Feature Extraction and Lithium Battery Cycling Curve Prediction via Machine Learning

Laisuo Su[1], Shuyan Zhang[2], B. Reeja-Jayan[2], Alan J. H. McGaughey[2], Arumugam Manthiram[1]

[1]Materials Science and Engineering Program & Texas Materials Institute, The University of Texas at Austin, Austin, TX 78712, United States.

[2]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States.

*e-mail: manth@austin.utexas.edu

Abstract

Real-time onboard state monitoring and estimation of the battery over its lifetime is indispensable for the safe and durable operation of battery-powered devices. In this study, we develop a methodology to predict the entire constant-current cycling curve with limited input information that can be collected in a short period of time. A total of 10,066 charge curves of $LiNiO_2$-based batteries at a constant C-rate are collected. With the combination of a feature extraction step and a multiple linear regression step, the method can accurately predict an entire battery charge curve with an error of < 2% using only 10% of the charge curve as the input information. The method is further validated across other battery chemistries ($LiCoO_2$-based) using open-access datasets (4,522 charge curves). The prediction error of the charge curves for the $LiCoO_2$-based battery is around 2% with only 5% of the charge curve as the input information, indicating the generalization of the developed methodology for predicting battery cycling curves. The developed method paves the way for fast onboard health status monitoring and estimation for batteries during practical applications.

Introduction

Lithium-ion batteries (LIBs) are becoming the dominant rechargeable batteries and are widely used in portable electronic devices, electric bikes, and electric vehicles (EVs) [1, 2]. Hundreds or even thousands of LIBs are connected to provide sufficient energy for EVs. For example, the Standard-range version of the Tesla Model 3 carries 2,976 LIBs arranged in 96 groups of 31 cells and the Long-range version contains 4,416 LIBs arranged in 96 groups of 46 cells[3]. The failure of one battery could propagate quickly through the entire battery pack, which triggers the malfunction of the battery system and may lead to safety issues like smoke, fire, and explosion[4]. Therefore, the states, such as state of charge (SOC) and remaining energy, and statuses, such as health condition, of batteries need to be accurately monitored to ensure their reliable and safe use.

A battery management system (BMS) is generally adopted to monitor the state of batteries, record battery usage information, analyze the status of batteries, and provide feedback and suggestion to customers[5]. The BMS can directly measure some key information with sensors, such as voltage, current, and temperature[6]. The combination of this information can further estimate the state of each battery, including SOC, remaining energy, and health conditions. Accurately estimating the health conditions of LIBs is very important but challenging to guide the use of batteries and at the same time prevent accidents and malfunctions[7]. The health condition of a battery is generally reflected by the decreased maximum capacity, the growth of internal resistance, and the appearance of fatal aging mechanisms such as the formation of lithium dendrite[8]. The assessment of these parameters is not trivial as BMSs typically only sample charging/discharging

current and voltage of batteries at a SOC range that is defined by customers' usage habits[9].

Many efforts have been made to estimate the health state of batteries in real applications. One common method is based on models, such as equivalent circuit models[10] and mechanism-based models[11], to simulate the behaviors of batteries, followed by various optimization algorithms and observations to identify the parameters in the models and the health states[12]. The estimation capability of battery health states relies on the accuracy of the models and the optimization algorithms. Therefore, building a representative model is crucial. Recently, data-driven methods are gaining increasing attention for battery health estimation and prediction due to their flexibility[13, 14]. As LIBs are nonlinear systems with complex degradation mechanisms that have not been fully understood, the nonlinear matching ability of data-driven methods makes them one of the most prominent approaches to estimating and predicting the health status for real applications[7]. The data-driven methods have been demonstrated to predict the state of health and remaining useful life of LIBs using impedance spectrscopy[15], to predict the cycle life using information from early cycles[14], and to predict the complete charge curve based on a part of the charging information[9]. In almost all these studies, the data-driven methods are treated as a "black box", which provides little help to deepen our understanding of the behavior of LIBs during cycling.

In this study, we combined unsupervised learning methods and supervised learning methods to predict the statuses of LIBs. Compared to human experts, unsupervised learning algorithms capture hidden features that can better represent the degradation of batteries. The physical meanings of the hidden features are discussed to help understand

the battery aging mechanisms. These hidden features are then used to predict the complect cycling curve, given a limit section on the curve. Finally, we expanded the developed methodology to predict open-source battery data with different chemistries.

Methods

Data generation

We collected the battery cycling curves in CR-2032 type coin cells with $LiNiO_2$ as the cathode and Li metal as the anode. The coin cells were tested at a C/10 rate three times after assembling, followed by a cycling test at room temperature with a C/2 charge rate and 1C discharge rate. The C/2 charge curves of these cells were collected for this study, and a total of 10,066 charge curves were selected with a minimum charge capacity of 160 mA h g^{-1}.

Open-source battery cycling data was also used to evaluate the developed methodology. A total number of 4,522 charge curves were taken from the Center for Advanced Life Cycling Engineering (CALCE) dataset (CS2_3, CS2_8, CS2_9, CS2_21, CS2_33~CS2_38) provided by A. James Clark School of Engineering at the University of Maryland[16]. The CALCE dataset was obtained from batteries with $LiCoO_2$ as the cathode material with trace elements of manganese, which is different from the $LiNiO_2$ cathode tested in our lab. The different chemistries of the two types of batteries lead to different shapes of the charge curves.

Feature extraction

Charge curves were selected for this study because the charging protocols are more controllable than discharge protocols to provide more consistent input in real-world applications. Three unsupervised learning algorithms were applied to extract hidden

features from the charge curves, which are principal component analysis (PCA), non-negative matrix factorization (NMF), and Autoencoder (AE). PCA and NMF are techniques that can decompose a matrix **Q** into two separated matrices **W** and **H**, such that it can be written as equation (1).

$$Q_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{p} W_{ia} H_{a\mu} \quad (1)$$

where Q is an $n \times m$ matrix that contains all the raw data information. **W** and **H** have dimensions of $n \times p$ and $p \times m$. The hyperparameter p is the number of features. The p columns of **W** can be interpreted as the hidden features of charge curves, which will be used to predict the complete charge curves by a supervised learning model. p is chosen based on the prediction accuracy of the validation set for all feature extraction algorithms. Each column of **H** contains the weights in a one-to-one correspondence with a basis hidden feature in **W**[17].

To obtain the elements of **W** and **H**, an optimization problem with the objective function $\|\mathbf{Q} - \mathbf{WH}\|_F$ is solved, where $\|\cdot\|_F$ is the Frobenius norm. The difference between PCA and NMF lies in the constraints on the optimization. In PCA, the columns of **W** are orthonormal, and the rows of **H** are orthogonal such that a unique solution is guaranteed[18]. In NMF, the elements of **Q**, **W,** and **H** are constrained to be non-negative. There is no unique solution because the problem is non-convex[19]. As such, we employed an initialization scheme called non-negative double singular value decomposition, which rapidly reduces the approximation error to a value that is lower than that using a random initialization[20]. PCA and NMF are performed using the Scikit-Learn package[21].

Autoencoder is an unsupervised learning method that adopts neural network architectures for the task of feature learning[22]. The neural network is constructed with a

bottleneck layer that enforces a compressed representation of the input layer, which has the same dimension as the output layer. The autoencoder with a single hidden layer implemented in this work is shown in Fig. S1 in the Supplemental Information. This autoencoder is similar to NMF in that the hidden layer contains the weight matrix (**H**) corresponding to the charge curves in the decoder weight matrix (**W**)[23], and the product of the two matrices approximates the input charge curve matrix. The autoencoder differs from NMF that there is no non-negativity constraint on the decoder weight matrix. It can also be easily extended by adding more fully connected layers or other layer types such as recurrent neural networks and convolutional neural networks[24]. The autoencoders are implemented using the Pytorch package[25].

Charge curve prediction

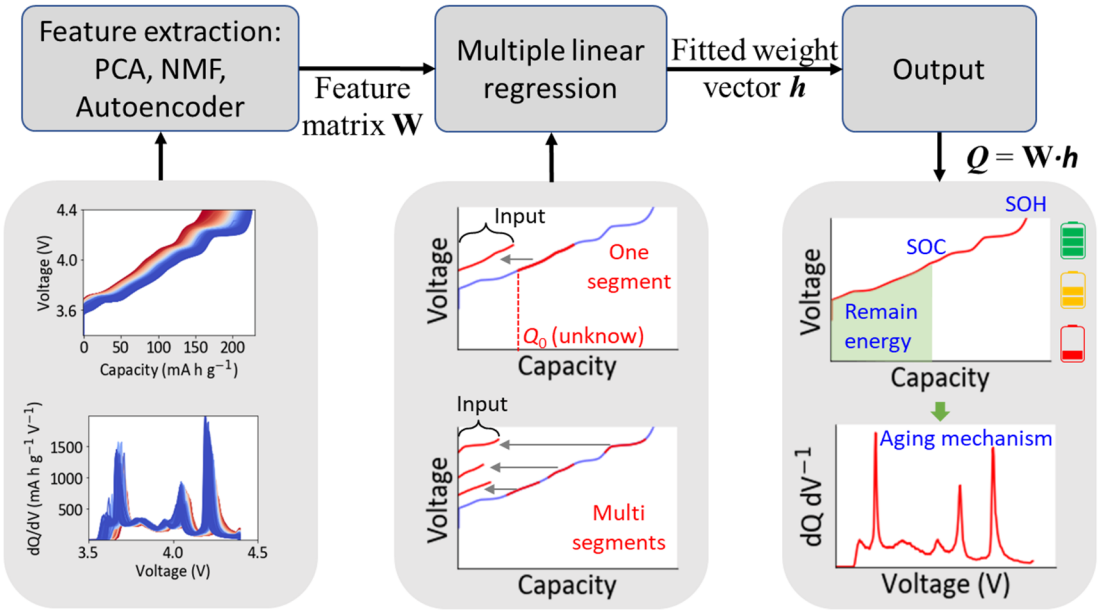
The 10,066 charge curves of the $LiNiO_2$ cells were randomly divided into a training set and a testing set with a ratio of 8: 2. Hidden features were extracted from the training dataset using PCA, NMF, and AE, from which we obtained the matrix **W** that contains p columns. Each column in matrix **W** represents a feature extracted from the training set.

Schematic 1 shows the workflow of this study. Given a partial charge curve with arbitrary starting voltage that corresponds to a starting capacity Q_0 and voltage window length that corresponds to a capacity range of $Q_{partial}$, we assume it can be approximated by the linear combinations of the same features in **W**. Thus, for the i^{th} point on the partial charge curve, we can write the following equation (2).

$$Q_0 + \Delta Q_i = \sum_{j=1}^{p} \mathbf{W}_j(V_i) \cdot h_j + \varepsilon_i \quad (2)$$

where Q_0 is the unknown starting capacity and is equivalent to the intercept of the linear regression, ΔQ_i is the incremental capacity relative to the starting capacity

Q_0, obtainable from experimental measurements. $\mathbf{W}_j$ is the j^{th} column of $\mathbf{W}$, h_j is the unknown weight corresponding to the features in $\mathbf{W}$, and ε is the error that follows a normal distribution.



Schematic 1. The workflow for predicting an entire charge curve of a battery based on a portion of the charge curve. Both a continuous segment and multiple separated segments can be used as the input. The output charge curve can derive many key states (SOC, SOH, and remaining energy) and even the aging mechanism of the battery.

The relationship between the incremental capacity ΔQ_i and the feature at a specific voltage $\mathbf{W}_j(V_i)$ can thus be modeled as a multiple linear regression problem. We use the L_1-norm as the regularization term to penalize the parameters and reduce overfitting. This regularizer (also called Lasso regularizer) can lead to some parameters being zero, i.e., removing the parameters for output evaluation[26]. Thus, the Lasso regularizer can also serve as a feature selection method. Given a single-segment partial charge curve with n points, the cost function can be defined as equation (3).

$$\mathcal{L}(\boldsymbol{h}, Q_0) = \sum_{i=0}^{n-1}\left(Q_0 + \Delta Q_i - \sum_{j=1}^{p} \mathbf{W}_j(V_i)\cdot h_j\right)^2 + \lambda \sum_{j=1}^{p}|h_j| \qquad (3)$$

where $\lambda \geq 0$ is the regularization parameter that controls the trade-off between approximation error and regularization strength[26]. The hyperparameter λ is determined through the validation set.

Solving this multiple linear regression will lead us to the $\boldsymbol{h}$ and Q_0 that minimize the cost function. To determine the complete charge curve $Q_{complete}$, we just need to take the linear combination of the charge curves **W** obtained from feature extraction and the weight vector $\boldsymbol{h}$ corresponding to the partial charge curve,

$$Q_{complete} = \mathbf{W} \cdot \boldsymbol{h} \tag{4}$$

The accuracy of the prediction is quantified by the root-mean-squared error (RMSE) between the predicted complete charge curve $Q_{complete}$ and the ground true charge curve Q_{true}, as shown in equation (5).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(Q_{complete}(i) - Q_{true}(i)\right)^2} \tag{5}$$

In practical applications, regenerative braking has been widely adopted in electric or hybrid vehicles to restore the wasted energy from the process of slowing down a car and using it to recharge the batteries[27]. The process of regenerative braking results in many separated charge segments instead of a continuous charge curve during the charging process. These separated segments can also be used as the input in the model to predict the entire charge curve.

Given m input segments with n points in total, we have m starting capacity $Q_{0,k}$ $(k = 1, 2, \ldots, m)$. This makes the multiple linear regression problem challenging to solve. For each segment, the linear relation at each point is displayed in equation (6).

$$Q_{0,k} + \Delta Q_{i,k} = \sum_{j=1}^{p} \mathbf{W}_j\left(V_{i,k}\right) \cdot h_j + \varepsilon_{i,k} \tag{6}$$

All input segments share the same $\boldsymbol{h}$ but have different starting capacity $Q_{0,k}$. This implies that the problem becomes *m* linear regressions with the same hyperplanes and different intercepts. One way to solve this is to center the **W**'s and ΔQ's for different segments at the origin by subtracting their means, then combine the centered values into new variables $\mathbf{W}^*$ and ΔQ^*. The *m* linear regressions are converted to a single formula (7).

$$\Delta Q^* = \sum_{j=1}^{p} \mathbf{W}^*(V_{i,k}) \cdot h_j + \varepsilon_i \quad (7)$$

The common $\boldsymbol{h}$ for the multiple input segments can then be solved.

Another way around it is to take the derivative of the input segments with respect to voltage, which generates the incremental capacity (IC) curves or dQ dV^{-1} curves. The dQ dV^{-1} analysis removes the unknown $Q_{0,k}$ because the analysis is based on the relative change of capacity. For this approach to work, we need to create another dataset with dQ dV^{-1} curves and perform feature extraction to obtain the dQ dV^{-1} curve matrix $\mathbf{W}_{IC}$. The i^{th} point on the *n* total points of the input can then be written as equation (8).

$$\left(\frac{dQ}{dV}\right)_i = \sum_{j=1}^{p} \mathbf{W}_{IC,j}(V_i) \cdot h_{IC,j} + \varepsilon_{IC,i} \quad (8)$$

where $\mathbf{W}_{IC,j}$ is the *j*th column of $\mathbf{W}_{IC}$. The cost function can be defined as equation (9).

$$\mathcal{L}_{IC}(\boldsymbol{h}_{IC}) = \sum_{i=0}^{n-1} \left(\left(\frac{dQ}{dV}\right)_i - \sum_{j=1}^{p} \mathbf{W}_{IC,j}(V_i) \cdot h_{IC,j} \right)^2 + \lambda \sum_{j=1}^{p} |h_{IC,j}| \quad (9)$$

The complete dQ dV^{-1} curve is obtained by $\left(\frac{dQ}{dV}\right)_{compleate} = \mathbf{W}_{IC}\boldsymbol{h}_{IC}$, and the complete charge curve is recovered by integrating the dQ dV^{-1} curve with respect to the voltage.

Results and Discussion

Charge curve feature extraction and reconstruction

The cycling stability of a battery largely depends on electrolytes that connect the two electrodes by providing Li^+ transport channels. Our group recently developed a novel electrolyte, namely, localized saturated electrolyte (LSE) that can significantly reduce the capacity fading rate of $LiNiO_2$-based batteries[28, 29]. For example, Figure 1a shows the evolution of the cell charge curve in the first 200 cycles tested in two different electrolytes, i.e., a conventional carbonate electrolyte (LP57) and an LSE. Compared to the LP57 electrolyte, the LSE slows down the shift speed of the charge curve to a lower capacity and a higher overpotential region (upper left) during cycling, which indicates that the LSE provided much better protection to the $LiNiO_2$ cathode during extended cycling.

Figure 1b normalizes all the charge curves in the aspect of the capacity for the batteries tested in both electrolytes. For the battery tested in the LSE, almost all normalized charge curves overlap with each other. The overlap of these normalized charge curves suggests that the maximum charge capacity is a dominant factor that describes the degradation of the battery during cycling. By comparison, there is a mismatch among the normalized charged curves for the battery tested in the LP57 electrolyte. The mismatch indicates that, in addition to the maximum charge capacity, there are other dominant factors that lead to the degradation of the battery. The maximum charge capacity and other dominant factors can be recognized as the expert-extracted features in the charge curve, which have physical meanings. For example, the maximum charge capacity can be correlated to the amount of $LiNiO_2$ active material. These expert-extracted features can not only be used to understand the degradation mechanisms of batteries during cycling but also to reconstruct the actual charge curve.

To examine the accuracy of the expert-extracted features to reconstruct the actual charge curve, we collected a total of 10,066 $LiNiO_2$-based battery charge curves from 52 cells. These cells used the same cathode ($LiNiO_2$) and anode (Li) but different electrolytes. All the charge curves were obtained at a rate of C/2 at room temperature. Figure 1c displays all the charge curves, and the capacity distribution of these curves is shown in Fig. S2. Figure 1d shows the normalized charge curves and the averaged normalized curve, which was calculated by taking the average of the normalized capacity for all the normalized curves at different voltages. The averaged normalized curve is considered as the expert-extracted feature, which is further used to reconstruct the actual charge curves.

Figure 1e shows the reconstruction of three charge curves based on the expert-extracted feature (averaged normalized curve in Figure 1d). The three charge curves were selected to be the ones with the maximum capacity (curve 3), the minimum capacity (curve 1), and the medium capacity (curve 2). The reconstructed curve 2 matches well with the measured curve, while there are noticeable differences between the reconstructed and the measured data for curve 1 and curve 3. Moreover, Figure 1f summarizes the distribution of the reconstruction errors of all the charge curves calculated from the equation (10),

$$Error = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(Q_{measured}(i) - Q_{reconstructed}(i)\right)^2} \quad (10)$$

where $Q_{measured}(i)$ is the actual capacity value at the sampling point i, $Q_{reconstructed}(i)$ is the reconstructed capacity value at the sampling point i, and n is the total number of sampling points. The average error is 4.7 mA h g^{-1} with a standard deviation of 2.0 mA h g^{-1}.

The averaged normalized curve in Figure 1d represents the characteristic voltage profile of $LiNiO_2$ during de-lithiation, while the charge capacity is determined by the number of active materials. If the loss of active materials is the only degradation mechanism for the capacity fading of $LiNiO_2$ cells, all the charge curves can then be accurately reconstructed using the extracted feature (the averaged normalized curve) and the corresponding parameter (maximum charge capacity). However, there are many other degradation mechanisms, such as impedance growth and reaction heterogeneity[30]. The impedance growth leads to an increased voltage polarization, which shifts the charge curve upwards. Moreover, the impedance of a battery is a function of the state of charge (SOC)[31], complicating the reconstruction process of the charge curve. Similarly, a $LiNiO_2$ electrode is composed of many secondary particles with a diameter of around 12 μm, which is further composed of hundreds of primary particles with a length of around 100 nm[28]. The extraction of Li^+ from these primary particles and secondary particles is nonuniform. The heterogeneity of the Li^+ extraction also depends on the SOC and different aging status, making the reconstruction process of the actual charge curve even more complicated. Therefore, the features that represent other degradation mechanisms of $LiNiO_2$ cells need be considered and extracted to predict the health status of batteries.

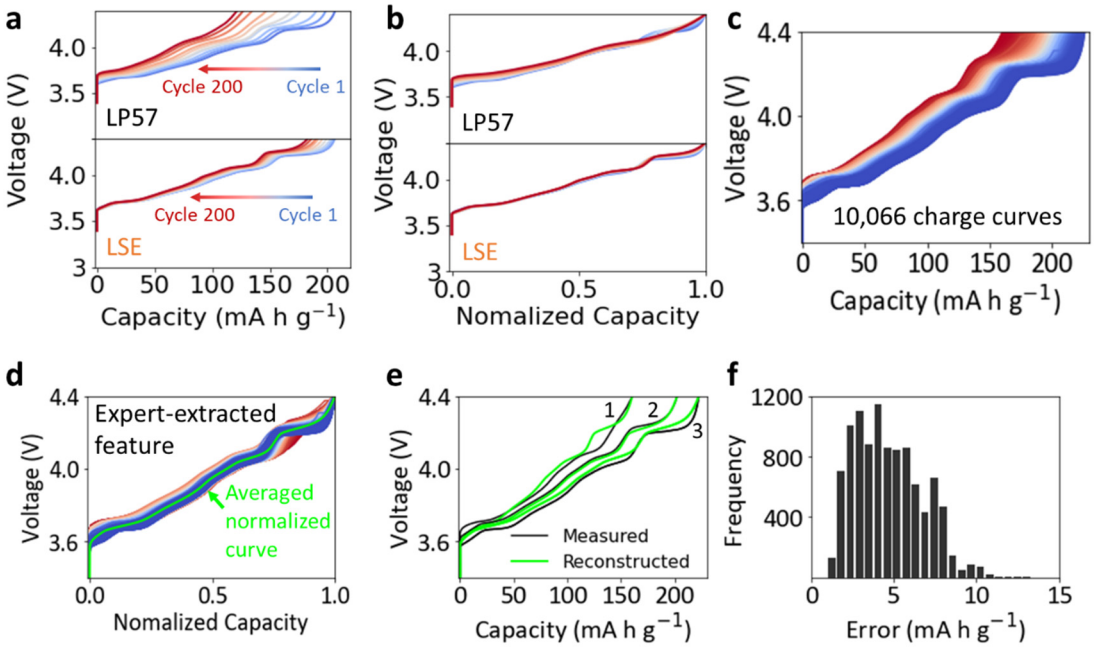


Figure 1 | Battery charge curves feature extraction and reconstruction. (a, b) The battery charge curves and normalized charge curves in the first 200 cycles tested with the LP57 electrolyte and the LSE. (c) The visualization of all the 10,066 charge curves used in this study. (d) The normalized charge curves and the averaged normalized charge curve of the total 10,066 profiles. The averaged normalized charge curve was calculated by taking the average of the normalized capacity for all the normalized curves at different voltages. (e) The comparison between three measured charge curves and the corresponding reconstructed charge curves. The reconstructed curves were obtained based on the averaged normalized curve shown in (d) and the maximum charge capacity. (f) The distribution of the reconstruction error of the 10,066 charge curves.

Unsupervised learning algorithms were applied to extract hidden features from the charge curves. Figure 2a shows the decomposition of all the 10,066 charge curves (Q) into two components (W, features) and the corresponding weights (W) using PCA. The mathematical principle of charge curve matrix decomposition can be found in the Method Section. Component 1 has a similar shape as the expert-extracted feature shown in Figure 1d. Thus, it represents the characteristic voltage profile of $LiNiO_2$ during de-lithiation and the corresponding weight 1 represents the number of active materials. Interestingly, component 2 shows a similar shape to dQ dV^{-1} analysis of the charge curve, where each phase transition corresponds to a peak in the dQ dV^{-1} curve[32]. It needs to be noted that the

Li^+ diffusion coefficient follows the phase transition of $LiNiO_2$, and each phase transition in the dQ dV^{-1} curve corresponds to a sharp decrease in the Li^+ diffusion coefficient. Thus, component 2 may represent the kinetic effect of $LiNiO_2$ during de-lithiation, and the corresponding weight 2 represents the kinetic contribution to the overall charge capacity. More hidden features can be extracted from the charge curve when increasing the number of components, but the physical meanings behind these features are hard to explain.

The accuracy of the reconstructed charge curve can be improved by increasing the number of components using PCA and NMF. For example, Figure 2b suggests that the charge curve can be accurately reconstructed when the number of components is increased to three ($p = 3$) in PCA, while there are noticeable differences between the measured charge curve and the reconstructed curves when the number of components is less than three ($p = 1, 2$). Thus, there are at least three different degradation mechanisms that significantly affect the shape of the charge curve. Similarly, Figure 2c suggests that three components ($p = 3$) are needed to well reconstruct the measured charge curve using NMF. It needs to be noted that the NMF algorithm does not converge when the number of components is one ($p = 1$), thus it is not included here. Figure 2d further shows the average reconstruction error of all the 10,066 charge curves with respect to the number of components using PCA and NMF, and the standard deviation of the errors is shown by the error bar. The reconstruction error decreases as the number of components increases. However, the physical meanings behind all these components (features) are hard to be explained when too many components are applied. Thus, the number of components should be chosen so that these components can accurately fit the charge curves at the same time provide explainable battery degradation mechanisms.

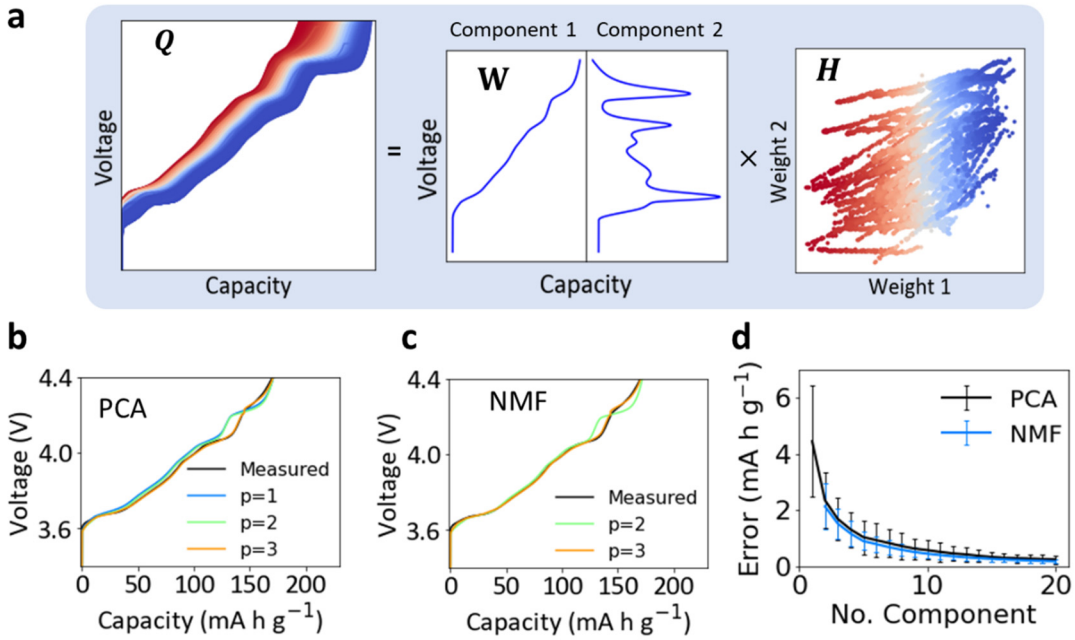


Figure 2 | Battery charge curve feature extraction and reconstruction using unsupervised learning algorithms. (a) Visual representation of charge curves matrix decomposition with the number of components in two. (b, c) The comparison between a measured charge curve and reconstructed charge curves based on (b) PCA and (c) NMF with different number of components (p). The charge curve was randomly picked just to demonstrate the effect of the number of components on the reconstruction accuracy. (d) The evolution of the reconstruction errors of all the 10,066 charge curves with the number of components in the PCA and NMF. The error bar represents the standard deviation of the errors.

Charge curve prediction based on a single input segment

The accurately reconstructed charge curves shown in Figure 2 indicate that the hidden features extracted by the unsupervised learning algorithms can well represent the degradation mechanisms of the tested battery. These hidden features are used by the multiple linear regression model to predict the entire charge curve, as introduced in the Methods section.

Figure 3 shows that the model based on PCA-extracted features can accurately predict the battery charge curve with limited input information. The averaged prediction errors of all the test data (2013 charge curves) are plotted in Fig. S3. The prediction

accuracy depends on the input data, which can be defined by the starting voltage (*viz.*, starting position) and the voltage window (*viz.*, length of the input data). The large prediction error at the bottom left corner in Fig. S3 is caused by the limited meaningful input information between 3.4 – 3.6 V, which feeds capacity value of zero into the model. In actual battery applications, the amount of input data depends on the charge time, which is directly related to the capacity rather than the voltage. No zero capacity values will be fed into the model, and the poor prediction performance at the bottom left region can be avoided. Fig. S4 further shows the evolution of the averaged prediction error with respect to the input length. The averaged prediction error reaches 6.5 mA h g^{-1} with only 20% of input length, corresponding to 3.0% of the relative error when normalized by the maximum specific capacity of 220 mA h g^{-1}.

Error! Reference source not found.a shows a selected input length from 40% to 100% of the charge curve to highlight the prediction accuracy. The inset shows the average prediction error with different input lengths, and the error bars represent the standard deviation of the errors across different starting positions. The result reveals that the model based on PCA-extracted features generally has a better prediction performance when the input sequence starts at a medium voltage, which may depend on the investigated cathode and anode materials that determine the shape of the charge curve. The maximum relative error is 1.8% (4.0 mA h g^{-1}) with 40% of input length, regardless of the starting position. The averaged relative error is only 1.3% (2.8 mA h g^{-1}) with 40% of input, and it further reduces to less than 1.0% (2.2 mA h g^{-1}) when the input goes beyond 50%. Moreover, the NMF-extracted features and AE-extracted features can also be used in the model for predicting the charge curves, and the performance of the model

is shown in Fig. S5 and Fig. S6. Both show high prediction accuracy, but slightly worse than the performance of the model based on the PCA-extracted features. Therefore, no further analysis was conducted on these two models.

To visualize the performance of the model based on PCA-extracted features, we show in **Error! Reference source not found.**b the prediction of the charge curves with the largest charge capacity (equivalent to the first cycle) and 80% of the largest charge capacity (equivalent to the last cycle), where 40% of the charge curve is chosen as the input. As the prediction accuracy depends on the starting position, the starting positions of the best and worst prediction are marked in **Error! Reference source not found.**a. **Error! Reference source not found.**b displays both the best and the worst prediction results of the charge curve in the first and the last cycle. The overlap between the predicted charge curves and both tested charge curves suggests the outstanding performance of the model.

Moreover, the corresponding $dQ\ dV^{-1}$ curves derived from the charge curves are shown in **Error! Reference source not found.**c,d, which are calculated from the equation (11). The $dQ\ dV^{-1}$ curve has been reported to be a versatile tool for diagnosing battery degradation mechanisms, such as loss of active materials, impedance increase, and lithium plating[33]. A good match among these $dQ\ dV^{-1}$ curves highlights the significance of the prediction method. Thus, a full charge curve at a constant current is no longer needed to evaluate the health status of batteries, which is time-consuming to collect and, in certain cases, unrealistic. Instead, a partial charge curve is sufficient to construct the full $dQ\ dV^{-1}$ curve for analyzing the degradation mechanisms of batteries.

$$(dQ\ dV^{-1})_k = \frac{Cap_{k+1} - Cap_k}{V_{k+1} - V_k}, (1 \leq k\ \leq N-1) \quad (11)$$

where k is the calculated point location and N is the total number of data points in a charge curve.

It is worth noting that dQ dV^{-1} curves are generally plotted at low rates (C/10 or below) to investigate the thermodynamical aspects of the battery. The accuracy of the dQ dV^{-1} analysis also depends on the quality of the measurement data[33]. For example, it is important to ensure environmental consistency during the test (temperatures and contacts), and the sampling rate should be reasonable to ensure enough data points for analysis and avoid large data files. Best practices for testing have been introduced in the literature[34]. In our study, the charging rate was C/2 for the batteries. The prediction accuracy of the dQ dV^{-1} curves is expected to increase with a slower charging rate, which warrants further investigation.

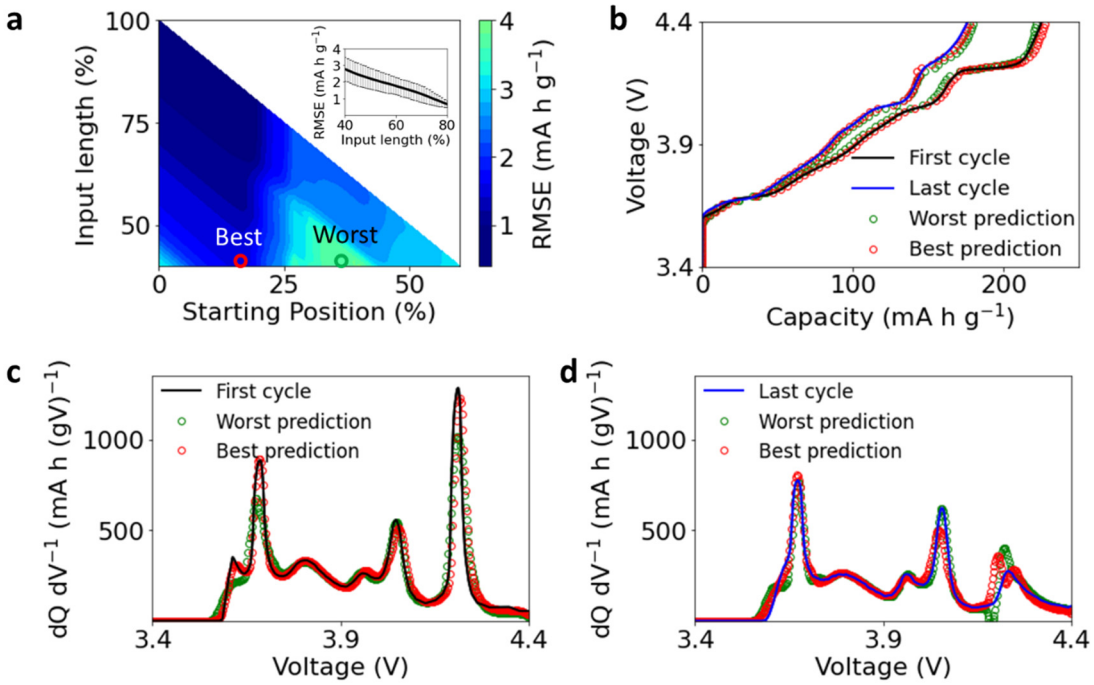


Figure 3 **Charge curve prediction based on PCA-extracted features.** (a) The average prediction error of the charge curves using the features captured by PCA. The x axis is the ratio between the voltage window of $(V_{start} - 3.4\,V)$ and the total voltage window $(4.4\,V - 3.4\,V)$, and the y axis is ratio between the voltage window of $(V_{end} - V_{start})$ and the total voltage window $(4.4\,V - 3.4\,V)$. The inset shows the average prediction error across different starting positions with respect to the input length. (b) The best and worst prediction results of the first charge curve and the last charge curve. The locations of the best and worst predictions are marked in (a). (c-d) The corresponding best and worst predictions of dQ dV^{-1} curves of the two charge curves shown in (b).

Charge curve prediction with multiple separated input segments

Regenerative braking has been widely adopted in electric or hybrid vehicles[27], during which many separated charge segments can be obtained. These separated segments can be used as the input in the developed model to predict the entire charge curve, providing an online tool to monitor the health status of the battery system. Moreover, the charge information at different times can also be combined as the input of the model to predict the entire charge curve, assuming little or no changes in the battery health status in the period.

Figure 4 shows the performance of the model using multiple separated input segments based on AE-extracted features. The segments were randomly selected on a charge curve, and the error is averaged throughout all the test data. The performance of the model based on PCA-extracted features and NMF-extracted features is shown in Fig. S7. As the model based on AE-extracted features shows the best prediction performance, the other two models are not further analyzed. However, the best feature extraction algorithms depend on the shape of the charge curve, which is determined by the materials of the two electrodes in a battery. Thus, other feature extraction algorithms may be applied for batteries with different types of electrodes to achieve optimal prediction performance.

Figure 4a suggests that the prediction error decreases with the increase in the number of segments and the total input length in the model. The model achieves high prediction accuracy when the number of segments is more than 15, even with a small input length. For example, the prediction error can be as low as 4.1 mA h g^{-1} (the relative error is 1.9%) with only 10% of input length when the number of segments reaches 20, which corresponds to around 12 mins of charge data collected at a C/2 rate. It should be mentioned that dQ dV^{-1} curves were predicted first when applying the multiple separated input segments, which were then used to calculate the charge curves by integrating the dQ dV^{-1} curves on voltage. We also examined the performance of the model by predicting the charge curves directly from the separated charge curve segments, but the prediction accuracy is not as good, as shown in Fig. S8.

Figure 4b,c compares the tested and predicted dQ dV^{-1} curves with the maximum capacity (Figure 4b, first cycle) and 80% of the maximum capacity (Figure 4c, last cycle).

Two types of inputs are selected for the model: a total input length of 20% with 10 segments and a total input length of 10% with 20 segments, which are marked in Figure 4a. Increasing the number of segments is more effective in improving the prediction accuracy than increasing the input length. For example, Figure 4a shows the prediction error of the model with 10% input and 20 segments (4.0 mA h g^{-1}) is smaller than that with 20% input and 10 segments (5.1 mA h g^{-1}). Figure 4b,c further shows that the peak positions and intensities of the dQ dV^{-1} curves are accurately predicted by the model that uses 20 segments and 10% of input length as the input. By comparison, there is a slight mismatch of peak intensities at 4.15 V (Figure 4b) and 3.65 V (Figure 4c) between the tested curves and the predicted ones with 10 segments and 20% of input length as the input. Such a mismatch can lead to over- or under-estimation of the total charge capacity, as shown in Figure 4d.

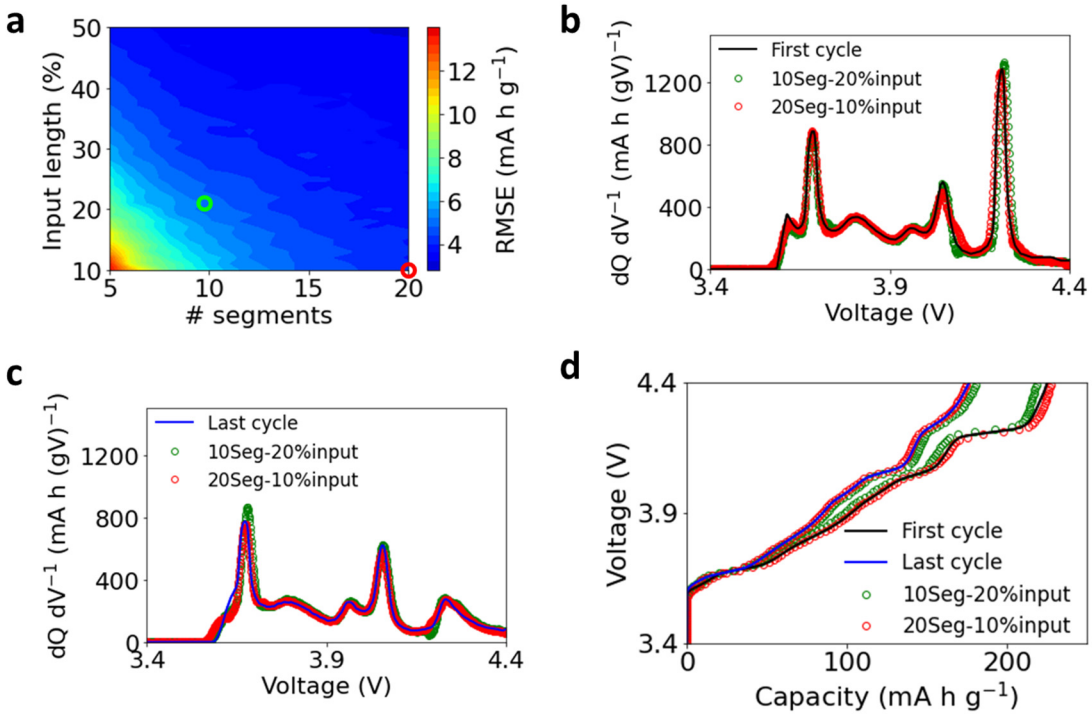


Figure 4 | Charge curve prediction based on multiple separated input segments. (a) The prediction error of the model based on AE-extracted features. The x axis is the

number of segments, and the y axis is the total length of the input sequence. (b, c) The comparison between measured and predicted dQ dV^{-1} curves of (b) the first cycle (with the maximum charge capacity) and (c) the last cycle (80% of the maximum charge capacity). Two different types of inputs are examined, which are marked in (a). (d) The corresponding charge curves calculated from the dQ dV^{-1} curves in (b) and (c).

Charge curve prediction for different batteries

To evaluate the applicability of the methodology developed in this work, we applied the workflow to open-source battery cycling data[16]. A total number of 4,522 charge curves were taken from the CALCE dataset. The CALCE dataset was tested from batteries with $LiCoO_2$ as the cathode material, which is different from the $LiNiO_2$ cathode, and thus, shows different shapes of the charge curves, as shown in Figure 5a. The accurate charge curve prediction of these batteries would illustrate the wide applicability of the developed methods.

The batteries from the CALCE dataset show a maximum charge capacity of ~ 1 Ah and a minimum capacity of 0.6 Ah. Figure 5a displays all the charge curves with the same charging rate (C/2) and the corresponding normalized charge curves. Figure 5b shows the charge capacity distribution of the 4,522 curves. The decrease in the charge capacity could be attributed to the loss of active materials during cycling. Moreover, the normalized charge curve shifts upwards, indicating the growth of the resistance that leads to a large overpotential during charging. There might be other degradation mechanisms, which can hardly be extracted by experts based on the simple analysis of the charge curves. Therefore, the three unsupervised learning algorithms (PCA, NMF, and AE) were applied to extract hidden features to be fed into the multiple linear regression model for predicting the health status of the battery.

Figure 5c and Fig. S9 show the performance of the model for predicting the overall charge curve based on features extracted from the three different algorithms. Overall, the model based on the PCA-extracted features outperforms the model based on the other algorithms-extracted features. Thus, the PCA-extracted features were used to predict the charge curves of the CALCE battery. Figure 5c suggests that the prediction error depends on the starting position and the length of the input data. A relatively large error appears in the bottom left corner that corresponds to 0 – 15% of the starting position, which is also shown in Fig. S10 with a full range of the input length from 1% to 100%. This large error is caused by the sharp voltage increase between 3.5 – 3.7 V (Figure 5a). As the input length was defined by the voltage range rather than the charge capacity, the starting position at around 3.5 V would lead to much less meaningful input information compared to that started at a higher voltage. To account for this drastic increase in the voltage region, we avoid this specific region when calculating the average prediction error with respect to the input length, as shown in the inset of Figure 5c. The average prediction error is less than 0.01 Ah when the input length goes beyond 50%, which corresponds to a relative error < 1.0% after being normalized by the maximum capacity of 1 Ah.

Figure 5d displays the prediction of the charge curves with the largest charge capacity (first cycle) and 80% of the largest charge capacity (last cycle). 50% of the charge curve was chosen as the input, and two different starting positions were selected as marked in Figure 5c to represent the best and worst performance of the model. The good agreement between the prediction curves and the tested curves indicates the outstanding performance of the model. Moreover, the corresponding dQ dV^{-1} curves

derived from the charge curves also match well between the prediction and the test data (Fig. S11), including the positions and intensities of all the peaks.

Figure 5e shows the performance of the model with multiple separated input segments based on the PCA-extracted features. The results show that the prediction accuracy increases with the increase in the number of segments and the input length. When the number of segments is more than 10, the prediction error is close to 0.02 Ah (relative error is 2.0%) with only 5% of the input length. Figure 5f displays the performance of the model to predict the charge curves in the first and the last cycle with 5% of the total input length and 10 separated segments. Fig. S12 further displays the corresponding dQ dV^{-1} curves derived from the charge curves. The almost perfect overlap between the predicted curves and the tested curves in both Figure 5f and Fig. S12 highlights the significance of the prediction method to diagnose the health status of batteries.

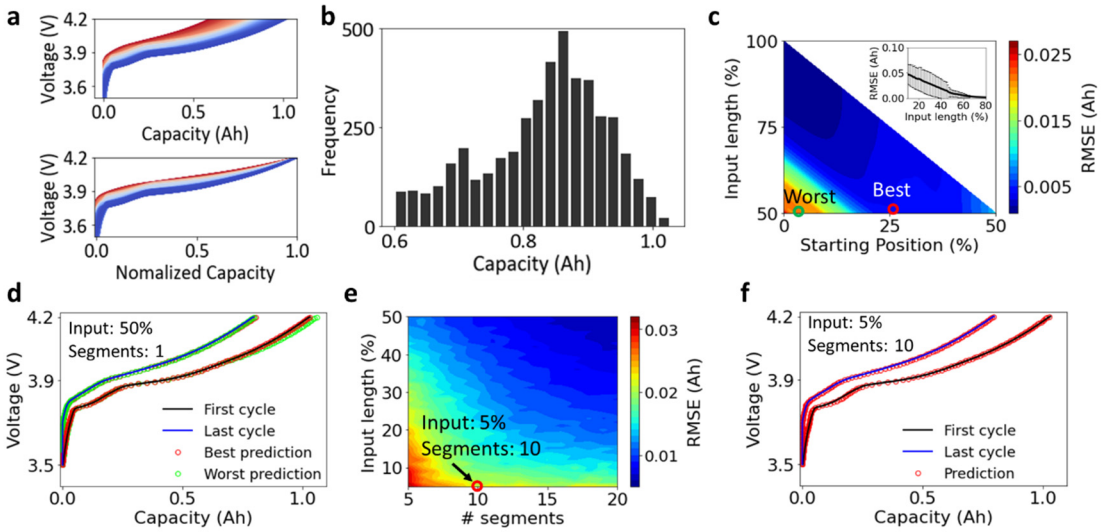


Figure 5 | Applying the methodology to other battery chemistries. (a) Charge curves and normalized charge curves of batteries taken from the Center for the CALCE dataset. A total number of 4,522 charge curves are taken in this study. (b) The distribution of the charge capacity of all the 4,522 charge curves. (c) The performance of the model based on PCA-extracted features and a single input segment. The inset shows the average

prediction error across different starting positions. (d) The best and worst prediction results of the first charge curve and last charge curve. The corresponding starting positions are marked in (c). (e) The performance of the model based on the PCA-extracted features and multiple separated input segments. (f) The prediction results of the first and last charge curves based on only 5% of the input length and 10 segments. The corresponding input condition is marked in (e).

Moving forwards to the real-world applications

The proposed methodology for predicting battery cycle life includes two steps: hidden feature extraction and multiple linear regression. Different from the literature that treats the data-driven method as a "black box"[7], the feature extraction step captures important information about the battery system that reflects the degradation mechanisms, such as loss of active materials, impedance growth, and increase of reaction heterogeneity. Moreover, the linear regression step provides the parameters that can be used to predict the health status of the battery, including remaining useful life, state of charge, and an entire charge curve. Therefore, the developed methodology has wide application in understanding the aging mechanisms, predicting health status, and providing advice to customers to optimizing the application of batteries in their devices. However, there are some gaps between this study and the real-world applications of the method in a BMS, which warrants further investigation.

Firstly, the charge curves in the study were collected at a constant C-rate (C/2) and at room temperature. But the current and temperature vary in real-world battery applications. Collecting and selecting the appropriate information to be used as the input will be an important step to improve the robustness of the model. An alternative solution is to develop a more robust model that can take all these information (current,

temperature, voltage, capacity, etc.) as the inputs. Neural networks could be a candidate for solving the problem.

Secondly, batteries with different types of chemistries (cathodes and anodes) have been widely used, which leads to different shapes of cycling curves. Although we examined two different batteries and demonstrated the applicability of the methodology in both cases, further evaluation of the model in other battery systems is needed. Moreover, the optimal feature extraction algorithms may differ from one battery system to another. More algorithms should be examined to obtain the model with the best prediction performance for a specific battery system.

Finally, correlating the algorithms-extracted features with the degradation mechanisms of a battery is an important step to deepen our understanding of the system. As a battery is a complex nonlinear system, the evolution of the electrodes (cathode and anode), electrolytes, and the interface between them could lead to the change of the capacity and resistance, which will be reflected in the cycling curve. Uncovering the battery degradation mechanisms and quantifying their effect on the shape of the charge curve could help build physics-informed models to reach an optimal prediction of battery performance.

Conclusion

Data-driven methods have a superior ability to capture hidden features in cycling curves. The hidden features can be correlated to the aging mechanism of batteries, such as loss of active materials, growth of resistance, an increase of reaction heterogeneity, etc. Moreover, these hidden features can be combined with a multiple linear regression model

to predict a complete cycling curve based on a limited portion of it. We demonstrate that both single continuous segment and multiple separated segments can be used as the input to predict the complete cycling curve. The model achieves a 2% prediction error of an entire charge curve using only 10% of the curve as the input for the $LiNiO_2$-based batteries and achieves the same accuracy with only 5% of the curve as the input for the $LiCoO_2$-based batteries. The complete charge curve can be used to evaluate the health status of batteries, which can not only guide the use of batteries but prevent accidents and malfunctions.

Data availability

The data that support the plots within this paper are available from the corresponding author upon reasonable request.

Conflicts of interest

The corresponding author (A. M.) is a co-founder of TexPower, Inc., a start-up company focusing on cobalt-free cathode materials for lithium-based batteries.

Acknowledgment

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Vehicle Technologies of the U.S. Department of Energy through the Advanced Battery Materials Research (BMR) Program (Battery500 Consortium) award number DE-EE0007762. (Prof. Jayan, Prof. McGaughey –Please add your funding sources here)

References

1 A. Manthiram, A reflection on lithium-ion battery cathode chemistry, *Nat. Commun.*, 2020, **11**.

2 H. Yaghoobnejad Asl and A. Manthiram, Toward sustainable batteries, *Nat. Sustain.*, 2021, **4**, 379-380.

3 T. Model and M. Reichweite, Tesla Model 3, *Access October 28, 2022*.

4 J. Zhang, L. Su, Z. Li, Y. Sun and N. Wu, The Evolution of Lithium-Ion Cell Thermal Safety with Aging Examined in a Battery Testing Calorimeter, *Batteries*, 2016, **2**, 12.

5 M. Nizam, H. Maghfiroh, R. A. Rosadi and K. D. Kusumaputri, Battery management system design (BMS) for lithium ion batteries,2020.

6 J. Huang, S. T. Boles and J. Tarascon, Sensing as the key to battery lifetime and sustainability, *Nat. Sustain.*, 2022, **5**, 194-204.

7 Y. Li, K. Liu, A. M. Foley, A. Zülke, M. Berecibar, E. Nanini-Maury, J. Van Mierlo and H. E. Hoster, Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review, *Renewable and Sustainable Energy Reviews*, 2019, **113**, 109254.

8 A. Farmann, W. Waag, A. Marongiu and D. U. Sauer, Critical review of on-board capacity estimation techniques for lithium-ion batteries in electric and hybrid electric vehicles, *J. Power Sources*, 2015, **281**, 114-130.

9 J. Tian, R. Xiong, W. Shen, J. Lu and X. Yang, Deep neural network battery charging curve prediction using 30 points collected in 10 min, *Joule*, 2021, **5**, 1521-1534.

10 Z. Liu, Z. Li, J. Zhang, L. Su and H. Ge, Accurate and Efficient Estimation of Lithium-Ion Battery State of Charge with Alternate Adaptive Extended Kalman Filter and Ampere-Hour Counting Methods, *Energies*, 2019, **12**, 757.

11 X. Yang, Y. Leng, G. Zhang, S. Ge and C. Wang, Modeling of lithium plating induced aging of lithium-ion batteries: Transition from linear to nonlinear aging, *J. Power Sources*, 2017, **360**, 28-40.

12 L. Lu, X. Han, J. Li, J. Hua and M. Ouyang, A review on the key issues for lithium-ion battery management in electric vehicles, *J. Power Sources*, 2013, **226**, 272-288.

13 L. Su, M. Wu, Z. Li and J. Zhang, Cycle life prediction of lithium-ion batteries based on data-driven methods, *eTransportation*, 2021, **10**, 100137.

14 K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh and R. D. Braatz, Data-driven prediction of battery cycle life before capacity degradation, *Nat. Energy*, 2019, **4**, 383-391.

15 Y. Zhang, Q. Tang, Y. Zhang, J. Wang, U. Stimming and A. A. Lee, Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning, *Nat. Commun.*, 2020, **11**.

16 Open Source Battery Research Data, https://calce.umd.edu/data#INR.

17 D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, 1999, **401**, 788-791.

18 J. C. Liao, R. Boscolo, Y. Yang, L. M. Tran, C. Sabatti and V. P. Roychowdhury, Network Component Analysis: Reconstruction of Regulatory Signals in Biological Systems, *Proceedings of the National Academy of Sciences - PNAS*, 2003, **100**, 15522-15527.

19 K. Huang, N. D. Sidiropoulos and A. Swami, Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition, *IEEE Trans. Signal Process.*, 2013, **62**, 211-224.

20 C. Boutsidis and E. Gallopoulos, SVD based initialization: A head start for nonnegative matrix factorization, *Pattern Recognit.*, 2008, **41**, 1350-1362.

21 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, Scikit-learn: Machine learning in Python, *the Journal of machine Learning research*, 2011, **12**, 2825-2830.

22 I. Goodfellow, Y. Bengio and A. Courville. *Deep learning*, MIT press, 2016.

23 P. Smaragdis and S. Venkataramani, A neural network alternative to non-negative audio models,2017.

24 P. Smaragdis and S. Venkataramani, A neural network alternative to non-negative audio models,2017.

25 A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, Automatic differentiation in pytorch, 2017.
26 R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, **58**, 267-288.
27 S. R. Cikanek and K. E. Bailey, Regenerative braking system for a hybrid electric vehicle,2002.
28 L. Su, E. Jo and A. Manthiram, Protection of Cobalt-Free LiNiO2 from Degradation with Localized Saturated Electrolytes in Lithium-Metal Batteries, *ACS Energy Lett.*, 2022, 2165-2172.
29 L. Su, X. Zhao, M. Yi, H. Charalambous, H. Celio, Y. Liu and A. Manthiram, Uncovering the Solvation Structure of LiPF6 - Based Localized Saturated Electrolytes and Their Effect on LiNiO2 - Based Lithium - Metal Batteries, *Adv. Energy Mater.*, 2022, 2201911.
30 L. Su, J. L. Weaver, M. Groenenboom, N. Nakamura, E. Rus, P. Anand, S. K. Jha, J. S. Okasinski, J. A. Dura and B. Reeja-Jayan, Tailoring Electrode – Electrolyte Interfaces in Lithium-Ion Batteries Using Molecularly Engineered Functional Polymers, *ACS Appl. Mater. Interfaces*, 2021, **13**, 9919-9931.
31 L. Su, J. Zhang, J. Huang, H. Ge, Z. Li, F. Xie and B. Y. Liaw, Path dependence of lithium ion cells aging under storage conditions, *J. Power Sources*, 2016, **315**, 35-46.
32 L. de Biasi, A. Schiele, M. Roca Ayats, G. Garcia, T. Brezesinski, P. Hartmann and J. Janek, Phase Transformation Behavior and Stability of LiNiO2 Cathode Material for Li - Ion Batteries Obtained from In Situ Gas Analysis and Operando X - Ray Diffraction, *ChemSusChem*, 2019, **12**, 2240-2250.
33 M. Dubarry and D. Anseán, Best practices for incremental capacity analysis, *Front. Energy Res.*, 2022, **10**.
34 M. Dubarry and G. Baure, Perspective on Commercial Li-ion Battery Testing, Best Practices for Simple and Effective Protocols, *Electronics*, 2020, **9**, 152.