less accurately quantified. As expected, there is a trade-off between the number of experiments performed and the estimation error, and based on the users' preference, one can choose what fraction level to use while doing their experiments. Fig. 3(b,c) demonstrates that the stopping criteria can save more than 50% of experiments while incurring less than 10% error.

3 Application to Experimental Data

Now that the GP model is tailored for our synthetic battery data, we show its effectiveness on experimental data. We use two sets of battery data found in the literature. The first dataset comprises 20 nominally identical pouch cells cycled similarly and retrieved from Harris et al. [45]. In this dataset, we are interested in fitting the battery energy as a function of the cycle number. (Although capacity fade is generally used in the field, this parameter ignores voltage fade, without which battery degradation cannot be properly evaluated.) The second dataset comprises 48 cells also cycled similarly and retrieved from Baumhöfer et al. [30]. For this dataset, we are interested in fitting the battery capacity as a function of the cycle number. (The energy data is not available to us.) Using those two datasets, we show that our approach is not only applicable to different datasets but also agnostic to the type of QoI (energy or capacity) used to quantify battery degradation. Both datasets are for lithium batteries with capacity degradation trends similar to the synthetic data used to tailor the GP model in the previous Section. Other types of batteries might have other capacity degradation trends, and the GP model may need to be modified to account for it.

3.1 Experimental Dataset 1

The GP model is tailored and trained to fit the experimental data from [45] to predict failure probability distributions and demonstrate the performance of the proposed stopping criterion. In our fitting process, we simulate real-world experiments, assuming that we use four concurrent channels. We collect data sequentially during cycling and update our GP model every 12 cycles, with each update considered a GP modeling step. Each step includes training the hyperparameters via MLE and predicting the posterior distribution across the input domain. To evaluate the performance of our stopping criterion, we run the complete simulation twice, with and without implementing the stopping criterion.

We incorporate domain knowledge to tailor our GP model for battery experimental data. We assume there are only two failure mechanisms, the first evident from the initial cycles, while the second becomes visible at the knee. We choose the 2-element piecewise polynomial function, equation 3, to model the GP prior mean. We also assume that the rate of energy degradation is constant before and after the knee. This results in a 2-element piecewise linear function with $p_1 = p_2 = 1$. For the noise model, we choose the power-law model, equation 7, to estimate the uncertainty. To account for non-stationarity and accurately estimate the probability of failure distribution, we use the deep kernel, equation 9.

Fig. 4 shows the GP fitting progression with implementing the stopping criterion. The blue markers represent the data points, the red line represents the GP predicted mean, and the gray region represents the uncertainty of the prediction. Fig. 4(a-c) shows the fitting of the data of the first set of four batteries at Step 5 (60 cycles), step 30 (360 cycles), and Step 50 (600 cycles). Starting with data within the first 60 cycles, Fig. 4(a) shows how the model accurately fits those data points in that region and follows the same path well beyond the 60 cycles mark until it hits the knee. At that point, the model randomly chooses the second slope, as our data does not provide any information about the slope beyond the knee, and therefore, any value is equally probable. This shows that with only a few data points, the GP model

learned that batteries degrade with cycling and that degradation will likely continue beyond the tested region. However, the GP compensates for its inaccurate second slope estimation by increasing its uncertainty bounds to show that the prediction may not be accurate beyond the knee. With the addition of new data points, Fig. 4(b), the posterior mean further follows these points and keeps fitting the data well as before, but now with a better understanding of how the degradation curve continues beyond the currently available experimental results. The posterior uncertainty also adapts to the data, accounting for the heteroscedastic variability as a function of cycle number. The bounds of the gray region increase with data variability. Comparing Fig. 4(b) and (c), with the latter representing the GP fit after finishing the first set of four batteries with 600 cycles, it can be seen that even though they were only at cycle 360, the GP predictions of the degradation were accurate. After completing the tests of these four batteries, four new batteries were tested. Fig. 4(d) shows the results after adding the data of those new batteries, where the predictions become slightly more accurate, and the uncertainty increases to account for the added variability in the data, as appropriate. This is also seen in Fig. 4(e) after adding the data for the third set of batteries. However, adding the third set of batteries does not significantly change the predictions. This is also seen in Fig. 4(f), which contains the data for all twenty batteries. This demonstrates that the testing should have been stopped earlier. These new data points did not add any information. We note that the failure probability density function is generated by cutting the gray-shaded region horizontally at the chosen failure point, as seen in Fig. 4(f) at 80% of the initial energy. The failure distribution is computed based on the GP model's normal probability density function values corresponding to the 80% failure level at each cycle number. We note that the progression in Fig. 4 depends on the order of the tested batteries. The results might differ during the initial steps if another sequence was used.

Applying the stopping criterion saves testing resources while keeping an accurate estimation of the failure probability. We repeat the earlier experiments with the same experimental setup with the proposed stopping criterion. We terminate battery testing when the expected information gain drops below half the average previous information gain. Fig. 5 compares the GP fits with and without applying the stopping criterion. Fig. 5(a) replicates Fig. 4(f). Fig. 5(b) shows the GP fit, the experimental data used when implementing the stopping criterion, and the corresponding failure probability. To better compare the failure probabilities, (c) plots the two failure distributions with and without the stopping criterion and the region at which 5% of batteries fail. Comparing both fits, it is evident that applying the stopping criterion significantly lowered the number of experiments while accurately predicting the failure distribution. As evident in Fig. 5(b), the batteries not expected to improve the GP predictions were stopped early on. Only the informative batteries were tested until failure. The results indicate that around 70% of the experiments were eliminated while incurring less than 3% error when estimating the cycle number until 5% of batteries fail.

3.2 Experimental Dataset 2

To showcase the framework's agnosticism to the battery data type, we perform the same analysis for the data by Baumhöfer et al. [30]. The available data from [30] is the batteries' discharge capacity as they are cycled. In our fitting process, we also assume four channels running concurrently. We collect data sequentially during cycling and update our GP model every 150 cycles, with each update considered as a GP modeling

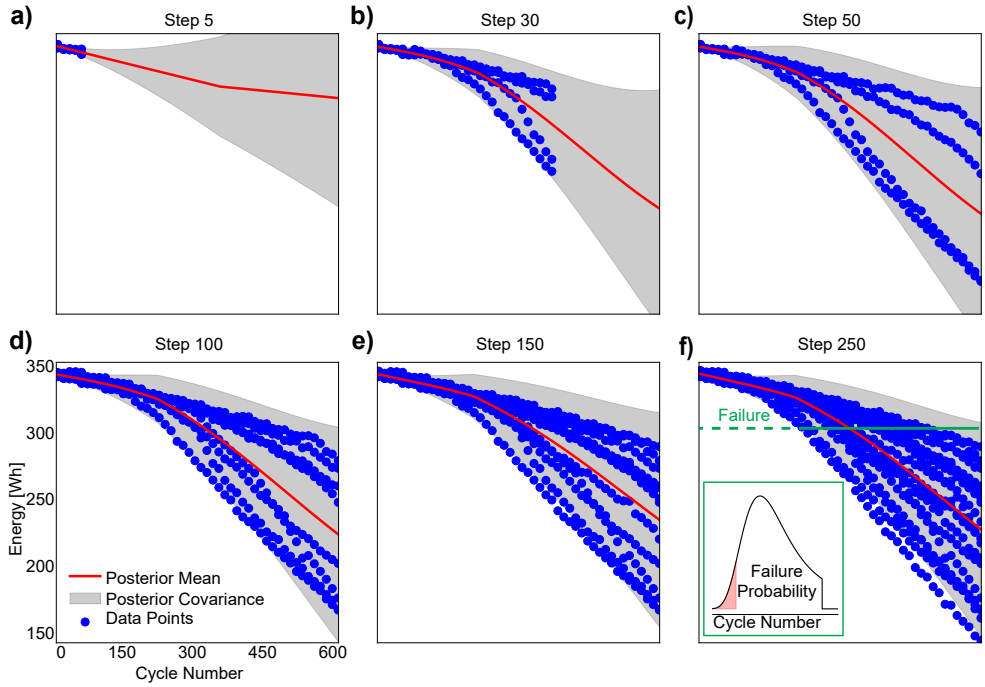

Fig. 4 GP model fitting progression, with the mean shown in the red line and uncertainty in the gray region, with the availability of the experimental data that is shown in the blue markers. Sets of four batteries are tested simultaneously, and data is collected sequentially at different steps. (a) Step 5, (b) Step 30, and (c) Step 50 represent the fitting for the first set of four batteries. (d) Step 100, (e) Step 150, and (f) Step 250 show the addition of the completed tests for the second, third, and fifth sets of batteries, respectively.

step. This process repeats 200 times to test all 48 batteries. Since the data here exhibit similar trends as earlier, we use the exact GP tailoring.

Fig. 6 shows the GP fitting sequence of the data as it is being collected. The blue markers represent the collected discharge data, the red line represents the posterior mean, and the gray region represents the uncertainty. Fig. 6(a-c) shows the progression as the first four batteries are added at (a) step 5 (800 cycles), (b) step 10 (1600 cycles), and (c) step 15 (2000 cycles). Similar to the results in Section 3.1, tailoring the GP model and incorporating domain knowledge provides a good prediction accuracy of the discharge curve as a function of cycles even before the cycling of these first four batteries ended at step 15. However, with these four batteries, the uncertainty is not accurately quantified as it increases significantly in Fig. 6(d-e) after completing the testing of 8 and 24 batteries at steps 30 and 100, respectively. This increase in uncertainty results from more variability in the experimental data. Fig. 6(f) shows the results of all 48 batteries with the failure probability assuming battery failure occurs at 80% of initial discharge capacity, where there is minimal improvement in the predictions compared to Fig. 6(e). This shows that beyond the 24 batteries, additional data did not provide sufficiently extra information to the model, encouraging the use of the proposed stopping criterion.

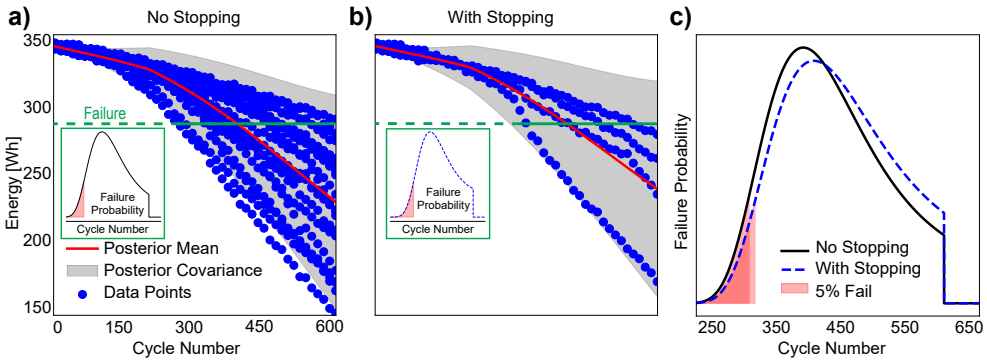

Fig. 5 Comparing the GP models and the failure probabilities when (a) the stopping criterion is not implemented — duplicating Fig. 4(f)–, and (b) when it is implemented. The blue markers represent the data used to train the GP model represented by the red line and gray region, corresponding to the posterior mean and covariance. To compare the failure probabilities, (c) plots the two distributions with and without the stopping criterion. The results show a significant decrease in the number of experiments (70% decrease) while keeping accurate predictions ($< 3\%$ error).

Applying the stopping criterion, with the same fraction as before, to the dataset by Baumhöfer et al. [30] showed significant experimental resource savings while keeping accurate predictions. Fig. 7 compares the GP fits with and without applying the stopping criterion. Fig. 7(a) replicates Fig. 6(f). Fig. 7(b) shows the GP fit and the experimental data when implementing the stopping criterion, along with the corresponding failure probability. To better compare the failure probabilities, Fig. 7(c) plots the two failure distributions with and without the stopping criterion and the region at which 5% of batteries fail. The results show that applying the stopping criterion resulted in 70% fewer experiments while incurring less than 5% error when estimating the cycle number until 5% of batteries fail. This also shows the effectiveness of our modeling and proposed stopping criterion in efficiently quantifying the failure probability accurately while being agnostic to the type of battery data used.

4 Discussion and Conclusions

Efficient early prediction of failure distributions for energy storage systems is crucial for utilities. Considerable research has been done to predict the expected life of batteries early on. However, even a perfect prediction for the expected life provides no insight into the failure distribution, which means that the predictions provide no information on how to price a warranty or estimate viability for a second life. In this work, we developed a framework based on GP modeling that integrates domain knowledge of the expected degradation and variation in the performance with cycling to allow for accurate extrapolation and quantification of failure distributions. We also developed a stopping criterion to avoid testing uninformative batteries, where an explicit trade-off between experimental efficiency and accuracy is found. This allows for the accurate early estimation of failure distributions with minimal testing.

We discussed in Section 2 how the performance of batteries can degrade with cycling at an increasing rate as more failure mechanisms occur. We integrated this knowledge into the GP model by customizing the prior mean function with a 2-element piecewise linear function, which flexibly models this degradation pattern. Comparing the

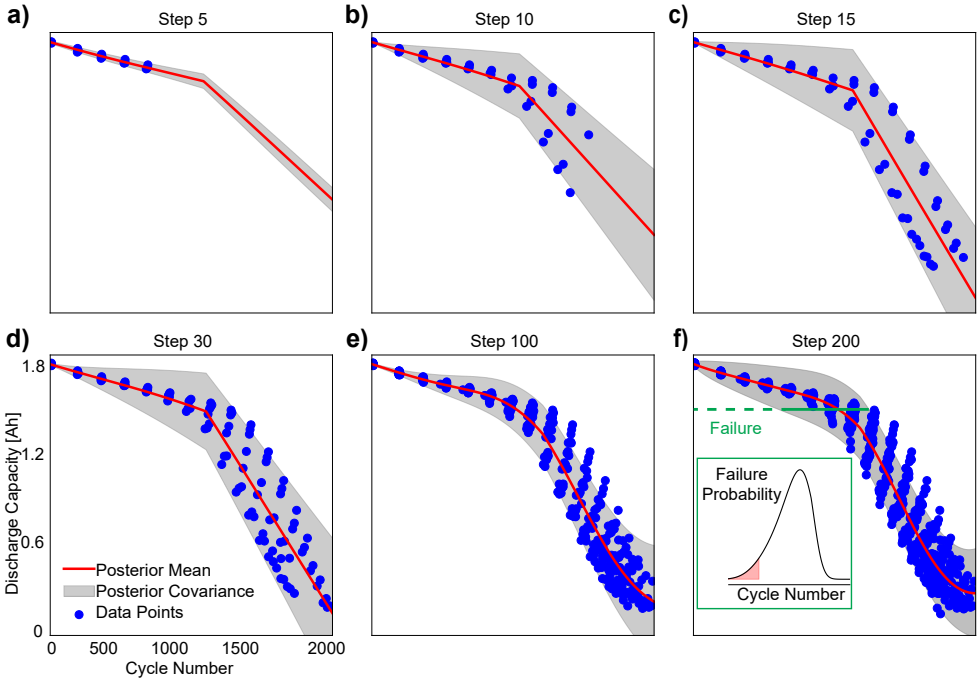

Fig. 6 GP model fitting progression, with the mean shown in the red line and uncertainty in the gray region, and the corresponding experimental data at each step shown in the blue markers. Sets of four batteries are tested simultaneously, and data is collected sequentially at different steps. (a) Step 5, (b) Step 10, and (c) Step 15 represent the fitting for the first set of four batteries. (d) Step 30, (e) Step 100, and (f) Step 200 show the addition of the completed tests for the second, sixth, and twelfth sets of batteries, respectively.

performance of the standard model — constant prior mean — with the present model showed significant improvement in the extrapolation capabilities as shown in the comparison of Fig. 1(c) and (d). Using this prior mean function to fit the experimental data sequentially also showed that even before finishing the testing of the first set of batteries, we had an accurate estimation of the expected performance degradation of the batteries. However, expected performance is not useful for setting warranties, which utilities would require to use new storage technologies. So, we tailored the noise function of the GP model to account for data variability as subject-domain-experts predict it to be. When comparing multiple functions, we chose the power-law function due to its flexibility and accuracy in fitting the data variability. Introducing this function as the noise function significantly enhanced the variability predictions as illustrated in Fig. 1(d) and (e). This is also demonstrated in the fits of the experimental data, where an accurate estimation of the posterior covariance was made early on. We continued improving our predictions by introducing a DNN non-stationary kernel function as shown in Fig. 1(e) and (f). The improvements due to non-stationary kernels depend largely on the data characteristics. We only saw slight improvements in our predictions because our dataset had little non-stationarity. However, we argue that using this DNN kernel was beneficial as it accounted for the slight non-stationarity in the data and improved the predictions. Using this kernel was also beneficial because it is a reference for interested readers to implement it for their applications.

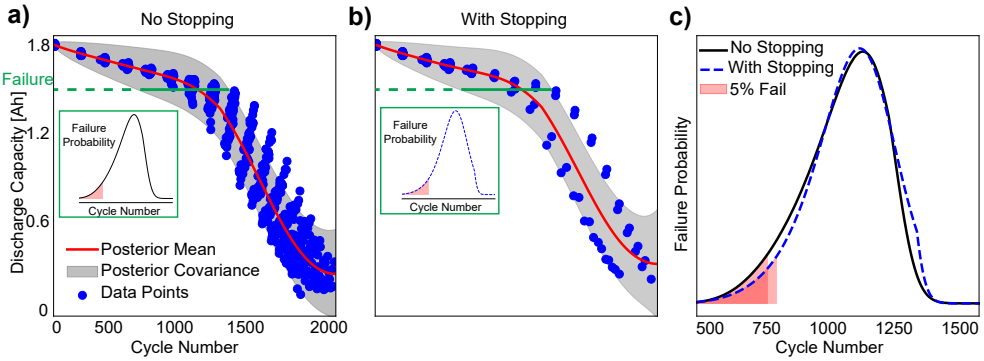

Fig. 7 Comparing the GP models and the failure probabilities when (a) the stopping criterion is not implemented — this is the same as Fig. 6(f)–, and (b) when it is implemented. The blue markers represent the data used to train the GP model represented by the red line and gray region, corresponding to the posterior mean and covariance. To compare the failure probabilities, (c) plots the two distributions with and without the stopping criterion. The results show a significant decrease in the number of experiments while keeping accurate predictions.

Tailoring the GP model to our application does not prevent it from being agnostic to the type of data used. As seen in Section 3, we used the same GP model to fit battery energy and discharge capacity as a function of cycle number. These two measures are different, but they have the same trend in terms of degradation and variability. We tailored our GP model to account for degradation and variability while keeping a flexible GP model, as we did not specify the values of the hyperparameters of the prior mean, noise, and the DNN kernel function. We allowed the model to learn these hyperparameters based on the data. This does not mean that our model will work for all applications. The GP model will likely need to be modified for other applications where the shape of the QoI and the variability are different. However, we showed how, intuitively, domain knowledge can be integrated to improve the GP model. Previous work considered modifying the GP model using physics-based degradation models [28, 60, 61]. However, these models are usually developed to model a specific failure mechanism [62]. We argue that using these in the GP model would bias the predictions according to the failure mechanism of the physics-based model used in the GPR. Here, we aim to develop an agnostic framework that can accurately predict failure distributions regardless of the underlying failure mechanism in the data and free the GP predictions from any possible bias. This is achieved using the general trend models discussed in Section 2.

To decrease the number of experiments, developing a stopping criterion for when additional battery testing is not informative was crucial. We based our stopping criterion on the expected information gain and showed its performance with different thresholds. The performance was quantified in terms of the number of experiments and errors in estimating the cycle number until 5% of batteries fail, which might be a warranty criterion. Applying the stopping criterion on the real-world experimental data showed up to 70% decrease in the number of experiments with less than 3% estimation error. These significant savings are also due to our use of the modified GP model, as it allowed us to predict, early on, accurate posterior mean and covariance with the least amount of data. Many more data points would have been needed if the GP model had not been modified.

Recent work considered efficient prediction and optimization of battery failure. For example, Attia et al. [63] developed a framework to predict battery life and optimize battery charging protocols using features from the first 100 cycles, following the work of Severson et al. [42]. In this approach, Severson et al. [42] and Attia et al. [63] focused on the average degradation of batteries and did not consider the variability in battery degradation when using the same charging protocols. They also did not quantify the failure probability for each charging protocol. In our work, we can quantify the variability of battery degradation and estimate the failure distribution.

Jiang et al. [31] extended the work of Attia et al. [63] to be one of the earliest to estimate battery failure distributions. Their approach required data from multiple charging protocols and their distributions, along with early testing data from a new protocol, to predict the distribution of this new protocol. Although our approach can learn the failure distribution of batteries of different charging protocols by intuitively extending the model to multiple input dimensions, it does not require data from other protocols, only data from the considered protocol. Additionally, their approach can only estimate a discrete failure distribution with the number of levels chosen *a priori* based on an assumed distribution family. Our approach does not have these constraints, as it estimates a continuous failure distribution without being restricted to a specific distribution family.

The work presented here is just one step toward fast validation of energy storage systems, and more work is needed. Here, we consider that degradation depends only on cycling. However, previous research showed that several other parameters could affect degradation, such as temperature, depth of discharge, and charging and discharging rates [64, 65]. Future work must generalize the framework developed here to account for multiple parameters simultaneously. In addition, efficient frameworks are needed to quantify the durability of batteries when these parameters are considered. Since battery tests are resource-intensive, testing all possible combinations of parameters would be prohibitively expensive. Moreover, the current framework will need to be compared to other approaches in terms of prediction accuracy and speed.

In conclusion, accurately predicting long-duration energy systems' failure probabilities is crucial for their integration into the grid to fight global warming. Utilities require the failure probabilities distributions as they are interested in estimating warranties. Although much work has been done on estimating the expected degradation of battery performance using either physics-based modeling or data-driven approaches, it does not help estimate the failure probabilities. Here, we integrated both approaches to estimate these failure probabilities early on with the minimum number of experiments. The key outcomes of this work are:

- An agnostic framework that integrates domain knowledge with a data-driven GP modeling
- A framework that has accurate extrapolation and uncertainty quantification
- Accurate predictions of failure probabilities with minimum testing
- A stopping criterion based on expected information gain that significantly saves on resources while keeping accurate predictions

Acknowledgments

This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy contract no. DE-AC02-05CH11231. We also thank Thorsten Baumhöfer and his colleagues for providing their data.

Conflict of Interest

All authors declare that they have no conflicts of interest.

Data and Code Availability

The data and the codes will be available upon the publication of the manuscript.

References

[1] Tianmei Chen, Yi Jin, Hanyu Lv, Antao Yang, Meiyi Liu, Bing Chen, Ying Xie, and Qiang Chen. Applications of lithium-ion batteries in grid-scale energy storage systems. *Transactions of Tianjin University*, 26(3):208–217, 2020.

[2] Jiangtao He, Zhongbao Wei, Xiaolei Bian, and Fengjun Yan. State-of-health estimation of lithium-ion batteries using incremental capacity analysis based on voltage–capacity model. *IEEE Transactions on Transportation Electrification*, 6(2):417–426, 2020.

[3] Jiabo Li, Min Ye, Yan Wang, Qiao Wang, and Meng Wei. A hybrid framework for predicting the remaining useful life of battery using gaussian process regression. *Journal of Energy Storage*, 66:107513, 2023.

[4] Xin Xiong, Yujie Wang, Kaiquan Li, and Zonghai Chen. State of health estimation for lithium-ion batteries using gaussian process regression-based data reconstruction method during random charging process. *Journal of Energy Storage*, 72:108390, 2023.

[5] Manh-Kien Tran, Manoj Mathew, Stefan Janhunen, Satyam Panchal, Kaamran Raahemifar, Roydon Fraser, and Michael Fowler. A comprehensive equivalent circuit model for lithium-ion batteries, incorporating the effects of state of health, state of charge, and temperature on model parameters. *Journal of Energy Storage*, 43:103252, 2021.

[6] Qi Zhang, Yunlong Shang, Yan Li, Naxin Cui, Bin Duan, and Chenghui Zhang. A novel fractional variable-order equivalent circuit model and parameter identification of electric vehicle li-ion batteries. *ISA transactions*, 97:448–457, 2020.

[7] Yang Li, Mahinda Vilathgamuwa, Troy W Farrell, Ngoc Tham Tran, Joseph Teague, et al. Development of a degradation-conscious physics-based lithium-ion battery model for use in power system planning studies. *Applied Energy*, 248:512–525, 2019.

[8] Zachary M Konz, Brendan M Wirtz, Ankit Verma, Tzu-Yang Huang, Helen K Bergstrom, Matthew J Crafton, David E Brown, Eric J McShane, Andrew M

Colclasure, and Bryan D McCloskey. High-throughput li plating quantification for fast-charging battery design. *Nature Energy*, pages 1–12, 2023.

[9] Yizhao Gao, Kailong Liu, Chong Zhu, Xi Zhang, and Dong Zhang. Co-estimation of state-of-charge and state-of-health for lithium-ion batteries using an enhanced electrochemical model. *IEEE Transactions on Industrial Electronics*, 69(3):2684–2696, 2021.

[10] Xi Zhang, Yizhao Gao, Bangjun Guo, Chong Zhu, Xuan Zhou, Lin Wang, and Jianhua Cao. A novel quantitative electrochemical aging model considering side reactions for lithium-ion batteries. *Electrochimica Acta*, 343:136070, 2020.

[11] Gaizka Saldaña, José Ignacio San Martín, Inmaculada Zamora, Francisco Javier Asensio, Oier Oñederra, and Mikel González. Empirical electrical and degradation model for electric vehicle batteries. *IEEE Access*, 8:155576–155589, 2020.

[12] Shuoqi Wang, Dongxu Guo, Xuebing Han, Languang Lu, Kai Sun, Weihan Li, Dirk Uwe Sauer, and Minggao Ouyang. Impact of battery degradation models on energy management of a grid-connected dc microgrid. *Energy*, 207:118228, 2020.

[13] Matthew B Pinson and Martin Z Bazant. Theory of sei formation in rechargeable batteries: capacity fade, accelerated aging and lifetime prediction. *Journal of the Electrochemical Society*, 160(2):A243, 2012.

[14] Felix Heinrich and Marco Pruckner. Virtual experiments for battery state of health estimation based on neural networks and in-vehicle data. *Journal of Energy Storage*, 48:103856, 2022.

[15] Xing Shu, Jiangwei Shen, Guang Li, Yuanjian Zhang, Zheng Chen, and Yonggang Liu. A flexible state-of-health prediction scheme for lithium-ion battery packs with long short-term memory network and transfer learning. *IEEE Transactions on Transportation Electrification*, 7(4):2238–2248, 2021.

[16] Weihan Li, Neil Sengupta, Philipp Dechent, David Howey, Anuradha Annaswamy, and Dirk Uwe Sauer. Online capacity estimation of lithium-ion batteries with deep long short-term memory networks. *Journal of power sources*, 482:228863, 2021.

[17] Gae-Won You, Sangdo Park, and Dukjin Oh. Diagnosis of electric vehicle batteries using recurrent neural networks. *IEEE Transactions on Industrial Electronics*, 64(6):4885–4893, 2017.

[18] Noman Khan, Fath U Min Ullah, Amin Ullah, Mi Young Lee, Sung Wook Baik, et al. Batteries state of health estimation via efficient neural networks with multiple channel charging profiles. *Ieee Access*, 9:7797–7813, 2020.

[19] Fan Xu, Fangfang Yang, Zicheng Fei, Zhelin Huang, and Kwok-Leung Tsui. Life prediction of lithium-ion batteries based on stacked denoising autoencoders. *Reliability Engineering & System Safety*, 208:107396, 2021.

[20] Laisuo Su, Mengchen Wu, Zhe Li, and Jianbo Zhang. Cycle life prediction of lithium-ion batteries based on data-driven methods. *ETransportation*, 10:100137, 2021.

[21] Xiaoyu Li, Changgui Yuan, Xiaohui Li, and Zhenpo Wang. State of health estimation for li-ion battery using incremental capacity analysis and gaussian process regression. *Energy*, 190:116467, 2020.

[22] Benjamin Larvaron, Marianne Clausel, Antoine Bertoncello, Sébastien Benjamin, and Georges Oppenheim. Chained gaussian processes to estimate battery health degradation with uncertainties. *Journal of Energy Storage*, 67:107443, 2023.

[23] Zicheng Fei, Fangfang Yang, Kwok-Leung Tsui, Lishuai Li, and Zijun Zhang. Early prediction of battery lifetime via a machine learning based framework. *Energy*, 225:120205, 2021.

[24] Xiaoyu Li, Changgui Yuan, and Zhenpo Wang. Multi-time-scale framework for prognostic health condition of lithium battery using modified gaussian process regression and nonlinear regression. *Journal of Power Sources*, 467:228358, 2020.

[25] Zhiyuan Wei, Changying Liu, Xiaowen Sun, Yiduo Li, and Haiyan Lu. Two-phase early prediction method for remaining useful life of lithium-ion batteries based on a neural network and gaussian process regression. *Frontiers in Energy*, pages 1–16, 2023.

[26] Sean Buchanan and Curran Crawford. Probabilistic lithium-ion battery state-of-health prediction using convolutional neural networks and gaussian process regression. *Journal of Energy Storage*, 76:109799, 2024.

[27] Kailong Liu, Yi Li, Xiaosong Hu, Mattin Lucu, and Widanalage Dhammika Widanage. Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries. *IEEE Transactions on Industrial Informatics*, 16(6):3767–3777, 2019.

[28] Robert R Richardson, Michael A Osborne, and David A Howey. Gaussian process regression for forecasting battery state of health. *Journal of Power Sources*, 357:209–219, 2017.

[29] Stephen J Harris and Marcus M Noack. Statistical and machine learning-based durability-testing strategies for energy storage. *Joule*, 7(5):920–934, 2023.

[30] Thorsten Baumhöfer, Manuel Brühl, Susanne Rothgang, and Dirk Uwe Sauer. Production caused variation in capacity aging trend and correlation to initial cell performance. *Journal of Power Sources*, 247:332–338, 2014.

[31] Benben Jiang, William E Gent, Fabian Mohr, Supratim Das, Marc D Berliner, Michael Forsuelo, Hongbo Zhao, Peter M Attia, Aditya Grover, Patrick K Herring, et al. Bayesian learning for rapid prediction of lithium-ion battery-cycling protocols. *Joule*, 5(12):3187–3203, 2021.

[32] Katharina Rumpf, Maik Naumann, and Andreas Jossen. Experimental investigation of parametric cell-to-cell variation and correlation based on 1100 commercial lithium-ion cells. *Journal of Energy Storage*, 14:224–243, 2017.

[33] Andrew Weng, Peyman Mohtat, Peter M Attia, Valentin Sulzer, Suhak Lee, Greg Less, and Anna Stefanopoulou. Predicting the impact of formation protocols on battery lifetime immediately after manufacturing. *Joule*, 5(11):2971–2992, 2021.

[34] Zihao Zhou and David A Howey. Bayesian hierarchical modelling for battery lifetime early prediction. *IFAC-PapersOnLine*, 56(2):6117–6123, 2023.

[35] Feng Leng, Cher Ming Tan, and Michael Pecht. Effect of temperature on the aging rate of li ion battery operating above room temperature. *Scientific reports*, 5(1):12967, 2015.

[36] AJ Smith, JC Burns, S Trussler, and JR Dahn. Precision measurements of the coulombic efficiency of lithium-ion batteries and of electrode materials for lithium-ion batteries. *Journal of The Electrochemical Society*, 157(2):A196, 2009.

[37] Madeleine Ecker, Jochen B Gerschler, Jan Vogel, Stefan Käbitz, Friedrich Hust, Philipp Dechent, and Dirk Uwe Sauer. Development of a lifetime prediction model for lithium-ion batteries based on extended accelerated aging test data. *Journal of Power Sources*, 215:248–257, 2012.

[38] Karl Ezra Pilario, Mahmood Shafiee, Yi Cao, Liyun Lao, and Shuang-Hua Yang. A review of kernel methods for feature extraction in nonlinear process monitoring. *Processes*, 8(1):24, 2019.

[39] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[40] Marcus M Noack and Kristofer G Reyes. Mathematical nuances of gaussian

process-driven autonomous experimentation. *MRS Bulletin*, 48(2):153–163, 2023.

[41] Weihan Li, Haotian Zhang, Bruis van Vlijmen, Philipp Dechent, and Dirk Uwe Sauer. Forecasting battery capacity and power degradation with multi-task learning. *Energy Storage Materials*, 53:453–466, 2022.

[42] Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fraggedakis, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.

[43] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[44] Marcus M Noack, Hengrui Luo, and Mark D Risser. A unifying perspective on non-stationary kernels for deeper gaussian processes. *arXiv preprint arXiv:2309.10068*, 2023.

[45] Stephen J Harris, David J Harris, and Chen Li. Failure statistics for commercial lithium ion batteries: A study of 24 pouch cells. *Journal of Power Sources*, 342:589–597, 2017.

[46] Suyeon Sohn, Ha-Eun Byun, and Jay H Lee. Two-stage deep learning for online prediction of knee-point in li-ion battery capacity degradation. *Applied Energy*, 328:120204, 2022.

[47] Paula Fermín-Cueto, Euan McTurk, Michael Allerhand, Encarni Medina-Lopez, Miguel F Anjos, Joel Sylvester, and Gonçalo dos Reis. Identification and machine learning prediction of knee-point and knee-onset in capacity degradation curves of lithium-ion cells. *Energy and AI*, 1:100006, 2020.

[48] Valentin Meunier, Matheus Leal De Souza, Mathieu Morcrette, and Alexis Grimaud. Design of workflows for crosstalk detection and lifetime deviation onset in li-ion batteries. *Joule*, 7(1):42–56, 2023.

[49] Elisa Braco, Idoia San Martín, Alberto Berrueta, Pablo Sanchis, and Alfredo Ursúa. Experimental assessment of cycling ageing of lithium-ion second-life batteries from electric vehicles. *Journal of Energy Storage*, 32:101695, 2020.

[50] Ruqing Fang, Peng Dong, Hao Ge, Jiangtao Fu, Zhe Li, and Jianbo Zhang. Capacity plunge of lithium-ion batteries induced by electrolyte drying-out: Experimental and modeling study. *Journal of Energy Storage*, 42:103013, 2021.

[51] Weiping Diao, Jonghoon Kim, Michael H Azarian, and Michael Pecht. Degradation modes and mechanisms analysis of lithium-ion batteries with knee points. *Electrochimica Acta*, 431:141143, 2022.

[52] Peter M Attia, Alexander Bills, Ferran Brosa Planella, Philipp Dechent, Goncalo Dos Reis, Matthieu Dubarry, Paul Gasper, Richard Gilchrist, Samuel Greenbank, David Howey, et al. "knees" in lithium-ion battery aging trajectories. *Journal of The Electrochemical Society*, 169(6):060517, 2022.

[53] Renato Miyagusuku, Atsushi Yamashita, and Hajime Asama. Gaussian processes with input-dependent noise variance for wireless signal strength-based localization. In *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–6. IEEE, 2015.

[54] Christopher Paciorek and Mark Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems*, 16, 2003.

[55] Paul D Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

[56] Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new

class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.

[57] Yalong Yang, Siyuan Chen, Tao Chen, and Liansheng Huang. State of health assessment of lithium-ion batteries based on deep gaussian process regression considering heterogeneous features. *Journal of Energy Storage*, 61:106797, 2023.

[58] Xizhe Wang, Xufeng Hong, Quanquan Pang, and Benben Jiang. Deep kernel learning-based bayesian optimization with adaptive kernel functions. *IFAC-PapersOnLine*, 56(2):5531–5535, 2023.

[59] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[60] Moritz Streb, Mathilda Ohrelius, Aamer Siddiqui, Matilda Klett, and Göran Lindbergh. Diagnosis and prognosis of battery degradation through re-evaluation and gaussian process regression of electrochemical model parameters. *Journal of Power Sources*, 588:233686, 2023.

[61] Jianwen Meng, Meiling Yue, and Demba Diallo. A degradation empirical-model-free battery end-of-life prediction framework based on gaussian process regression and kalman filter. *IEEE Transactions on Transportation Electrification*, 2022.

[62] Jorn M Reniers, Grietus Mulder, and David A Howey. Review and performance comparison of mechanical-chemical degradation models for lithium-ion batteries. *Journal of The Electrochemical Society*, 166(14):A3189–A3200, 2019.

[63] Peter M Attia, Aditya Grover, Norman Jin, Kristen A Severson, Todor M Markov, Yang-Hung Liao, Michael H Chen, Bryan Cheong, Nicholas Perkins, Zi Yang, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578(7795):397–402, 2020.

[64] Jacqueline S Edge, Simon O'Kane, Ryan Prosser, Niall D Kirkaldy, Anisha N Patel, Alastair Hales, Abir Ghosh, Weilong Ai, Jingyi Chen, Jiang Yang, et al. Lithium ion battery degradation: what you need to know. *Physical Chemistry Chemical Physics*, 23(14):8200–8221, 2021.

[65] Bibaswan Bose, A Garg, BK Panigrahi, and Jonghoon Kim. Study on li-ion battery fast charging strategies: Review, challenges and proposed charging framework. *Journal of Energy Storage*, 55:105507, 2022.