Predicting Discharge Reference Capacity of Lithium-ion Batteries

Allen Lewis, Angel Barrera, Anna White,
Joshua Owusu, Nathaniel Lancaster, Taylor Y. Boles

Executive Summary

This study explores methods to predict the reference discharge capacity of lithium-ion batteries without relying on their cycle history. By analyzing the voltage rebound curve—a phenomenon observed during rest periods after discharge—this approach provides a practical and cost-effective way to estimate capacity.

The research identifies key challenges, including overfitting issues when predicting capacity loss on a cycle-by-cycle basis. A broader, generalized modeling strategy is recommended to improve prediction reliability. Additionally, the study emphasizes the need for careful preprocessing of battery datasets, which often lack standardization. Ensuring that temporal dependencies are preserved during data preparation and evaluation is essential to avoid data leakage and to maintain model accuracy.

This work offers insights into battery performance modeling and highlights the importance of robust data practices and generalized approaches for accurate predictions.

- Battery data is often inconsistently collected, which requires thorough preprocessing before analysis.
- Discharge reference capacity can potentially be predicted independently of cycle history using individual voltage rebound curves.
- Overfitting makes it impractical to estimate capacity loss on a per-cycle or per-time measurement basis; generalized approaches are necessary.
- Time-series data must be handled carefully to avoid data leakage and ensure temporal dependencies are preserved during preprocessing and model evaluation.

1 Introduction and Motivation

Battery technology plays a critical role in modern energy storage solutions and powers everything from electric vehicles to renewable energy grids [16]. A crucial performance metric for battery cells is their capacity—the amount of energy they can store and deliver over time. Yet, as batteries undergo continuous cycles of charging and discharging, their capacity declines. This degradation is influenced by various factors including the number of cycles a battery experiences, the temperature conditions during operation, and specific charging protocols [7].

The accurate prediction of battery capacity, especially when dealing with missing data and inconsistent charge/discharge cycles, is essential for optimizing battery performance and predicting lifespan. In many real-world datasets, critical measurements such as capacity are often missing, particularly for cells that were partially charged or discharged. This missing data poses a significant challenge when attempting to model battery behavior over time.

In this report, we aim to address two key objectives: (1) predict the capacity of fully charged or discharged battery cells by utilizing multiple battery sources and (2) explore the potential for predicting capacity degradation in cells that were only partially charged or discharged. By employing machine learning techniques such as Random Forest modeling and advanced data imputation methods, we seek to fill the gaps in our datasets and provide actionable insights into battery performance. The findings will contribute to the growing body of research aimed at extending battery life and improving the reliability of energy storage systems in various applications. This study provides a detailed methodology for handling missing data, predicting discharge reference capacity for fully or partially discharged cells, and investigating degradation patterns under non-ideal charging conditions. The results have potential implications for industries reliant on battery performance, particularly those focused on sustainability and energy efficiency, such as electric vehicle manufacturers and renewable energy providers.

Battery capacity degradation is a crucial factor in determining the lifespan and efficiency of rechargeable batteries, especially in applications such as electric vehicles, consumer electronics, and energy storage systems. Accurate prediction of battery degradation can help optimize battery usage, prevent failures, and enhance the design of next-generation battery systems [13].

2 Data

The data explored in this study come from variety of sources, including the National Aeronautics and Space Administration (NASA) [12] [4], Oxford [2] [3], Toyota Research Institute [20], Berkeley[9], University College of London [10], and the Center for Advanced Life Cycle Engineering (CALCE) [5] at the University of Maryland.

2.1 NASA

The first NASA dataset contains battery aging data collected from a series of charge and discharge cycles organized as a MATLAB array. The data aims to capture key performance and health metrics of the batteries over time [12].

The dataset provides detailed cycle-specific information, including the time, voltage, battery and ambient temperatures, current, and the type of operation being performed on the cell. Additionally, data related to cycle time and reference capacity are included for each cycle [12]. Charging was carried out in a constant current (CC) mode at 1.5A until the battery voltage reached 4.2V and then continued in a constant voltage (CV) mode until the charge current dropped to 20mA. Discharge was carried out at a constant current (CC) level of 1A or 2A until the battery voltage fell to a specified voltage [12]. The ambient temperature is a constant 4 or 24 degrees Celsius. Some important details are missing or unclear. The type of batteries being tested are lithium ion, but no further information is given above the cell's chemistry [15]. Additionally, the experimental setup is lacking details on environmental control, such as humidity or pressure conditions.

The second NASA dataset contains cycling information for four 18650 LCO cells with 2.1 Ah capacity [1]. The are 24 additional batteries in this dataset that were not considered for this analysis. This group was cycled at a constant temperature of 23.78 degrees Celsius.

There were two types of cycling for each battery—Random Walk (RW) and pulsed load cycling. After every 50 RW cycles, a pulsed load cycle was carried out on the battery. RW cycling involves the current for both charge and discharge changing after a specified amount of time. This time is approximately every 5 minutes of the battery cycling. The current is randomly selected from a set of values: 0.5, 1, 1.5, 2, 2.5, 3, 3.5, or 4 A. The pulsed load cycling included a discharge profile that consisted of a rest period of 20 minutes followed by a loading period of 10 minutes at 1A. Randomized loads were applied during discharge to simulate more realistic battery usage (i.e. partial discharge).

2.2 Oxford

The Oxford dataset, also given in a mat file like the two NASA datasets, contains battery aging data from 8 Koham manufactured, small lithium-ion pouch cells [3], which each have a capacity of about 740 milliampere-hours. This data aims to mimic cells used to power electric vehicles, so measurements were taken every 100 drive cycles [2]. The information from the constant-current-constant-voltage profile includes time in seconds, voltage in volts, charge in milliampere-hours, and measured temperature in Celsius [3].

2.3 Toyota Research Institute

The Toyota Research Institute (TRI) dataset originally included data on 124 LFP/graphite cell batteries made by A123 Systems [20]. The cells had a nominal capacity and voltage respectively of 1.1 Ah and 3.3 V. They were cycled under 72 different fast-charging profiles. Each profile used varying charging rates and strategies. Comparatively, all of the cells used a consistent discharge method. During the cycling of the cells, the ambient temperature in a forced convection temperature chamber was kept constant at 30 degrees Celsius.

Data collection began at cycle 2 and continued until the battery reached 80% of its original capacity–coined as end of life (EOL). In-cycle measurements of temperature (Celsius), current (Amps), voltage (Volts), charge capacity (Ah), discharge capacity (Ah), internal resistance, and charge time (seconds) were recorded during this time [20].

The dataset was split into three parts based on different experimental groups. Each batch included approximately 48 cells where the experimental setup minimally varied between them. This analysis focuses on 5 cells (17, 22, 32, 33, and 46) from batch 3. This batch followed a two-step fast charging policy with a set charging time of 10 minutes to reach 80% state of charge (SOC). State of charge refers to how charged a battery is and is given by

$$SOC = \frac{100 \times C_t}{C_0} \tag{1}$$

where C_t represents the current capacity of the battery and C_0 represents the maximum capacity. The cells had 5-second rest periods placed at different periods during the cycling process–post-80% SOC, after an internal resistance test, before and after discharging. The cutoff current for charging and discharging was set at C/20 [20].

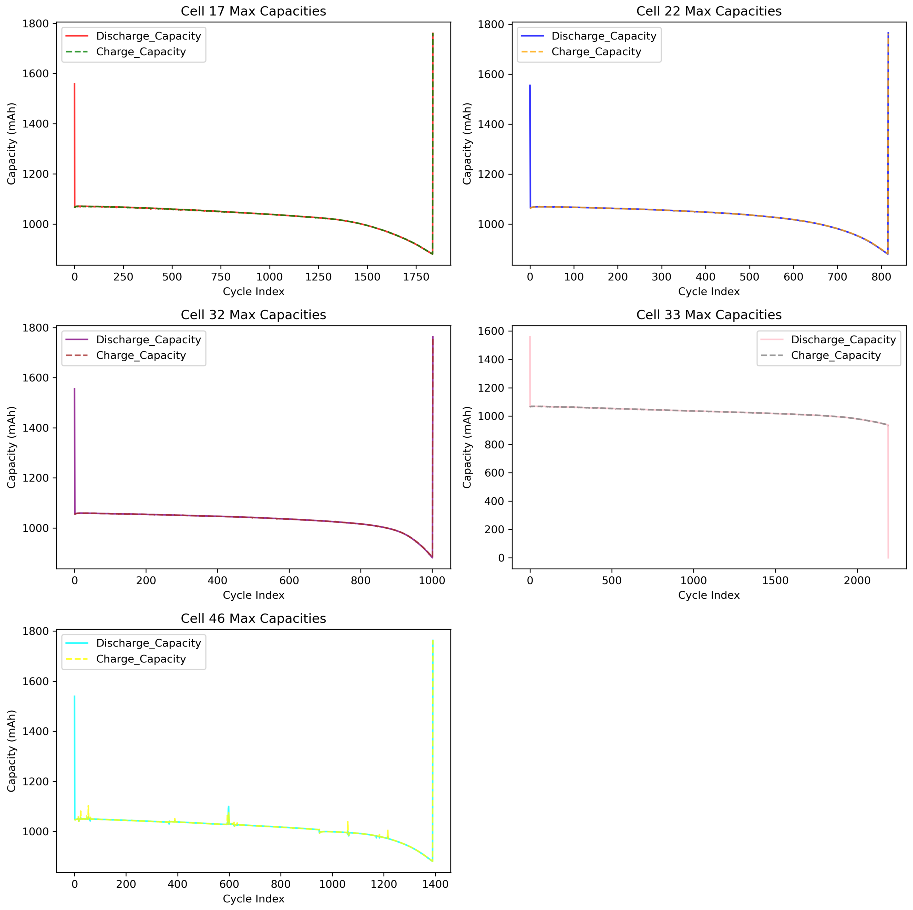


Figure 1: Each battery cell has a maximum charge and discharge capacity per cycle.

As shown in Figure 1, the maximum charge and discharge capacities for the five cells are tracked over their complete cycle life. In general, both capacities exhibit a gradual decline across all cells until a specific cycle point is reached. After that point, the decline becomes more pronounced until there is a sudden increase. This pattern is typical in battery degradation behavior. This EOL surge can be observed in all cells except for 33.

2.4 Berkeley

The Berkeley data set describes 42 different battery cells going through a series of different types of charging cycles [9]. Constant Current Constant Voltage (CCCV) involves charging the battery at a constant current until it reaches a certain voltage, then charging at a constant voltage. Constant Power Constant Voltage (CPCV) refers to charging at a constant power output up until a specific voltage level and then switching to a constant voltage. Power is given by

$$P = V \times I \tag{2}$$

where V is voltage and I is current. Multi-Current Charge is when the current varies over time during the cycle. Boostcharge involves rapidly charging the battery [17]. Where appropriate, the cycle types were further described by listing their current rate (such as 1 C, 1.2 C), which means the current was enough to charge 1 or 1.2 times the maximum capacity in one-hour respectively. These differing cycle types were used to study how varying charging methods impact variables like capacity fade and temperature.

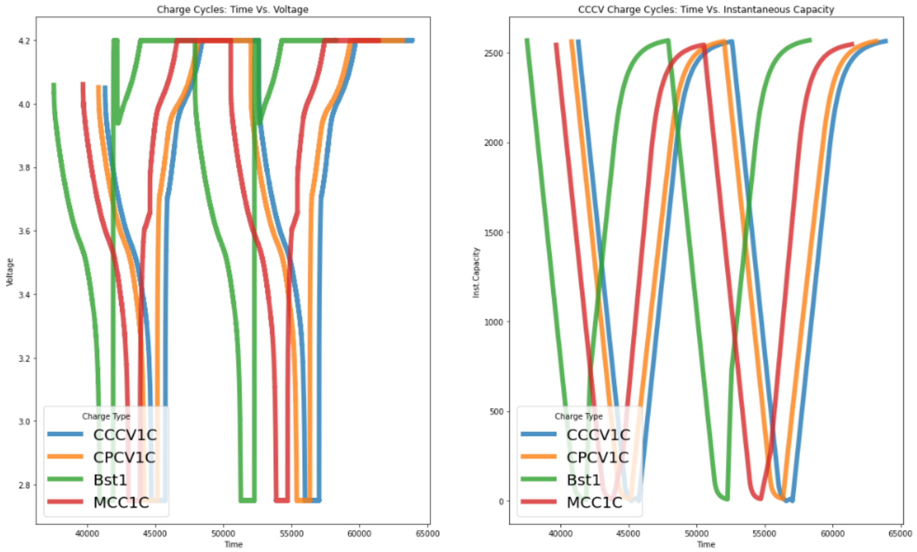


Figure 2: Differing charge methods result in varying frequencies and “shapes” of current and voltage.

The above figure is an example of how charge methods differ from each other.

2.5 University College of London

The University College of London data comes from a single lithium-ion battery cell with a nickel rich cathode and a graphite silicon anode [10]. The single cell was kept at an ambient temperature of 24 degrees Celsius and was cycled 400 times with a constant current of 1.5 amps for a charge until the cell reached 4.2 V. Then, voltage was held constant until the current fell to 100 milliamps. Lastly, the cell was discharged from 4.0 to 2.5 V. Within each of the 400 cycles, measurements of temperature (in Celsius), voltage (in Volts), and capacity (in mAh) were taken of the battery. On a per-cycle basis, the charge capacity (in mAh) and discharge capacity (in mAh) were recorded as well.

2.6 Center for Advanced Life Cycle Engineering

The Center for Advanced Life Cycle Engineering (CALCE) dataset describes 4 cylindrical battery cells with capacity of 2000 mAh. The batteries were charged at different temperatures to study how this would affect variables such as state of charge and open circuit voltage. Open circuit voltage or OCV represents the voltage of a battery when it is at rest, particularly after an adequate amount of time has passed since it was charged. Temperature can affect how batteries charge in a variety of mediums such as by altering the resistance. Figures 3 and 4 display how temperature can affect battery charge.

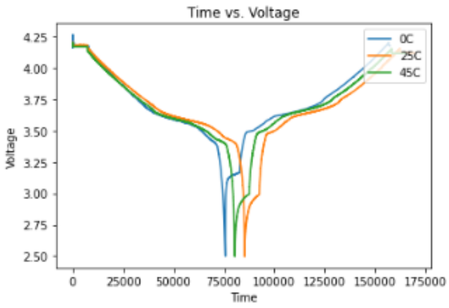


Figure 3: Temperature substantially affects the argmin.

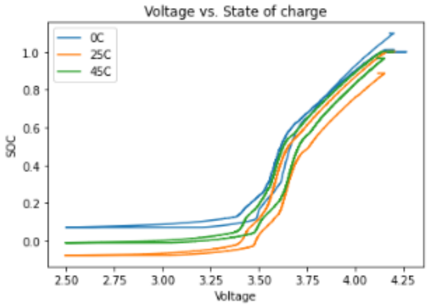


Figure 4: A colder battery has higher charge at terminal voltage.

2.7 Schema of Datasets

Table 1: Schema of Datasets

Variable	Type	Description
Source	String	Origin or dataset from which the data was collected.
Chemistry	String	The chemical composition of the battery (e.g., Lithium-ion).
Cell	String	The identifier for the specific battery cell being tested.
Cycle	Integer	A complete charge and discharge process of the battery.
Step	Integer	A stage within a battery cycle, such as charge, rest, or discharge.
Cycle Type	String	The classification of the cycle, typically distinguishing between charge and discharge cycles.
Time	Float	The timestamp or duration of the test step or cycle in seconds (s).
Voltage	Float	The electrical potential difference measured in volts (V) across the battery.
Instantaneous Capacity	Float	The battery capacity at a specific point during the cycle, measured in milliampere-hour (mAh).
Charge Capacity	Float	The total amount of electrical charge stored during the charging process measured in milliampere-hours (mAh).
Discharge Capacity	Float	The total amount of electrical charge released during the discharging process measured in milliampere-hours (mAh).
Temperature	Float	The temperature of the battery during the testing process, measured in degrees Celsius (C).
Ambient Temperature	Float	The surrounding environmental temperature during the test, measured in degrees Celsius (C).
Current	Float	The flow of electrical charge through the battery, measured in amperes (A).
Reference Capacity	Float	A benchmark capacity used for comparison measured in milliampere-hours (mAh).

Each dataset commonly recorded several lithium-ion battery attributes that will make them easy to adapt to a common schema. Almost all of them include measurements for voltage, current, ambient or cell temperature, and charge and discharge capacity for each cycle. These features are crucial for understanding a lithium-ion battery's performance across cycles. Current, time, and voltage are especially important when calculating both the reference and instantaneous capacity. These capacities will serve as the foundation in our analysis.

Table 1 shows the desired features in each of the cleaned datasets with their associated descriptions. Various features will be excluded from the datasets due to them being unnecessary for calculating and describing reference and instantaneous capacity. Excluding these features helps the datasets to remain focused on achieving the two goals of the analysis–(1) predicting the capacity of fully charged or discharged battery cells by utilizing multiple battery sources and (2) exploring the potential for predicting capacity degradation in cells that were only partially charged or discharged. The aim in any machine learning prediction should be to remove features that cause noise and provide no information.

Instantaneous capacity of a battery cell is calculated by integrating the current over the time of each charge or discharge cycle. This method approximates the total charge transferred during each cycle, providing a measure of how much energy has been accumulated or discharged at specific time intervals.

Instantaneous Capacity for a given cell is given by

$$C_t = \int_{t_0}^{t} I(t)\, dt \tag{3}$$

where

- C_t is the instantaneous capacity of the battery cell. This is measured in milliampere-hours (mAh). It represents the amount of charge that has been discharged or stored to a certain time t.
- t_0 is the starting time of the charging or discharging process.
- t is a specific time point within the process where the capacity is being recorded or measured.
- $I(t)$ is the current as a function of time. This is measured in amperes (A). It shows the range of charge flow at a given time during the process.
- $\int_{t_0}^{t} I(t)\, dt$ is the definite integral that calculates the charge accumulated or discharged from t_0 to a specific time t. It provides the amount of charge up to that specific time point rather than the total over the entire cycle.

Different battery data sets and battery cells have varying cycle types over different voltage and current ranges. To establish a consistent baseline for comparison between different cycles and datasets, the capacity is also measured at specific voltage points common among all sets referred to as the upper and lower voltage bounds. The upper bound typically represents a fully charged state, while the lower bound corresponds to a discharged state. The upper bound U can be given by

$$U = min\{max(V_1), max(V_2), ..., max(V_n)\} = sup(\bigcap_{k=1}^{n} V_k) \tag{4}$$

where each V_k represents the range of voltage values for cell k. Similarly, the lower bound L can be given by

$$L = max\{min(V_1), min(V_2), ..., min(V_n)\} = inf(\bigcap_{k=1}^{n} V_k) \tag{5}$$

This reference capacity provides a standardized measure of the energy delivered by the cell between these voltage levels, making it useful for comparing performance across different cycles and datasets. Reference Capacity for a given cell is given by

$$C_{\text{ref}} = \int_{t_U}^{t_L} I(t)\, dt \tag{6}$$

where

- C_{ref} is the reference capacity of the battery cell. This is measured in milliampere-hours (mAh). It represents the total charge that the battery can hold during a standard discharge cycle.
- t_U is the starting time of the discharge when the voltage is equal to the upper bound.
- t_L is the end time of the discharge process when voltage is equal to the lower bound.
- $I(t)$ is the current as a function of time. This is measured in amperes (A). It represents the rate at which charge is being discharged from the battery over time.
- $\int_{t_U}^{t_L} I(t)\, dt$: This integral calculates the total charge discharged from the battery over the time interval from t_U to t_L. It sums the current over time to see the total amount of charge released.

This integral can be rewritten in terms of the instantaneous capacity. Given that $\frac{d\hat{I}(t)}{dt} = I(t)$,

$$\begin{aligned}
C_{t_L} &= \int_{t_0}^{t_L} I(t)dt = \hat{I}(t_L) - \cancel{\hat{I}(t_0)} \\
C_{t_U} &= \int_{t_0}^{t_U} I(t)dt = \hat{I}(t_U) - \cancel{\hat{I}(t_0)} \\
C_{t_L} - C_{t_U} &= \hat{I}(t_L) - \hat{I}(t_U) \\
\hat{I}(t_L) - \hat{I}(t_U) &= \int_{t_U}^{t_L} I(t)dt = C_{ref} \\
C_{t_L} - C_{t_U} &= C_{ref}
\end{aligned}$$

Therefore, by determining the instantaneous capacities at these two points, a reference capacity was calculated as the difference between the instantaneous capacity at the lower voltage and the instantaneous capacity at the upper voltage.

It is important to note that the reference capacity was calculated in a slightly different manner for the NASA RW dataset. There was no need to calculate upper and lower voltage bounds. Both reference charge and discharge steps were isolated from the dataset associated with each cell and used in the integration of current over time for each unique step. Merging these reference capacities back into the main dataset presented issues because of the step filtering prior to the calculation of reference capacity. This issue was partially resolved by taking the average of each group type and applying it to the empty reference capacity values in each group. Any other empty reference capacity values were filled by linearly interpolating between the first non-empty reference capacity value in each group.

3 Modeling Data

3.1 Voltage Rebound Analysis

One topic of interest in helping determine the factors contributing to capacity degradation in lithium-ion battery cells is voltage rebound. This phenomena occurs during the rest state after a discharge. After a cell reaches the voltage where the discharge process is ceased, the current reaches zero and then the voltage will rebound. Looking at figure 5, it is clear that the voltage is discharged until about 2.5 V at a constant negative current. Once the voltage reaches 2.5 V, the current changes from a rate of about -2 amps to 0 amps. Then, when there are no longer any electrons flowing out of the cell, the voltage increases. The increase follows along a curve until it levels off at a certain asymptote.

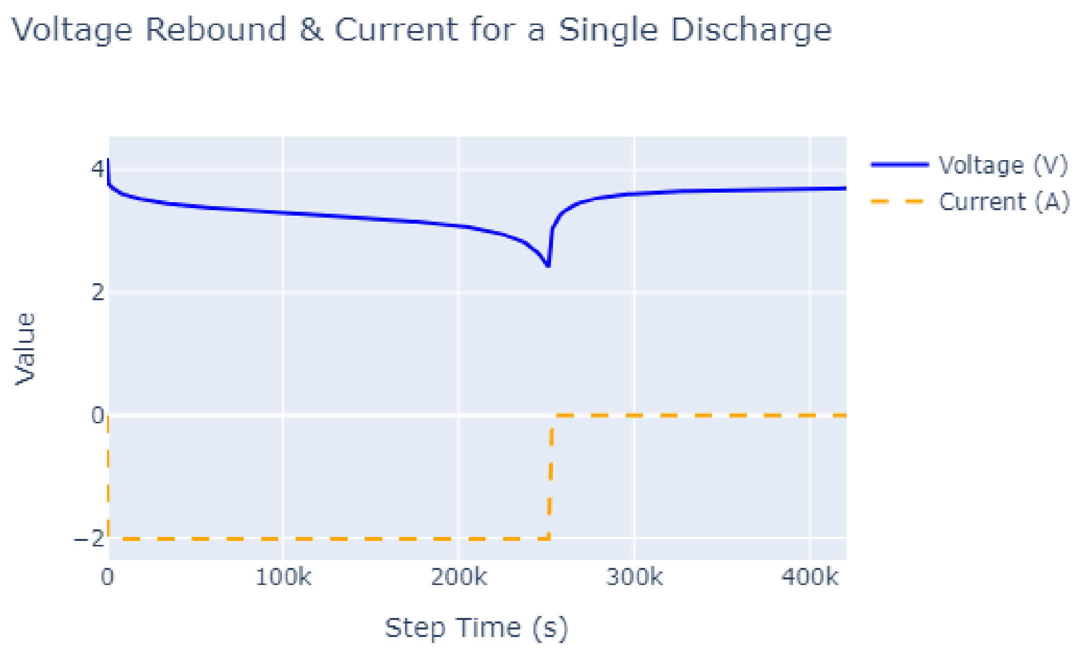


Figure 5: Voltage rebound observed after a single discharge step in a NASA battery.

The significance of modeling voltage rebound comes from its practical use cases. If, after a discharge, the cell always enters a state of rest when current is 0, and the voltage rebounds in a way that indicates the amount of capacity left in the cell, this could be done frequently and effectively at low-cost as a way of checking the capacity of a cell. Because of this simple, practical use-case, the objective of this analysis is to develop a method to predict the discharge reference capacity of a cell using properties of the voltage rebound after the discharge.

It is important to note that there is a lot of variety in when experimenters stopped measuring voltage and temperature after a discharge. This means that the full rebound is often missing, often only the beginning stages of rest after a discharge are observed. Because of this issue with data collection, the full rebound must be estimated. First, a function class must be determined as an adequate representation of the rebound across collected cells and their discharge steps. The goal is to define the appropriate function that can approximate the rebound using only a few points, which we can then use to derive the asymptote of the voltage rebound curve, as well as the speed it takes the cell to fully rebound, both of which are believed to be good indicators of the discharge reference capacity of the cell.

To summarize, a voltage rebound analysis will provide insight into the processes happening within the lithium-ion batteries after a discharge cycle. Though there may be challenges to this with incomplete rebound data, modeling the rebound curve provides a solid approach for estimating the current capacity of a battery cell. This analysis has great potential to be a straightforward and cheap way to understand and predict discharge reference capacity.

3.2 Functional Data Analysis and Functional Principal Component Analysis for Battery Capacity Degradation

Functional Data Analysis (FDA) is a statistical framework that treats data as continuous functions rather than discrete observations [14]. This approach is particularly useful for analyzing time-series data, where the underlying process is smooth and continuous. By modeling entire functions, FDA provides a more comprehensive understanding of dynamic systems, such as battery charge-discharge cycles. In FDA, data is represented as functions $X(t)$, where t is a continuous variable (e.g., time).

Each function is approximated using a set of basis functions $\phi_k(t)$:

$$X(t) = \sum_{k=1}^{K} c_k \phi_k(t) + \epsilon(t)$$

In functional data analysis, we approximate continuous curves using basis expansions where the coefficients c_k determine the weight of each basis function in the linear combination, the basis functions $\phi_k(t)$ provide the fundamental building blocks that can represent complex functional forms when combined and the error term $\epsilon(t)$ captures the difference between our approximation and the true underlying function.

A common representation of functional data uses a finite linear combination where c_k are estimated from observed data points, $\phi_k(t)$ are carefully selected to match the properties of the data, and $\epsilon(t)$ is minimized through optimization techniques.

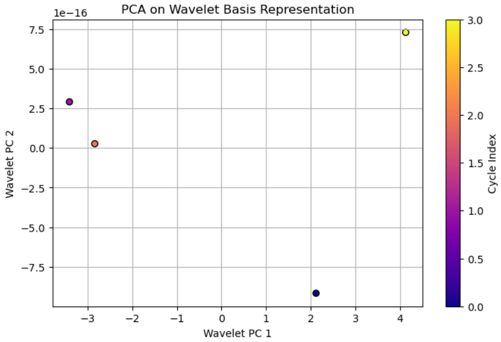


Figure 6: Wavelets

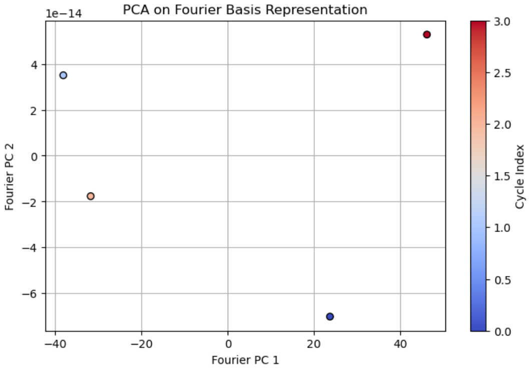


Figure 7: Fourier basis

Basis functions are building blocks for representing functional data. Common choices include B-splines, Fourier basis, and wavelets. B-splines are smooth, piecewise polynomials ideal for capturing localized features; whereas, Fourier basis is suitable for periodic data, such as temperature variations. Wavelets, on the other hand, are effective in analyzing signals with abrupt changes.

3.2.1 Functional Principal Component Analysis (FPCA)

Functional Principal Component Analysis (FPCA) is a dimensionality reduction technique for functional data [14]. It decomposes a set of functional observations $X_i(t)$ into principal components:

$$X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik} \phi_k(t) + \epsilon_i(t)$$

where $\mu(t)$ is the mean function, $\phi_k(t)$ are the principal component functions, ξ_{ik} are the principal component scores, K is the number of retained components and $\epsilon_i(t)$ is the error term.
To find this decomposition, you first need to center the data by subtracting the mean function from each observation. Then you compute the covariance function of your centered data, which captures how values at different points along the function covary. This covariance function is then decomposed using eigendecomposition (similar to PCA for multivariate data) to find the eigenfunctions (the $\phi_k(t)$) and eigenvalues that explain the most variation in the data.
Principal component scores ξ_{ik} are calculated by projecting each centered observation onto the principal component functions. These scores tell you how much of each component is needed to represent a particular observation.

The beauty of FPCA is that it often allows one to represent complex functional data using just a few principal components, making it much easier to analyze and visualize high-dimensional functional data. The number of components K is typically chosen based on the proportion of variance explained or using cross-validation approaches.

FPCA is particularly useful for battery data analysis for several reasons. Battery data exhibits functional characteristics: voltage, current, and temperature vary continuously over time or cycle number. When analyzing such data, FPCA effectively captures the main patterns of variation across cycles while filtering out measurement noise. Principal components extracted from battery data typically have physical interpretations. The first component often represents the primary capacity fade pattern, while subsequent components may capture specific degradation mechanisms or operational variations. By reducing complex battery time series to a small set of scores, FPCA simplifies comparison between multiple batteries. This reduction in dimensionality preserves essential features while making the analysis computationally feasible, even for large battery arrays.These principal component scores serve as effective features for machine learning models that predict remaining useful life or estimate state-of-health. Significant deviations in these scores from established patterns can indicate emerging failures before they become obvious through conventional monitoring methods. This capability makes FPCA valuable for battery management systems in applications ranging from electric vehicles to grid-scale energy storage, where monitoring numerous cells simultaneously is necessary for system reliability.

3.2.2 Functional Linear Model (FLM)

Functional Data Analysis (FDA) provides the theoretical foundation and methodological framework that naturally extends to Functional Linear Models (FLM). While FDA offers a comprehensive suite of techniques for representing and analyzing functional observations, FLM specifically addresses regression problems where either predictors, responses, or both are functional in nature. FLMs enable researchers to model relationships between functional variables through linear operators that map one function space to another, essentially extending classical linear regression to infinite-dimensional function spaces. In battery applications, this connection becomes particularly valuable—FDA techniques like FPCA can first extract the principal modes of variation from voltage curves or impedance spectra, and these functional components can then serve as inputs to FLMs that predict important outcomes such as remaining capacity or cycle life. This combined approach leverages FDA's strength in representing complex functional patterns while utilizing FLM's ability to establish predictive relationships between these functional features and scalar or functional responses of interest.

The relationship between functional predictors and battery degradation can be modeled using a Functional Linear Model (FLM) [14] is given by

$$Y = \alpha + \sum_{j=1}^{P} \int X_j(t)\beta_j(t)dt + \epsilon$$

where Y represents a scalar response like battery capacity fade, $X_j(t)$ are functional predictors such as voltage or current curves, $\beta_j(t)$ are functional coefficients, and ϵ is the error term [14].
In this equation, Y is the response variable that represents a key battery performance metric such as capacity loss or degradation rate. The functional predictors $X_j(t)$ are time-varying measurements collected during battery operation, including voltage profiles, current loads, and temperature fluctuations across the cell. The coefficient functions $\beta_j(t)$ quantify the influence of each functional predictor at different time points on the response variable. The error term ϵ accounts for random variations and measurement errors in the model.

By analyzing the coefficient functions $\beta_j(t)$, researchers can extract valuable insights about battery degradation mechanisms. The examination of these functions allows engineers to determine the impact of voltage fluctuations during specific portions of charge or discharge cycles on overall capacity loss. This approach enables battery scientists to understand the role of current variations during different phases of charge and discharge cycles, identifying particularly damaging operational regimes. Furthermore, the coefficient functions help researchers assess the effect of temperature changes throughout the battery's operation on its long-term performance and lifetime, revealing critical thermal thresholds that affect degradation rates.

In battery applications, this connection between FDA and FLM becomes particularly valuable—FDA techniques like FPCA can first extract the principal modes of variation from voltage curves or impedance spectra, and these functional components can then serve as inputs to FLMs that predict important outcomes such as remaining capacity or cycle life. This combined approach leverages

FDA's strength in representing complex functional patterns while utilizing FLM's ability to establish predictive relationships between these functional features and scalar or functional responses of interest.

3.2.3 Functional Random Forest Model

In addition to linear modelling, the utilization of non-linear regressors such as random forest can be applied. As a refresher, the random forest generates a series of decision trees that "vote" on the final outcome. This is a robust technique for both classification and regression, as it is able to handle the complex interactions and combinations between predictors that linear models cannot. However, the tradeoff comes in the fact that analyzing things such as p-values for indicators and other statistical benchmarks are harder to produce and analyze.

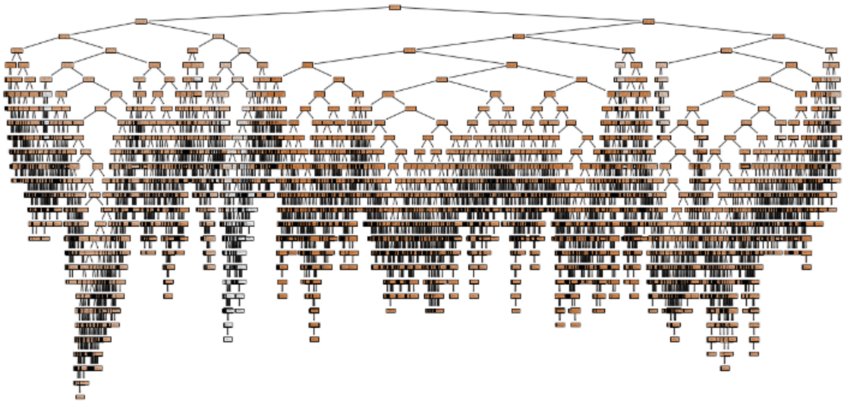


Figure 8: Example of a Decision tree.

4 Results

4.1 Voltage Rebound Analysis

4.1.1 Function Class to Fit the Rebound Curve

Lithium-ion battery datasets often stop measuring the voltage after a certain time has elapsed in seconds. This premature cutoff masks an observed potential voltage rebound phenomena in the cell. Only a few of the data sources allowed enough time to pass for the voltage rebound to be seen (NASA, NASA RW, and UCL). It is imperative that the voltage rebound of the other datasets be estimated for use in other analyses. The approach taken included finding the function class that best fit the rebound curve for the three full datasets. That function can potentially be used to estimate the remaining trajectory of the other datasets' voltage rebound curve in the future.

Four functions were fit to the datasets until one was found to have an excellent fit. The first function fit to the datasets was a power function and can be written as

$$f(a, x, b) = ax^b \tag{7}$$

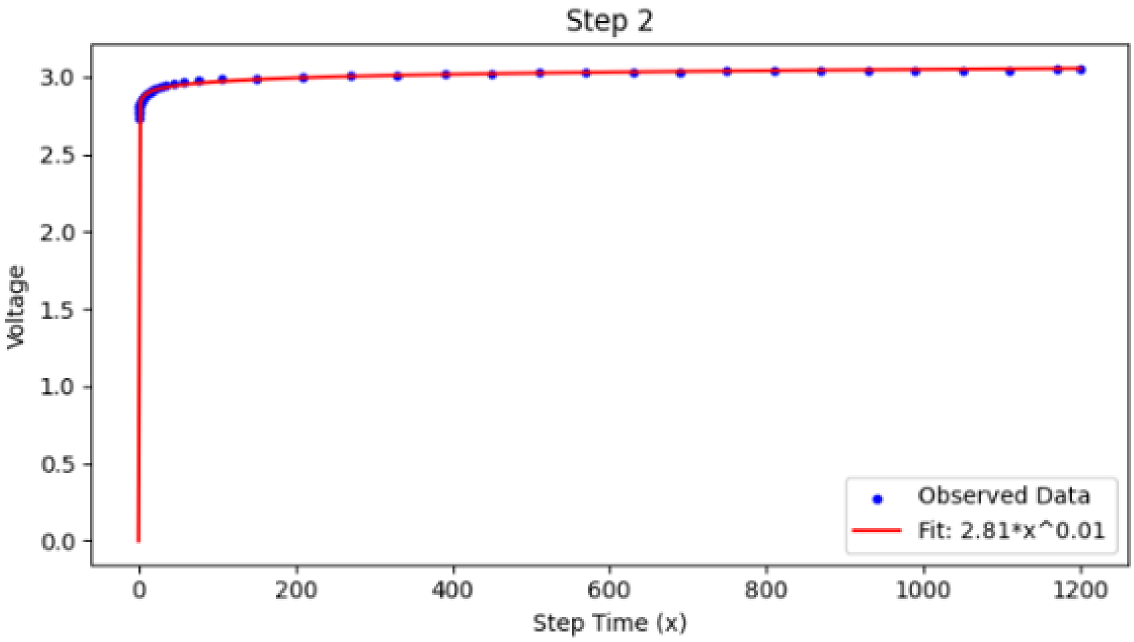


Figure 9: Power function on UCL.

Figure 9 shows this function fit on the UCL data. The fit of this function starts at zero, whereas the actual observed data does not begin until between 2.5 and 3.0 V. This function was ruled out as a good fit for the data. The second function fit to the datasets was another power function but with the addition of a constant. It can be written as

$$f(x, a, b, c) = ax^b + c \tag{8}$$

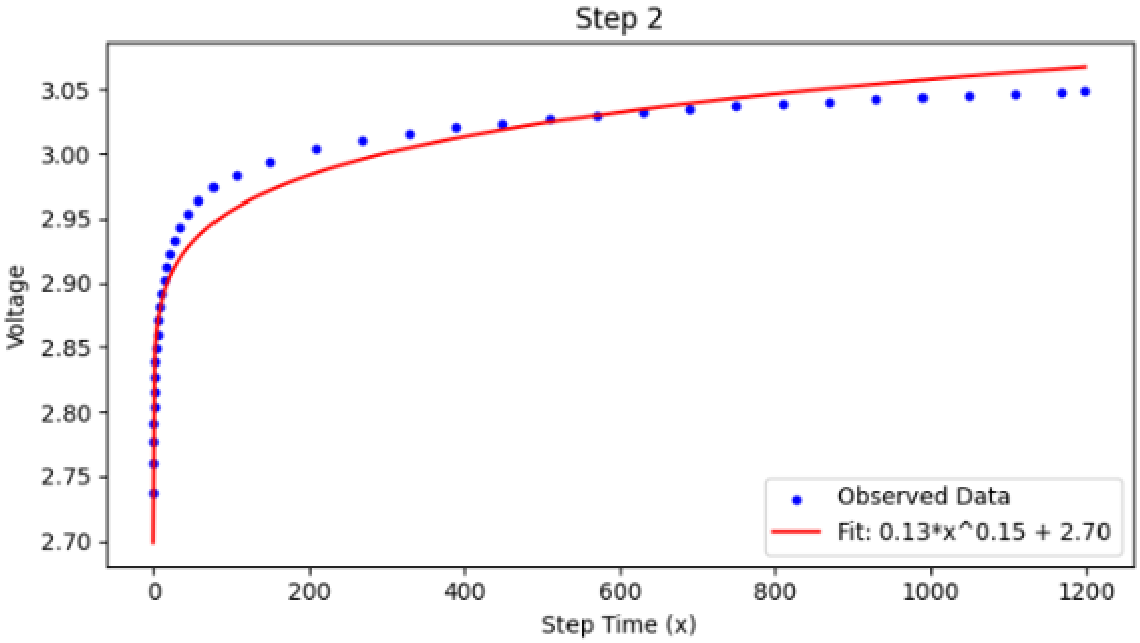


Figure 10: Power with constant function on UCL.

Figure 10 shows this function fit, again, on the UCL data. It underestimates the fit in the beginning, and as time progresses, it overestimates the fit to the data. This function was also ruled out as a potential estimator of the voltage rebound curve. The next function fit to the data was a type of modified exponential function. It can be written as

$$f(x, A, B, C) = A(1 - exp(-Bx)) + C \tag{9}$$

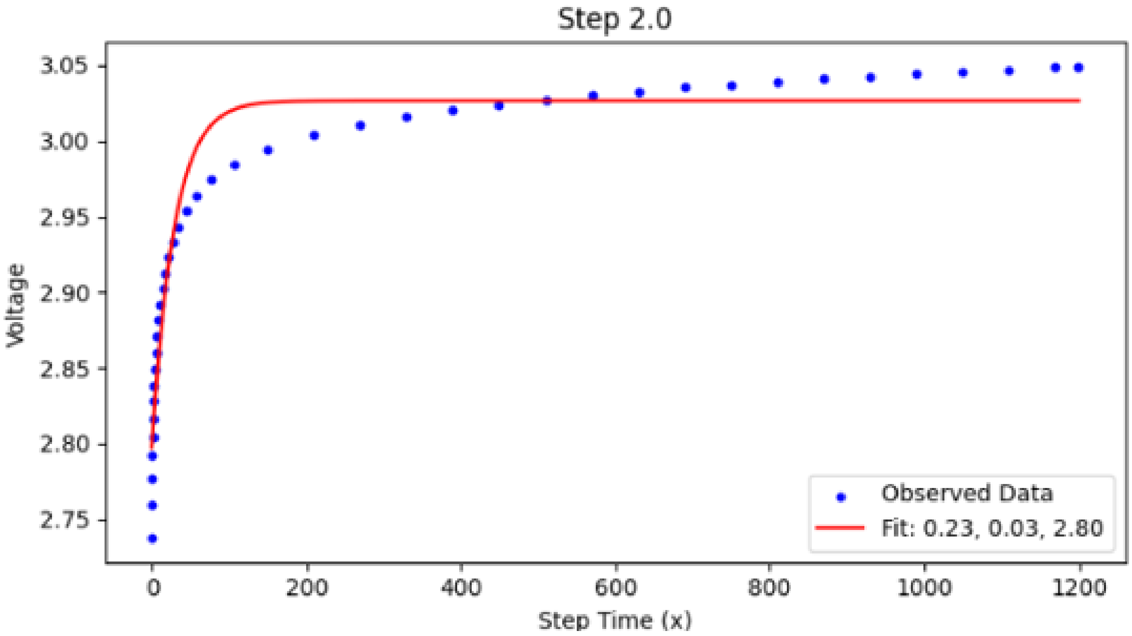


Figure 11: Modified exponential function on UCL.

Figure 11 shows this function fit on the UCL data. The fit overestimates the curve and then stays constant. This function was a poor fit to the data and was not considered any longer. The last function fit to the data was the logistic-like function. It can be expressed as

$$f(x, L, A, x_0, b) = L - \frac{A}{1 + \left(\frac{x}{x_0}\right)^b} \tag{10}$$

x represents the step time that begins at zero, L is the maximum value of the asymptote, A is the scaling factor, x_0 is the curve midpoint, and b is the growth rate of the curve. The scaling factor can also be described as the voltage at which the rebound begins. Additionally, the curve midpoint is the steepest point of the curve.

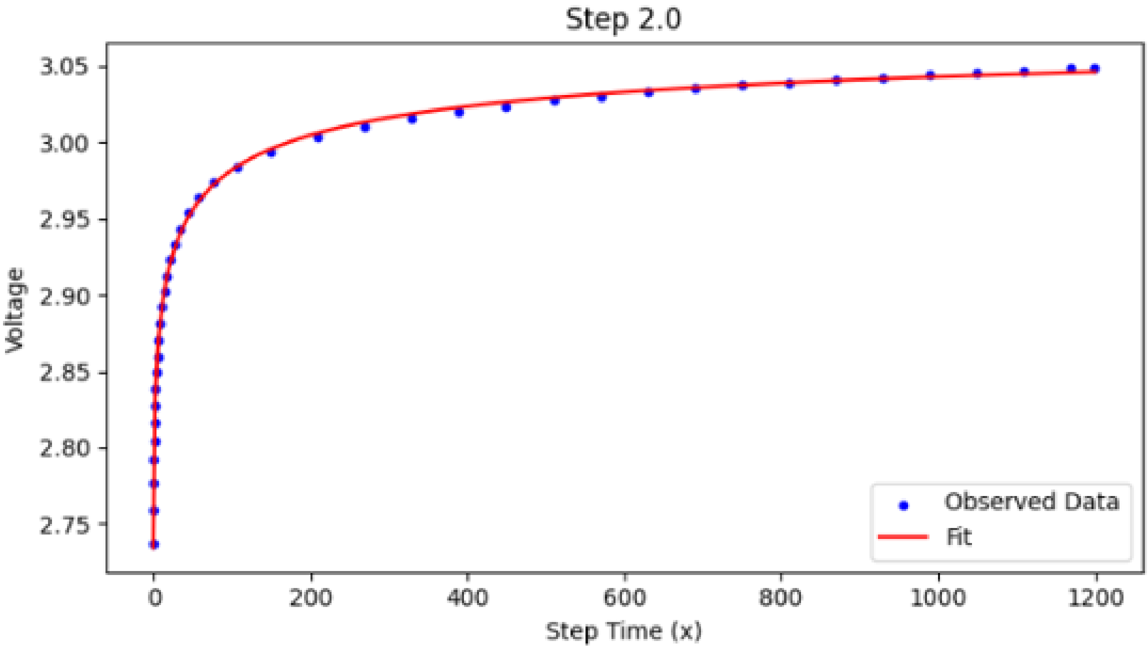


Figure 12: Logistic function on UCL.

Figure 12 fits this function to the UCL data. The fit is excellent and was considered for the two

NASA datasets, as well, where it fit even better. This type of logistic function was found to best estimate the voltage rebound curve in the full datasets. The question then pivoted to: approximately how much time (seconds) of a voltage rebound is needed for a good rebound curve estimation using the identified logistic-like function.

The question was answered by fitting the rebound curve on differently-sized subsets (variations of time within each step that experiences voltage rebound) of data for each qualifying NASA battery cell. Those cells included but were not limited to numbers 45, 47, and 48. The subsets of time include 30, 60, 120, and 180 seconds worth of time within each qualifying step. The estimated parameters derived from the logistic-like function for each subset were compared to those originating from the function fit on the full voltage rebound data. The percent differences between the subset parameters and the full data parameters were averaged to determine how well the function generally estimated the parameters across the subsets of time. Below are the results for each subset of time using the logistic-like function on the cell 45:

Time Window (s)	Average L % Difference	Average A % Difference	Average x_0 % Difference	Average b % Difference
30.0 (s)	5.73%	12.21%	0.96%	120.48%
60.0 (s)	9.18%	19.49%	20.68%	105.67%
120.0 (s)	5.85%	12.44%	15.21%	54.71%
180.0 (s)	3.77%	8.03%	10.62%	31.47%

4.1.2 Predicting Discharge Reference Capacity Using Voltage Rebound

This approach aims at using the estimated parameters from the fitted rebound curve as predictors for a single discharge step's reference capacity. This approach disregards a cell's previous history, so this could be implemented on a step-by-step basis as needed. After obtaining the fitted parameters from the logistic function, those parameters are used as features to predict reference capacity of the discharge step.

The first step in modeling was to check the predictors for outliers. An outlier in any of the fitted parameters indicates a poor fit of the logistic function to the rebound curve. A poor fit in this case is due to an unexpected shape of the rebound, that is shaped in ways that cannot be explained with the data available. Examining 13, the curve does not reach an asymptote in the observed data, so when the logistic function is fit, it says that the estimated maximum value of the asymptote (L) would be close to 1000. A voltage of 1000 is not a realistic value. So, cases like these should be removed from the reference capacity prediction.

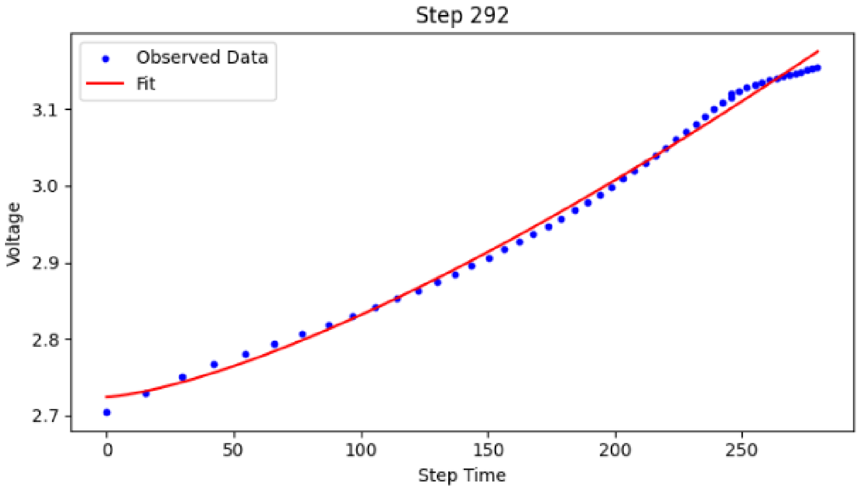


Figure 13: UCL - Poor Logistic Fit

After outliers were removed, that data was split into a training and testing set with 75% of the data assigned to the training set, and 25% of the data assigned to the testing set. It should be noted that the machine learning models were fit for the UCL, NASA, and NASA RW datasets separately, since these datasets followed different types of cells, with different charging protocols, and different nominal capacities.

4.1.3 Using Logistic Approximation of UCL Voltage Rebound (Full Discharge Steps)

Looking at the UCL data, the relationships between the logistic parameters and the target were close to linear Figure 14. This is especially true for the L and A parameters, but not as true for b and X_0. Additionally, there was a strong correlation between the L and A parameter (see Figure 15), so models that consider multicollinearity were considered as well.

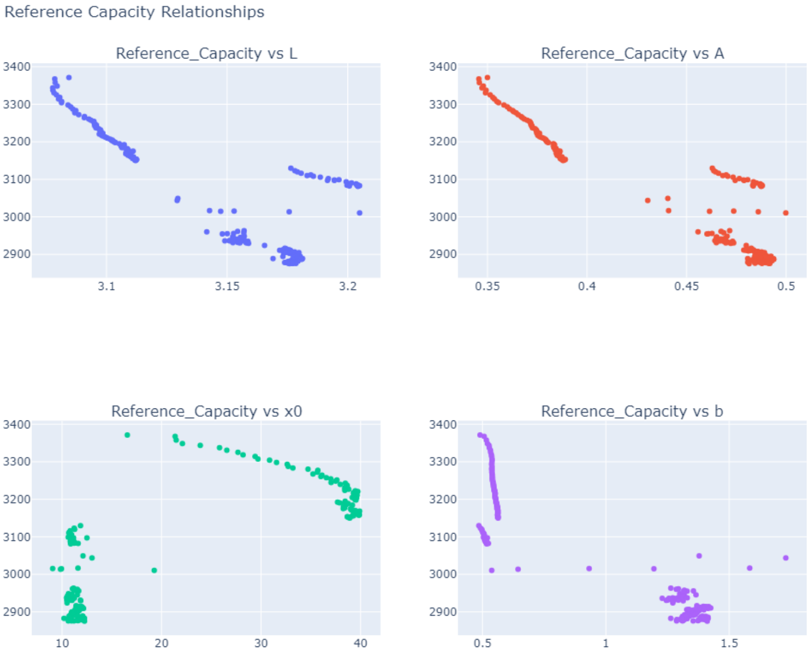


Figure 14: UCL - Reference Capacity versus Estimated Logistic Parameters

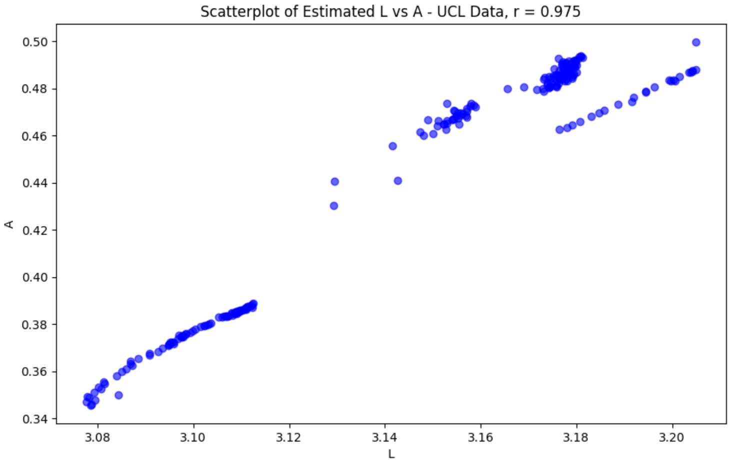


Figure 15: UCL - Estimated Logistic Parameters: L vs A

The models fit on the UCL data included linear regression, lasso regression, ridge regression, and elastic net regression. Linear regression is a statistical method that models a relationship between one or more independent variables and a continuous dependent variable by fitting a linear equation to the observed data [8]. To prevent overfitting and improve model generalization, regularization techniques such as ridge, lasso, and elastic net regression are employed. Ridge regression adds an $L2$ penalty term, which is proportional to the magnitude of the coefficient squared, thereby shrinking coefficients but not necessarily to zero [11]. Lasso regression introduces an $L1$ penalty term, in proportion to the absolute value of the coefficients. This results in some coefficients being exactly zero, which is like variable selection [11]. Elastic net regression unites both $L1$ and $L2$, so it is a hybrid model that balances the advantages of ridge and lasso regression [11]. The results on the testing set in table 2 show that the linear regression model outperformed other models. Next, in Table 2, we compare the estimated coefficients of the predictors between the linear regression and the lasso regression models.

Table 2: UCL - Machine Learning Modeling Results

ML Models	RMSE	% RMSE
Linear Regression	19.77	0.65%
Lasso Regression	21.87	0.72%
Ridge Regression	25.77	0.85%
Elastic Net	29.56	0.98%

Looking at the table 3 below, there are no strong differences in estimated coefficients of the linear and lasso regression models. Note that these models were fit on data that was normalized. So, this tells us that even after adjusting for possible multicollinearity, the linear model gives us the same idea of the relationship between the predictors and reference capacity as other models do. So, overall it looks as though the A parameter, which represents the scaling factor, is the most important for the UCL data reference capacity prediction.

Table 3: UCL - Linear & Lasso Regression Coefficients

Feature	Linear	Lasso
Maximum Temperature	13.71	12.56
b	-79.31	-73.75
X_0	-29.17	-32.98
A	-128.22	-154.25
L	23.14	41.25

4.1.4 Using Logistic Approximation of NASA Voltage Rebound (Full Discharge Steps)

Looking at the NASA data, the relationships between the logistic parameters and the target were not as close to linear like the UCL data (see Figure 16). This is true for the NASA RW data as well. Because of this, different machine learning models were used that are capable of handling non-linearity.

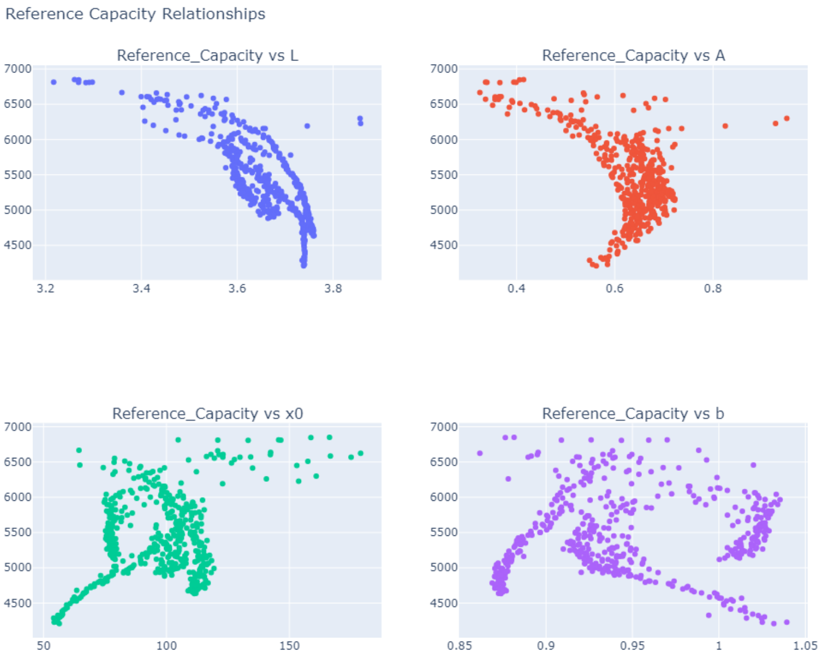


Figure 16: NASA - Reference Capacity versus Estimated Logistic Parameters

The models fit on the NASA and NASA RW datasets were tree-based methods including a decision tree regressor, random forest regressor, and an XGBoost regressor. A decision tress regressor is a non-linear model that uses a tree structure to split data into smaller segments based on feature thresholds [18]. A random forest regressor contains a forest full of individual decision trees to reach a conclusion [19]. XGBoost is a regressor that uses gradient boosting with regularization [6]. Note that all of these models were tuned as well. The results on the testing set in Table 4 NASA cells 5, 6, 7, and 18 show that while the decision tree performed the best, it was closely followed in performance by the random forest regressor. Additional cells in the NASA dataset yielded almost identical results with Random Forest always proving to be the best model.

Table 4: NASA (5, 6, 7, 18) - Machine Learning Modeling Results

ML Models	**RMSE**	**% RMSE**
Decision Tree	149.83	2.72%
Random Forest	152.52	2.77%
XGBoost	173.80	3.15%

4.1.5 Using Logistic Approximation of NASA Voltage Rebound (Subset Analysis)

The initial subset analysis results showed that the average percent difference between the full voltage rebound data and the subset of voltage rebound data generally increased as less time is used to estimate the voltage rebound curve. The ideal case would be that one could use as little as 30 seconds worth of voltage rebound data to estimate the parameters from the logistic-like function fit, and then use those in a predictive model to get a reliable prediction of the reference capacity. This would save time and money because less data during the rest step after discharge would have to be collected. Therefore, another question pivoted to whether or not using the parameters from subsets of less time would lead to increased error in the model. For this, cells 45 - 48 and a subset size of 30 seconds were considered in the model.

Table 5: NASA (Subset 30) - Machine Learning Modeling Results

ML Models	RMSE	% RMSE
Decision Tree	274.87	6.63%
Random Forest	155.05	3.74%
XGBoost	188.2	4.54%

Looking at Table 5, the error is still consistently low. It is slightly smaller than the full data error. This could be because the first 30 seconds worth of voltage rebound data contains less noise or because feature importance of the scaling parameter in the logistic-like function is high. Other subset sizes (60, 120, 180 seconds) in the model showed similar results.

4.1.6 Using Logistic Approximation of NASA Voltage Rebound (Partial Discharge Steps)

Another question to answer is whether the results from the prediction of the discharge reference capacity are consistent when only considering steps within cells that were partially discharged. The original NASA dataset always fully discharged cells. Due to this, another dataset had to be used for this partial discharge analysis–NASA Random Walk. The process for using this dataset to predict the discharge reference capacity was the same as the one with the other NASA dataset. First, we preprocessed the dataset in preparation for isolating the voltage rebound steps. This step included calculating both the instantaneous and reference capacities. The reference capacity had to be calculated using only the reference steps. Since this dataset included more than one type of step some rows did not have a value for the reference capacity. These empty rows had to be filled by linear interpolation between groups of step types. The second step in the process included isolating the steps where voltage rebound occurs during them. This was easier to do for this dataset because the rest after discharge steps were already separated from the rest of the steps. We had to make sure that the steps within the rest after discharge steps were filtered for voltage rebound by defining a function to return steps where voltage is increased and current was zero. Next, the logistic-like function was fit on the NASA random walk voltage rebound steps to obtain the same curve-estimation parameters. These parameters were then plugged into the prediction model to see how well they predicted the discharge reference capacity. The results were consistent with those from the full discharge data and can be observed in Table 6.

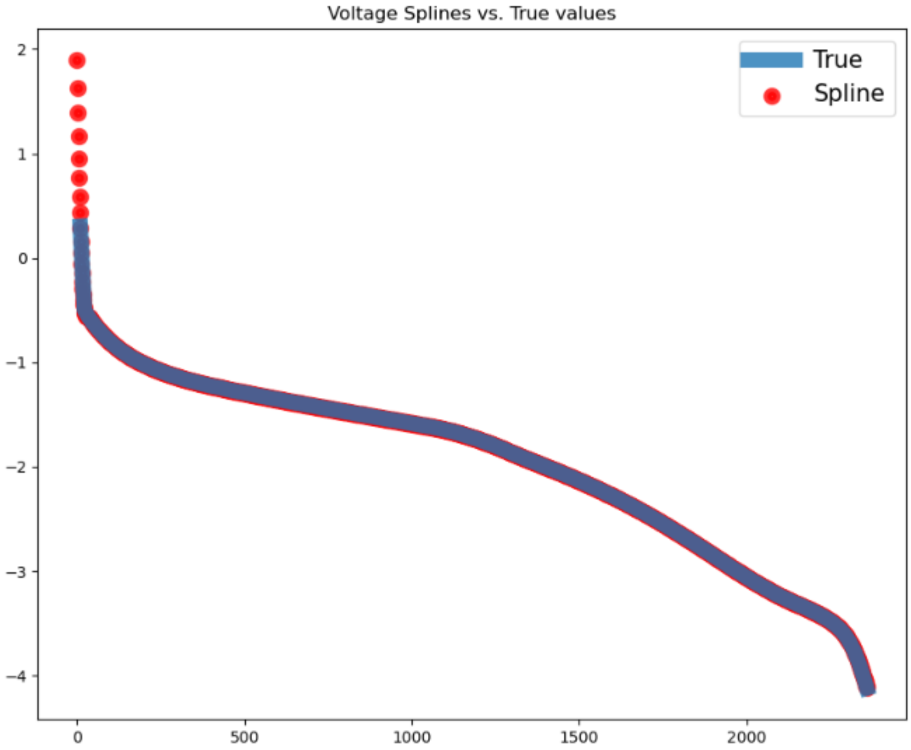
Table 6: Random Walk NASA - Machine Learning Modeling Results

ML Models	RMSE	% RMSE
Decision Tree	382.15	7.48%
Random Forest	282.38	5.53%
XGBoost	290.19	5.68%

4.2 Reference Capacity Prediction

4.2.1 Functional Principal Component Analysis

This analysis begins by taking each battery cycle and interpolating all of them into a common timespan using cubic splines. For example, one charge cycle with 20,000 observations can be made comparable to another cycle, such as one that is 4,000 observations long. This allows one to construct an $m \times n$ matrix with m being the number of samples per cycle (in this case 2,000) and n being the total number of cycles in the lifespan of the battery. Here we have an example of a spline approximation of the voltage of a battery during a discharge:



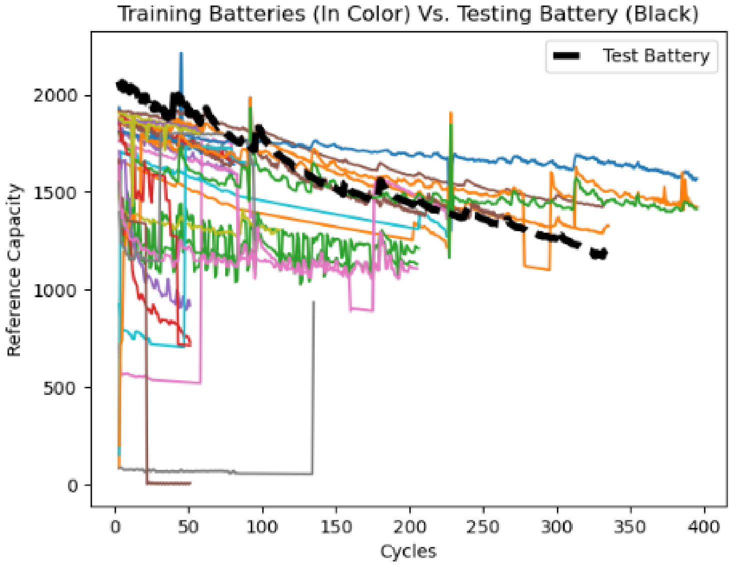
Note: The voltage of the battery is negative because the data has undergone standard scaling. Next, each column is converted into a sum of basis functions. Recall that each column X_i is treated as a function of time, such that

$$X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t) + \epsilon_i(t)$$

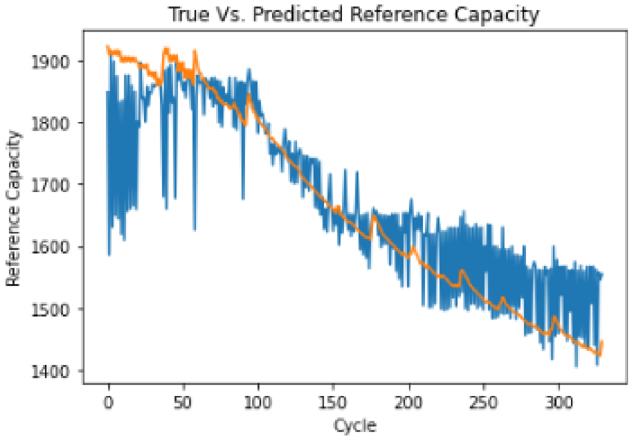
In other words, each column is treated as a linear combination of basis functions $\phi_k(t)$. From this, the coefficients ζ_{ik} can be extracted. Thus, the reference capacity for each cycle i is a function of coefficients ζ_{ik}.

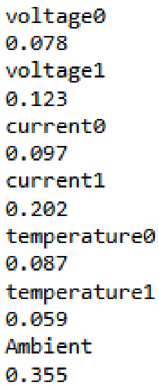
4.3 Random Forest Results

Because of the time-series nature of the data, training data is extracted from the entirety of the life-cycle of multiple batteries and can be used to estimate the reference capacity of a particular battery cycle. In this example, there are a multitude of different batteries with different conditions being used as a robust training set for the test battery as shown:



The primary reason for the wide variety in training sets has to do with batteries being tested at different temperatures, which is why ambient temperature is also included as an indicator. Shown as the predicted reference capacities using two components per indicator and with 18 random forest estimators.





```
voltage0
0.078
voltage1
0.123
current0
0.097
current1
0.202
temperature0
0.087
temperature1
0.059
Ambient
0.355
```

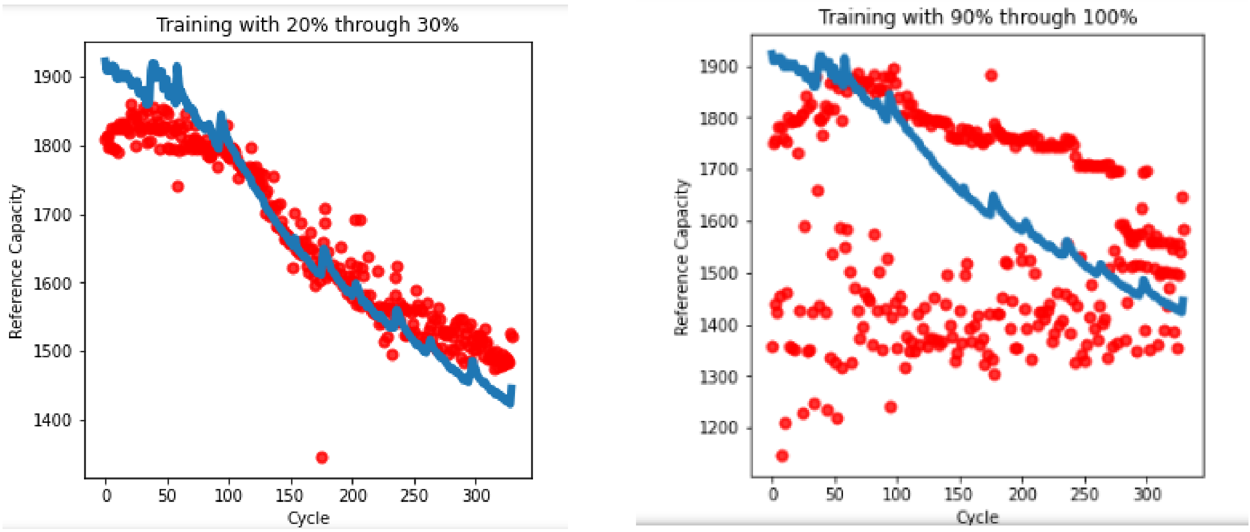
(a) Predicted Reference Capacity $R^2 = .77$

(b) Feature Importances

Figure 17

The model reasonably fits the general trend of the data, with ambient temperature being the most important predictor.

One additional investigative quandary is the portion of the cycle most indicative of battery damage. To investigate this, slices of the battery cycle were used to generate training and testing data rather than the full cycle.

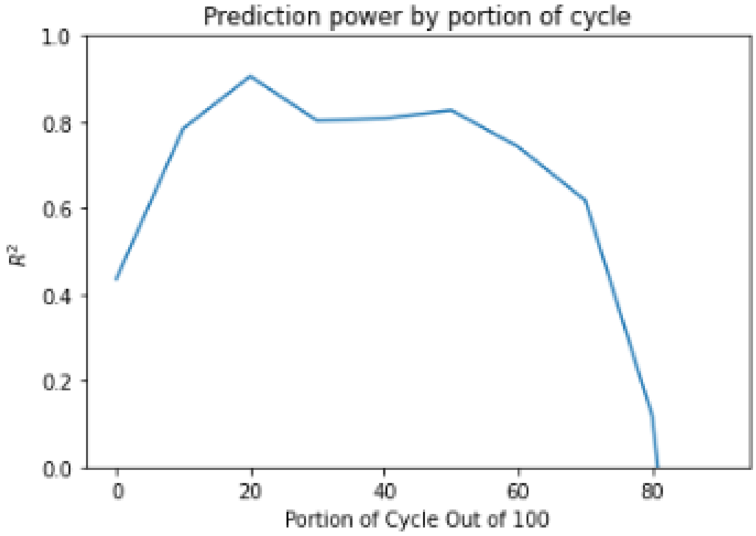


(a) Fairly accurate results for this slice

(b)

Figure 18: Excessive variance for this particular section

The full scores for each section are shown below:



The best section for prediction is 20% through 30% posessing an R^2 of .9, with the lowest being sections 90% through 100% that has a negative R^2 score.

5 Preview: Using State of Charge (SoC) to Understand Battery Health

This section introduces a preliminary exploration into calculating the State of Charge (SoC) from a Random Walk dataset. Instead of a finalized analysis, what follows is a forward-looking proposal, showcasing an early-stage demonstration of SoC estimation. The goal is to highlight SoC's potential as a compact, interpretable feature for future models aimed at assessing battery performance and degradation.

5.1 Why SoC Matters

State of Charge (SoC) is commonly used to represent the remaining energy in a battery, typically as a percentage of its total usable capacity. In electric vehicles and battery-powered systems, it provides a real-time estimate of how much charge is left. However, beyond operational monitoring, we believe SoC has untapped potential in predictive modeling—for instance, in estimating the battery's Reference Capacity, a key indicator of long-term health.

While this dataset does not include SoC directly, we propose calculating it using the following formulation:

$$\mathrm{SoC}(t) = \left(1 - \frac{Q_{\mathrm{used}}(t)}{Q_{\mathrm{max}}(t)}\right) \times 100$$

where $Q_{\mathrm{used}}(t)$ is the instantaneous capacity consumed up to time t, and $Q_{\mathrm{max}}(t)$ is an estimate of the reference (maximum) capacity. This equation allows for time-series computation of SoC from the raw current and time values recorded during each battery cycle.

5.2 Early Results: A Glimpse into the Possibilities

To illustrate this proposal, we present two initial results that offer a visual intuition for how SoC behaves in the early stages of the dataset.

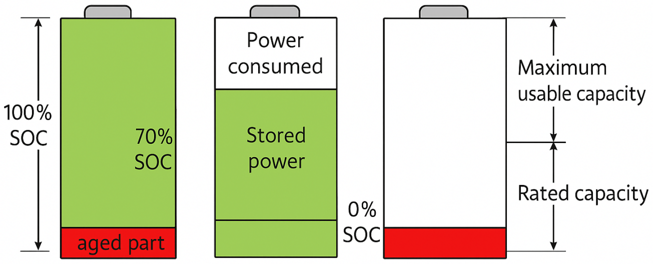


Figure 19: Conceptual illustration of SoC: stored energy, aging effects, and usable capacity.

Figure 19 serves as a visual aid to understand how SoC reflects usable battery energy over time. The diagram shows how aging reduces the effective capacity of the battery, leaving a portion unusable even when the battery appears full. This highlights the need for tracking SoC dynamically, especially as degradation progresses.

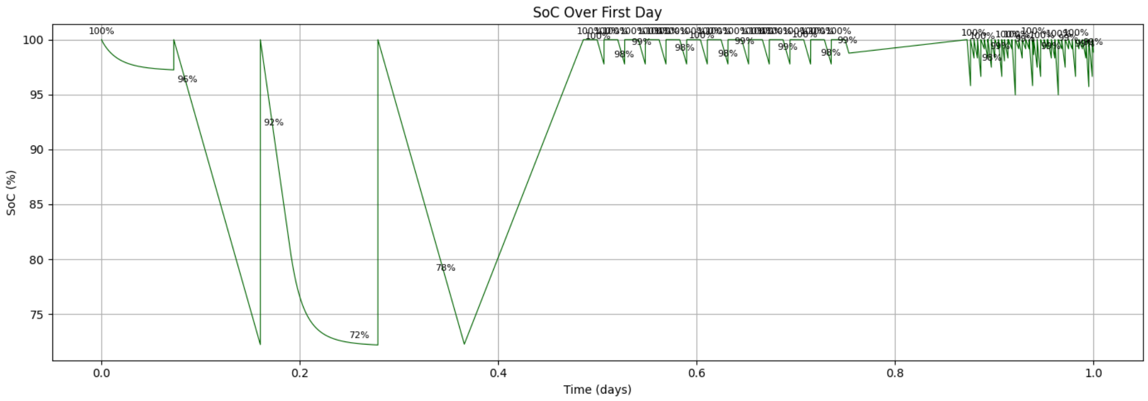


Figure 20: Zoomed-in preview: State of Charge (SoC) over the first day of cycling data.

In Figure 20, we plot the calculated SoC across the first day of the random walk cycling dataset. Even in this limited view, we observe fluctuations corresponding to charge and discharge events, showing potential for more detailed cycle-aware modeling.

5.3 Next Steps

This is only a starting point. In future work, we plan to integrate SoC with advanced statistical or machine learning models to explore its relationship with long-term degradation indicators such as Reference Capacity. The strong interpretability of SoC—combined with its alignment to physical battery behavior—makes it a promising feature for predictive applications.

6 Conclusions and Future Work

6.1 Analysis Conclusions

Concerning the voltage rebound analysis, as little as 30 seconds worth of voltage rebound data can be used in a predictive model to output the discharge reference capacity with low error. The average percent differences between the parameters obtained from fitting the logistic-like function on a subset of voltage rebound data versus the full data increase as smaller subsets worth of data are analyzed. Though, that does not lead to increased error in the predictive model results.

Overall, the voltage rebound shows promise for an easy and efficient way to predict discharge reference capacity using a logistic-like function as the estimation of the rebound curve.

6.2 Future Work

It was noticed that 30 seconds worth of voltage rebound data sometimes led to even lower error in the best model, random forest. This may be observed because after 30 seconds of voltage rebound data there is more noise. The subset analysis should be performed on the other two cell groups (5-18 and 53-56) from the NASA data to further investigate this.

References

[1] BatteryBits. Comparison of open datasets for lithium-ion battery testing. `https://medium.com/batterybits/comparison-of-open-datasets-for-lithium-ion-battery-testing-fd0de091ca2`, December 12, 2020. Medium post. Accessed: April 16, 2025.

[2] Birkl, C. R. Diagnosis and prognosis of degradation in lithium-ion batteries. *PhD thesis, Department of Engineering Science, University of Oxford* (2017).

[3] Birkl, C. R., and Howey, D. A. Oxford battery degradation dataset. `https://ora.ox.ac.uk/objects/uuid:03ba4b01-cfed-46d3-9b1a-7d4a7bdf6fac`, 2015.

[4] Bole, B., Kulkarni, C., and Daigle, M. Randomized battery usage data set. NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA, n.d. Dataset available at NASA Ames Research Center.

[5] CALCE. Battery-data: Cylindrical cells. `https://calce.umd.edu/battery-data`. n.d.

[6] Chen, T., and Guestrin, C. Xgboost documentation. `https://xgboost.readthedocs.io/`. Accessed: May 12, 2025.

[7] Devie, A., Baure, G., and Dubarry, M. Intrinsic variability in the degradation of a batch of commercial 18650 lithium-ion cells. *Energies 11*, 5 (2018), 1031.

[8] GeeksforGeeks contributors. Linear regression - geeksforgeeks, n.d. Accessed: 2025-05-01.

[9] Gun, D., Perez, H., and Moura, S. Fast charging tests. `https://datadryad.org/stash/dataset/doi:10.6078/D1MS3X#methods`. n.d.

[10] Heenan, T., Jnawali, A., and et al. Lithium-ion battery inr18650 mj1 data: 400 electrochemical cycles (eil-015). https://rdr.ucl.ac.uk/articles/dataset/Lithium-ion_Battery_INR18650_MJ1_Data_400_Electrochemical_Cycles_EIL-015_/12159462/1?file=23140433, 2020.

[11] Kumar, A. Ridge, lasso, and elastic net regression, 2020. Accessed: 2025-05-01.

[12] NASA. Nasa battery dataset, October 29 2022. Kaggle.

[13] Preger, Y., Barkholtz, H. M., Fresquez, A., Campbell, D. L., Juba, B. W., Romàn-Kustas, J., Ferreira, S. R., and Chalamala, B. Degradation of commercial lithium-ion cells as a function of chemistry and cycling conditions. *Journal of the Electrochemical Society 167*, 12 (2020), 120532.

[14] Ramsay, J. O., and Silverman, B. W. *Functional Data Analysis*. Springer, 2005.

[15] Reis, G. D., Strange, C., Yadav, M., and Li, S. Lithium-ion battery data and where to find it. *Science Direct* (2021). n.d.

[16] Research, S. Secondary battery market size, top share, report to 2032. https://straitsresearch.com/report/secondary-battery-market#:~:text=Market%20Overview,period%20(2024%E2%80%932032). n.d.

[17] ROHM. Electronics basics: Charging methods. https://www.rohm.com/electronics-basics/battery-charge/charging-method. n.d.

[18] scikit-learn developers. Decision trees – scikit-learn documentation. https://scikit-learn.org/stable/modules/tree.html. Accessed: May 12, 2025.

[19] scikit-learn developers. Ensemble methods (random forests) – scikit-learn documentation. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees. Accessed: May 12, 2025.

[20] Severson, K., Attia, P. M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M. H., Aykol, M., Herring, P. K., Fraggedakis, D., Bazant, M. Z., Harris, S. J., Chueh, W. C., and Braatz, R. D. Data-driven prediction of battery cycle life before capacity degredation. https://data.matr.io/1/projects/5c48dd2bc625d700019f3204, 2019.