# Electronic Supplementary Information (ESI)

# Data-driven insights into reaction mechanism of Li-rich cathodes

Jieun Kim[a,†], Injun Choi[b,†], Ju Seong Kim[a], Hyokkee Hwang[b], Byongyong Yu[a], Sang-Cheol Nam[a], and Inchul Park[a,*]

[a]LiB Materials R&D Laboratory, POSCO N.EX.T Hub, 100 Songdogwahak-ro, Yeonsu-gu, Incheon, 21985, Republic of Korea.

[b] Convergence Innovation Research Laboratory, Research Institute of Industrial Science & Technology, 67 Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, 37673, Republic of Korea

Email Address: inchul@posco-inc.com

*Corresponding author at: inchul@posco-inc.com

[†]These authors contributed equally to this work.

**Table S1.** Description of the collected experimental dataset

| Items | | Value & conditions |
|---|---|---|
| Dataset | # of cells | 390 |
| | # of cycles | 30,467 |
| Electrode | Cathode material | 92.5 wt% |
| | Conductor | 3.5 wt% |
| | Binder | 4.0 wt% |
| | Electrode density | Min: 2.53, Max: 3.18, Mean: 2.85 g/cm$^3$ |
| | Drying condition | 130 ℃, 20 min |
| Coin Half Cell | Size | 2032 |
| | Al-foil | 20 um thickness |
| | Anode | Li-metal |
| | Electrolyte | 1M $LiPF_6$ in (EC+EMC = 3 : 7 vol. ratio) |
| | Separator | Polypropylene |
| Test Sequence cc-mode | Standard capacity | 1C, 200mAh/g |
| | Temperature | 25 ℃ |
| | Formation & initial cycles | 0.1C/0.1C~2C (4.7/4.5-2.5V) |
| | Cycle test | 0.5C/0.5C (4.5-2.5V) |

# Supplementary Section 1. Detailed Preprocessing Criteria and Rationale (Figure S1 - 2)

This section provides a comprehensive overview of the filtering thresholds and interpolation choices discussed in the main text (Section 2.1 and Experimental Section). The primary goal is to ensure that only consistent and representative charge curves are retained for PCA, thereby minimizing noise and spurious signals arising from early activation, dendritic failures, or abnormal cycling behavior.

Although lower cutoffs (e.g., 2.0 - 2.5 V) and higher cutoffs (e.g., 4.8 - 4.9 V) can provide additional insights into $Mn^{3+/4+}$ and oxygen redox mechanisms, we limited our upper voltage to 4.5 V due to industrial considerations that aim to reduce electrolyte oxidation above 4.5 V. In addition, we emphasize charge curves rather than discharge curves because the charging process is more controllable, enabling a consistent representation of the electrochemical behavior and state of health.

## S1.1. Activation Considerations

1. **Exclusion of Cycles Before the 10th-Cycles**

• **Description**: We disregard the first nine cycles for each cell, beginning our analysis from the 10th cycle onward.

• **Rationale**: Many Li-rich layered oxide (LRLO) materials undergo “activation” during the initial few cycles, where capacity and average voltage can fluctuate significantly. By focusing

on cycles after this phase, we capture more stable behavior reflective of the material's steady-state operation.

2. **Minimum 10th-Cycle Capacity (> 160 mAh $g^{-1}$)**

• **Description**: If a cell's 10th-cycle capacity falls below 160 mAh $g^{-1}$, we exclude that cell entirely.

• **Rationale**: Persistently low capacity at this point suggests that either the cell is defective, or the material is failing to activate properly. We remove such cells to avoid confounding lower-limit outliers in the PCA analysis.

## S1.2. Capacity-Based Outlier Removal

1. **Capacity range (150–250 mAh $g^{-1}$)**

• **Description**: For any cycle, if the capacity is < 150 mAh $g^{-1}$ or > 250 mAh $g^{-1}$, we discard that specific cycle from further consideration.

• **Rationale**:

Empirical observations from operating LRLO-based cathodes in the 3.0–4.5 V range suggest that their capacities typically lie between roughly 150 and 250 mAh $g^{-1}$ after activation. Values far outside this range are likely attributable to measurement errors, sudden cell failures, or extraneous test conditions (*e.g.*, accidental short circuits).

**2. Minor Recovery Artifacts (Capacity > 1.025× Previous Cycle)**

• **Description**: If a cycle's capacity exceeds 1.025 times that of the immediately preceding cycle, we exclude the anomalous cycle only.

• **Rationale**: Slight capacity "jumps" can occur when different testing protocols introduce partial rest steps or small changes in temperature. These rest steps are often employed as a precautionary measure every 25 cycles; while they might transiently reverse some degradation markers, we observe that such steps do not substantially alter long-term patterns **(Figure S1)**. Because our focus is on continuous degradation trends, these abrupt bumps may not represent genuine electrochemical behavior and are excluded.

**3. Major Capacity Fluctuations (Capacity > 1.25× or < 0.8× Previous Cycle)**

• **Description**: If a cycle's capacity is more than 25% higher or 20% lower than the previous cycle, no **subsequent** cycles from that cell are included.

• **Rationale**: Large, sudden jumps or drops often reflect severe cell malfunctions—such as dendrite growth, electrolyte depletion, or internal shorting. Once such an event occurs, the long-term cycling profile for that cell typically remains unreliable, and those data are excluded from any further analyses.

## S1.3. Interpolation and Mean Subtraction

• **Interpolation**:

We **linearly interpolated** the retained charge curves onto a uniform voltage grid from 3.0 to 4.5 V, dividing the range into 10,000 equally spaced intervals. This method is

computationally straightforward and aligns with the general smoothness of voltage-capacity profiles under standard low to moderate C-rates.

• **Zero-Mean Transformation**:

After interpolation, we compute the mean capacity across all training curves at each voltage point and subtract it to ensure zero-mean data for PCA. This procedure isolates the relative differences in capacity-voltage behavior between cells.

## S1.4. Justification and Practical Implications

• **Why These Thresholds?**

The 1.025, 1.25, and 0.8 multipliers are empirically derived through extensive internal testing of LRLO coin cells. They reflect practical cutoffs beyond which anomalies are typically irreversible or unrepresentative of normal cycling.

• **Effect on Final Dataset**:

While these filters remove a notable fraction of cycles that are “edge cases”, this results in a more reliable and consistent dataset, ultimately improving the interpretability of PCA results.

• **Flexibility**:

If future studies employ different cycling protocols or operate in wider capacity ranges, these threshold values may be adjusted. The methodology remains general: the goal is to remove spurious data while capturing the true material degradation trends.

By applying these steps and thresholds **(Figure S2)**, we ensure that the final set of charge curves more accurately reflects the genuine electrochemical behavior of LRLO materials during cycling, mitigating artificial perturbations and noise.

**Note**: To reproduce the data further, please refer to the provided Jupyter notebooks available on GitHub. The detailed calculation process performed in this study is available at https://github.com/LRLO-PCA/Data-driven-insights-into-reaction-mechanism-of-Li-rich-cathodes

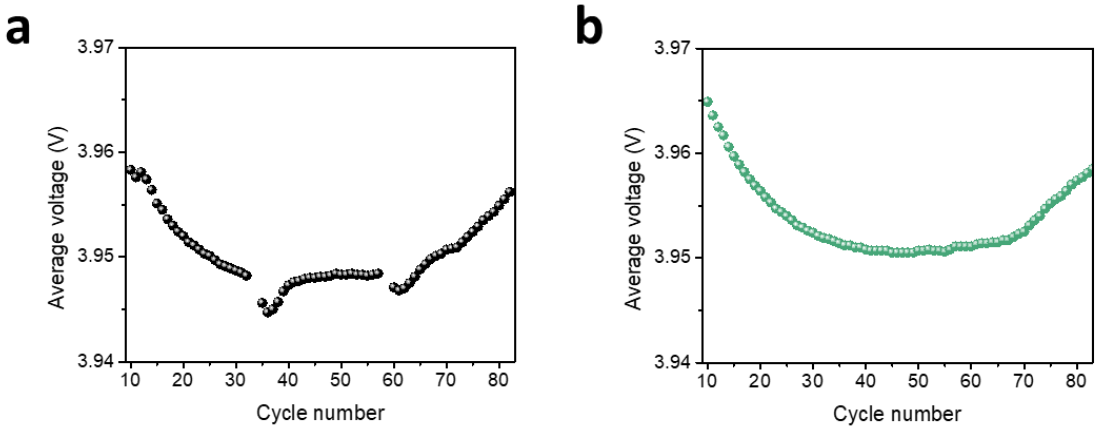


**Figure S1.** Average voltage **(a)** with recovery steps and **(b)** without recovery steps.

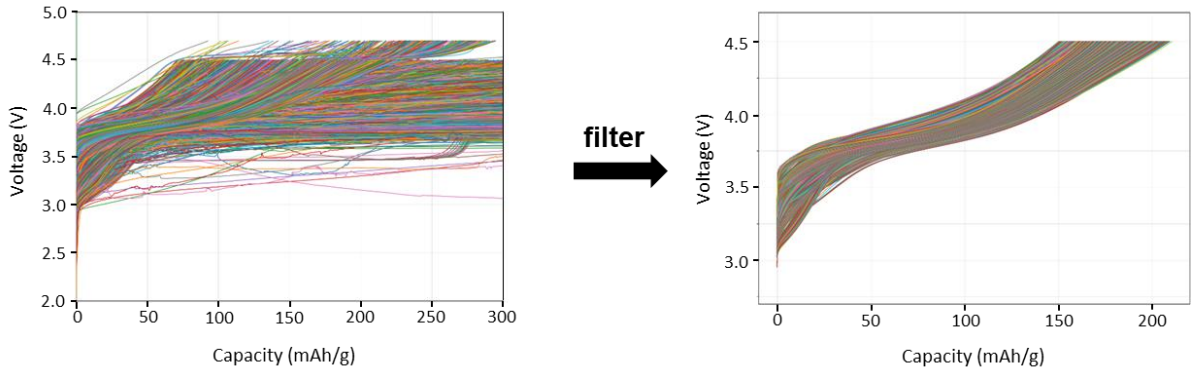


**Figure S2.** The charge curve dataset before and after filtering from LRLO used for PCA in this study

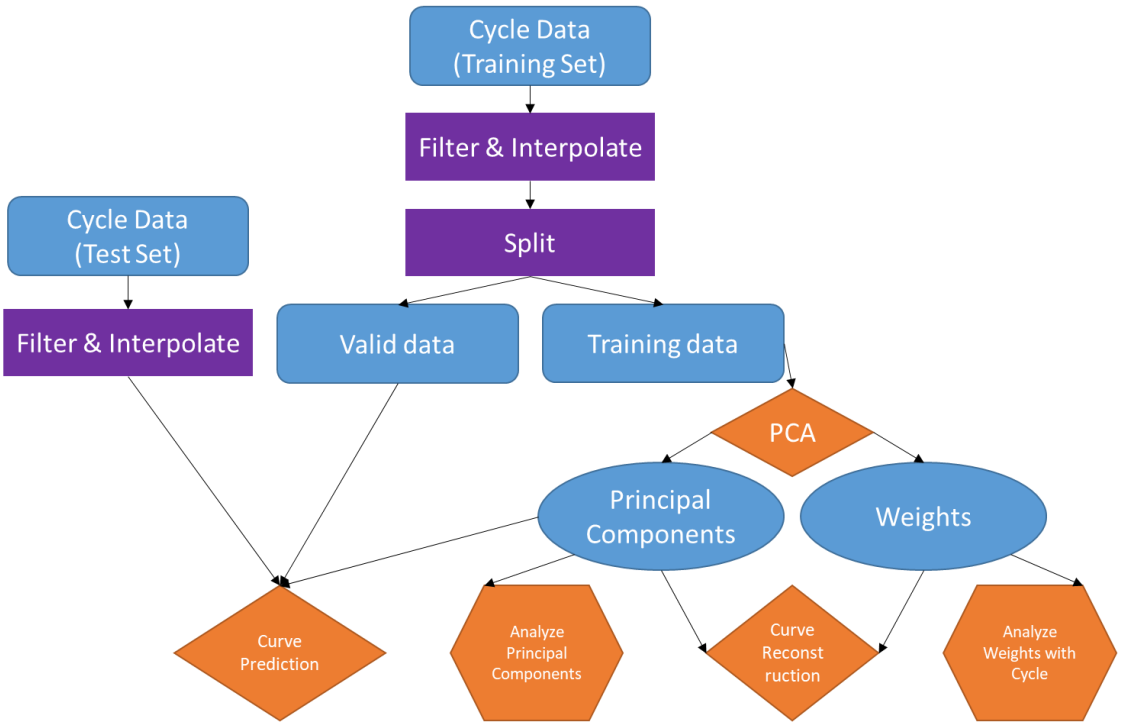


**Figure S3.** Flowchart of this research

## Supplementary Section 2. Evaluation of Alternative Dimensionality Reduction and Regression Methods (Figure S4, Table S2 - 4)

We employed NMF (Non-negative matrix factorization) as an alternative dimensionality reduction technique in addition to PCA. The results show that the component shapes are complex and noisy, which hinders the ability to derive insights into electrochemical processes (**Figure S4**). The reconstruction using NMF components also yield less accurate results compared to PCA (**Table S2**).

For the ablation study at the charge curve prediction phase, we assumed that some principal components we extract are not effective for predicting charge curves. To this, Lasso and Ridge regression were utilized instead of linear regression, revealing a higher mean of RMSE compared to linear regression, which employs all seven components (**Table S3**). This fact suggests that models with pushing component weights toward zero are not suitable in this system.

Also, to consider the latent non-linear characteristics, we tested non-linear models such as Support Vector Regression (SVR) with the Radial Basis Function (RBF) kernel and Light Gradient Boosting Machine (LightGBM) models which have non-linear discrimination power. As a result, both non-linear models undermine the prediction performance (**Table S4**), suggesting that the components derived from PCA, used as features in this study have a predominantly linear relationship with the charge curves.

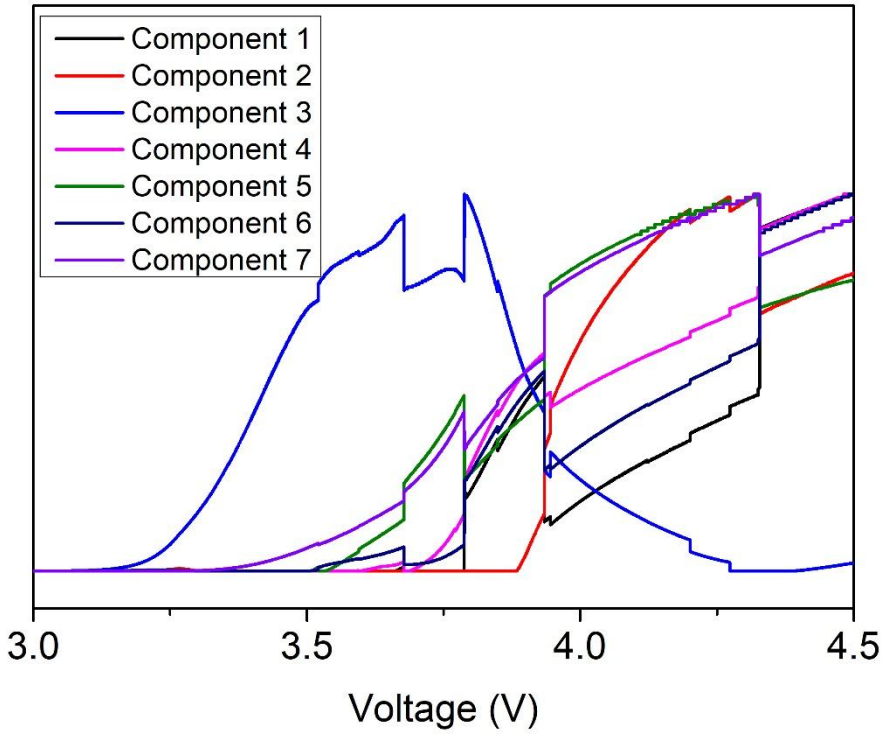


**Figure S4. (a)** The NMF components derived from the charge curves.

**Table S2.** Average RMSE of charge curve reconstruction/prediction from feature extraction models.

| Extraction Model | Reconstruction RMSE (mAh/g) | Prediction RMSE (mAh/g) |
|---|---|---|
| NMF | 1.7603 | 9.6122 |
| PCA | 0.1644 | 2.3895 |

**Table S3.** Accuracy of predicted curve via Lasso and Ridge regression.

| Regression Model | Number of components | α | Mean of RMSE (mAh/g) |
|---|---|---|---|
| Lasso | 7 | 0.01 | 4.08 |
| | | 0.1 | 6.16 |
| Ridge | | 0.01 | 7.17 |
| | | 0.1 | 6.28 |
| Lasso | 13 | 0.01 | 4.10 |
| | | 0.1 | 6.64 |
| Ridge | | 0.01 | 6.90 |
| | | 0.1 | 7.85 |

**Table S4**. Accuracy of predicted curve via Support Vector and LightGBM models.

| Regression Model | Average RMSE (mAh/g) |
|---|---|
| Linear | 2.3895 |
| Support Vector | 6.2410 |
| LightGBM | 6.8130 |

## Supplementary Section 3. Determining the Optimal Number of Principal Components for Charge Curve Reconstruction (Table S5, Figure S5 – 6)

The accuracy as a function of the number of PCA components is shown in **Figure S5** and **Table S5** for components ranging from n=1 to 10. With only one component, the average RMSE was 2.7 mAh/g, and the upper 95% confidence interval of the RMSE reached 5.2 mAh/g. This is an error of more than 2% for a typical capacity of 200 mAh/g. As the number of components increases, there is a consistent decrease in RMSE. Utilizing 4 components results in RMSE of 0.45 mAh/g. When the component count increases to 7, we observe an improvement in precision, as evidenced by the average RMSE declining to 0.16 mAh/g. Upon expanding the number of components to 10, the average RMSE further decreases to below 0.1 mAh/g. Moreover, in 95% of instances, the RMSE remains below 0.14 mAh/g. However, such high accuracy may indicate the possibility of overfitting, which could potentially reduce the prediction accuracy. Therefore, by measuring the prediction accuracy on a new test set, we confirmed that 7 principal components best describe the electrochemical characteristics of LRLO materials, and this number was chosen for use (see **Table S8**).

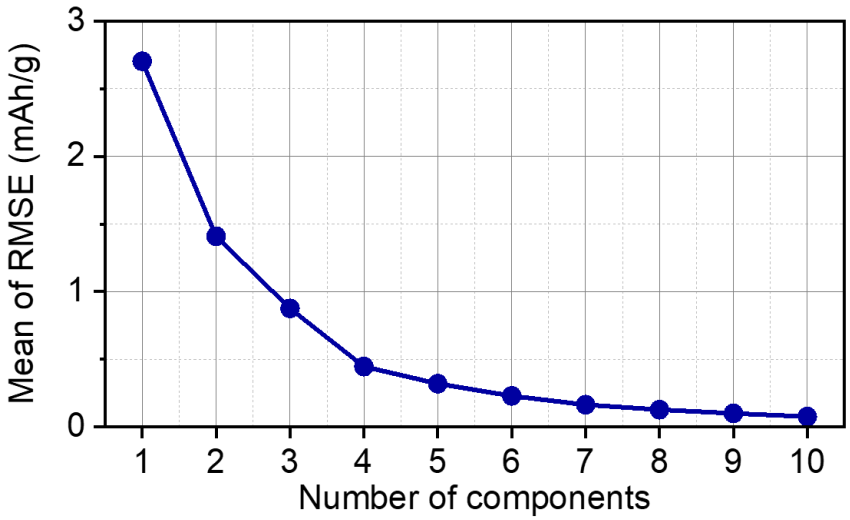


**Figure S5.** Accuracy plot of reconstructed curve by the number of PCA components

**Table S5.** Accuracy of reconstructed curve by the number of PCA components

| Number of components | RMSE | | | | | |
|---|---|---|---|---|---|---|
| | mean | median | 25% | 75% | 90% | 95% |
| 1 | 2.7017 | 2.4857 | 1.7290 | 3.4121 | 4.5449 | 5.1568 |
| 2 | 1.4088 | 1.2768 | 0.8706 | 1.8383 | 2.3914 | 2.6575 |
| 3 | 0.8757 | 0.7692 | 0.5280 | 1.0661 | 1.5176 | 1.9395 |
| 4 | 0.4469 | 0.4021 | 0.2720 | 0.5630 | 0.7556 | 0.8818 |
| 5 | 0.3201 | 0.2837 | 0.1893 | 0.3920 | 0.5223 | 0.6628 |
| 6 | 0.2301 | 0.1973 | 0.1425 | 0.2807 | 0.3874 | 0.4521 |
| 7 | 0.1644 | 0.1401 | 0.1069 | 0.1997 | 0.2680 | 0.3197 |
| 8 | 0.1278 | 0.1064 | 0.0844 | 0.1485 | 0.2158 | 0.2591 |
| 9 | 0.1011 | 0.0890 | 0.0694 | 0.1163 | 0.1584 | 0.1925 |
| 10 | 0.0772 | 0.0689 | 0.0537 | 0.0911 | 0.1177 | 0.1399 |

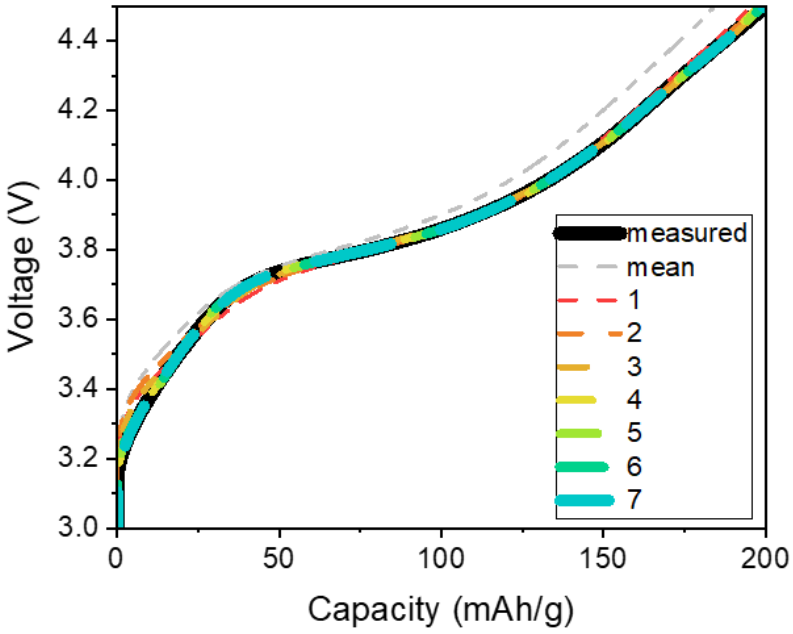


**Figure S6**. Charge curve reconstructed from PCA.

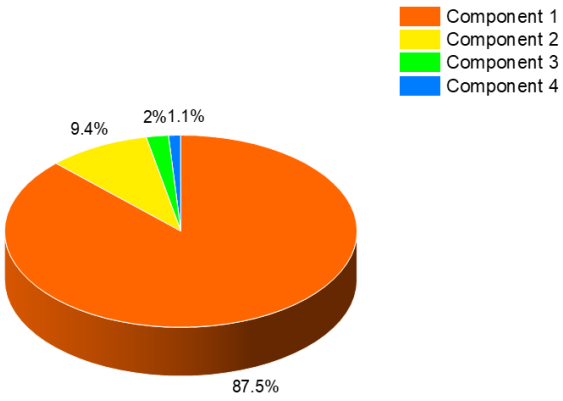


**Figure S7.** Interpretability of components

## Supplementary Section 4. Interpretation of higher-order principal components (PC3, 4) (Figure S8 - 9)

**Figure S8** plot the PC3 and PC4 weights against cycle number, respectively. These weights do not change appreciably in a uniform manner over cycles; instead, their values appear more strongly dependent on individual cell or material characteristics. In **Figure S9**, we plot PC1 weight on the x-axis against PC2, PC3, or PC4 weight on the y-axis (each figure separately). Unlike PC2, neither PC3 nor PC4 shows a clear degradation pattern that holds across the dataset, indicating that these components may reflect composition-driven variations that are less universally tied to LRLO degradation. However, these minor components may be unique to the current dataset, so they should not be overinterpreted.

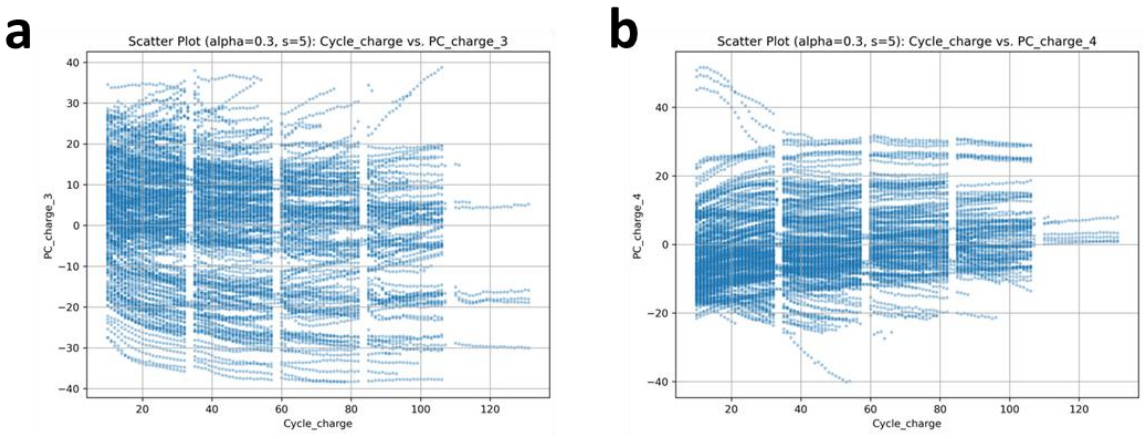


**Figure S8. (a)** PC3 weight and **(b)** PC4 weight vs. cycle number.

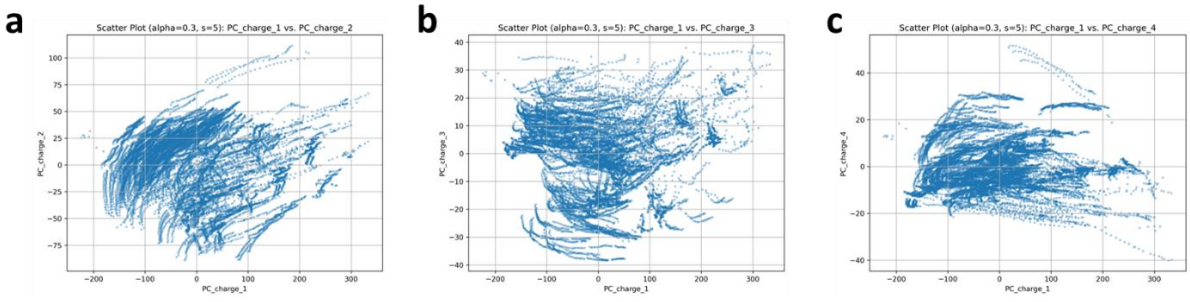


**Figure S9.** Scatter plot of PC1 weight (x-axis) vs. (a) PC2 weight, (b) PC3 weight and (c) PC4 weight (y-axis).

## Supplementary Section 5. Correlation Analysis of PCA Components with Capacity and Average Voltage across Cycles (Figure S10 - 13)

The Pearson correlation coefficients for all data were measured from the PCA coefficients, capacity, and average voltage across all cycles in the training dataset. To understand how the correlation coefficient changes with cell aging, the correlation coefficients of the PCA weights, capacity, and average voltage were measured specifically for data from certain cycle numbers within the training dataset.

**Figure S11** shows the correlation coefficient matrices for all data and specifically for the 10th, 20th, 30th, 40th, and 50th cycles. It was observed that there is no significant difference in the composition of the correlation coefficient matrix between the total cycle data and the individual cycle data. PC1 consistently shows a very high correlation with capacity, exceeding 0.95, regardless of the cycle. Although PC1 and average voltage have a strong correlation over all cycles with a coefficient of 0.86, it is lower compared to the correlation with capacity. Since the correlation between capacity and average voltage is high at 0.67, this likely contributes to the strong correlation between PC1 and average voltage. PC2 has a relatively high but weak correlation of 0.44 with average voltage compared to capacity. Further principal components, starting with PC3, show inconsistently low correlations. **Figures S12** and **S13** illustrate the cycle-wise relationship between the 1st and 2nd principal components and capacity and average voltage, respectively. These relationships appear to be consistent, almost independent of the number of cycles.

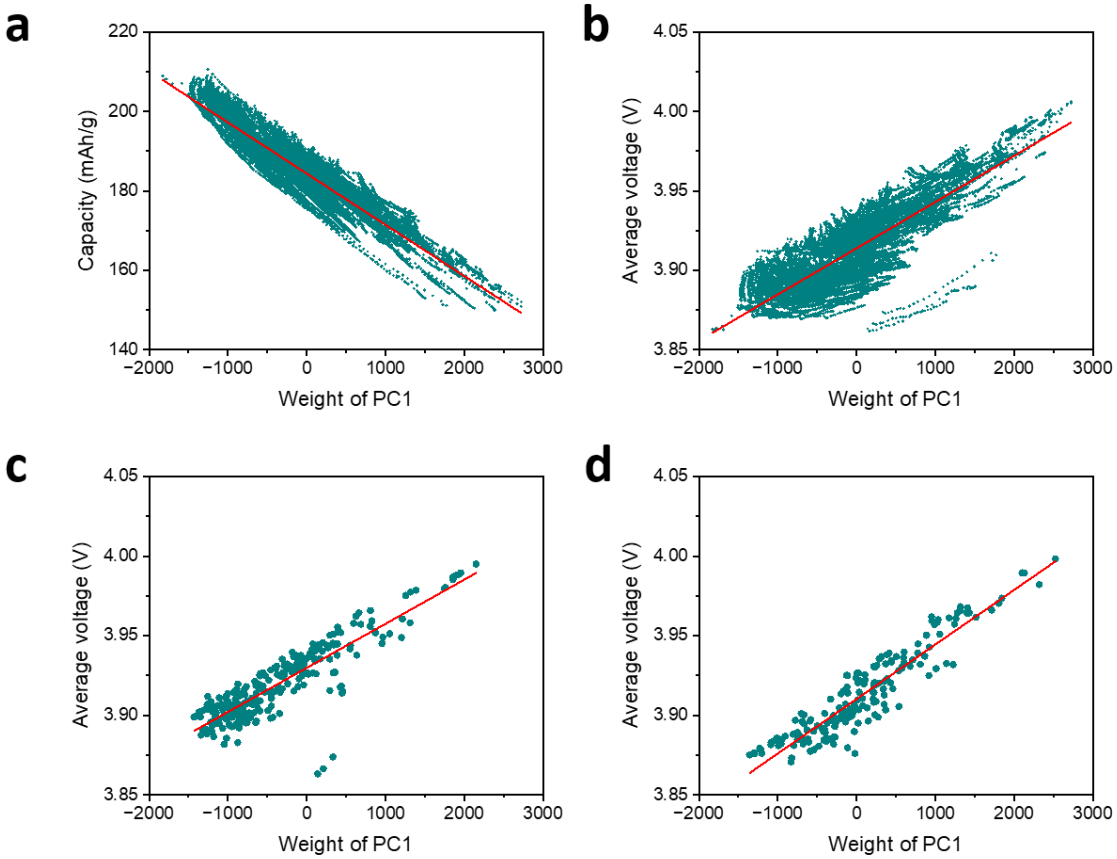


**Figure S10.** Relationship between the first principal component and **(a)** capacity for all cycles and average voltage for **(b)** all cycles, **(c)** fresh cells and **(d)** aged cycles.

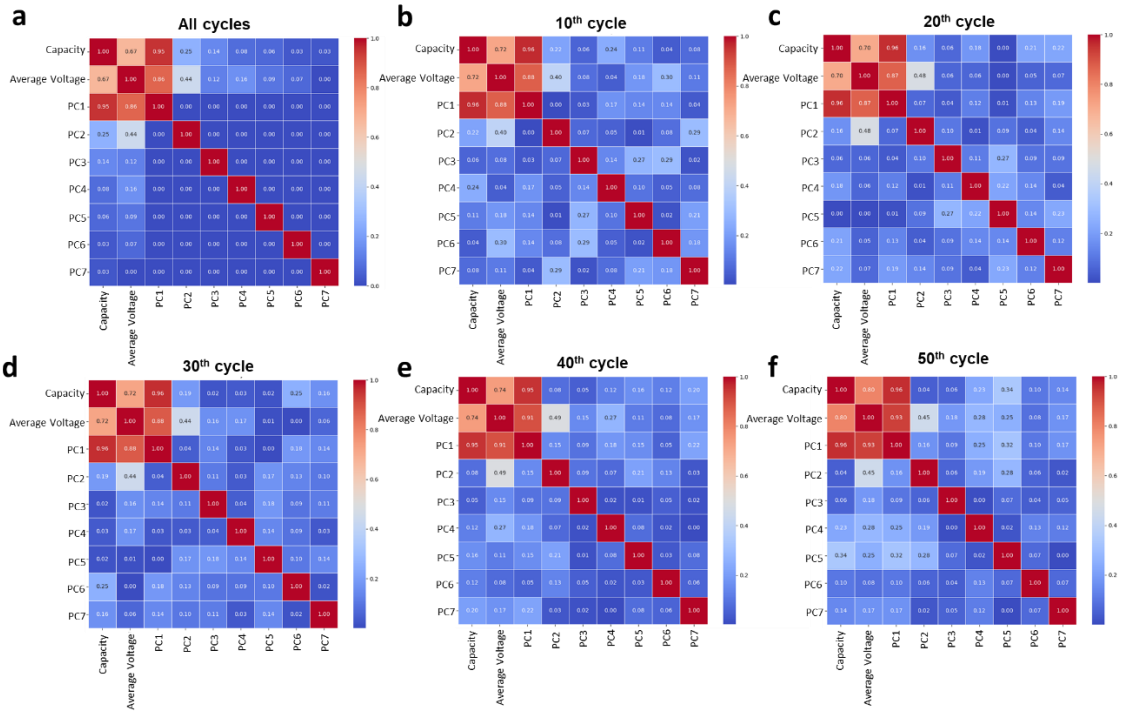


**Figure S11.** Correlation matrix between PCA components and battery performance for **(a)** all cycles, **(b)** $10^{th}$, **(c)** $20^{th}$, **(d)** $30^{th}$, **(e)** $40^{th}$ and **(f)** $50^{th}$ cycle.

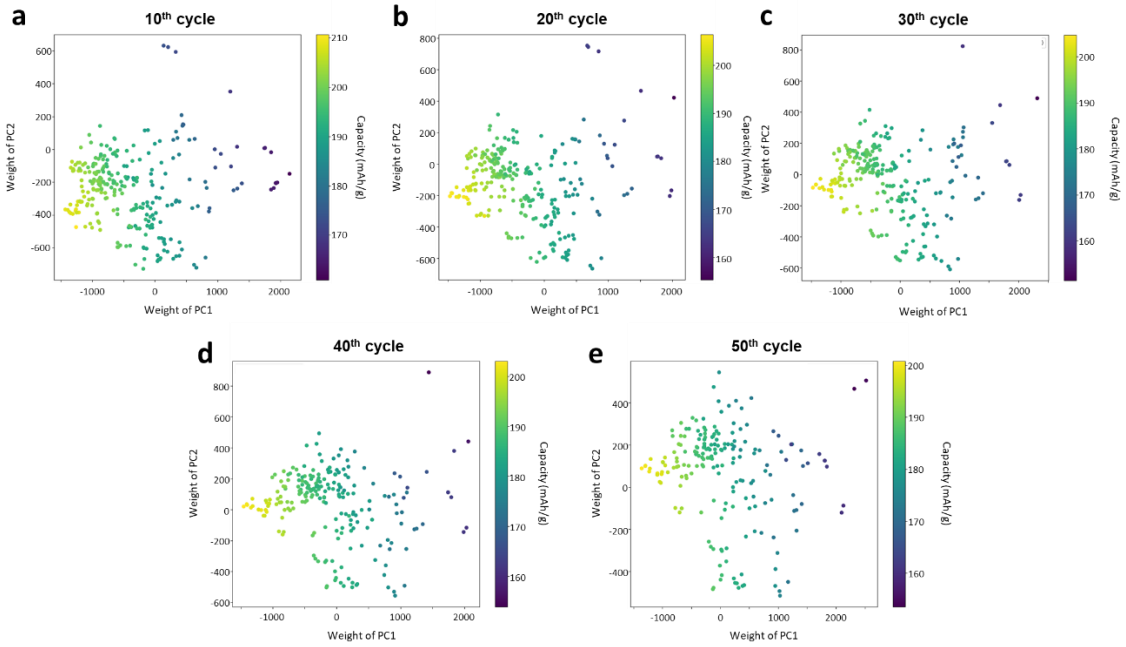


**Figure S12.** Relationship between the principal components and capacity for **(a)** 10$^{th}$, **(b)** 20$^{th}$, **(c)** 30$^{th}$, **(d)** 40$^{th}$ and **(e)** 50$^{th}$ cycle.

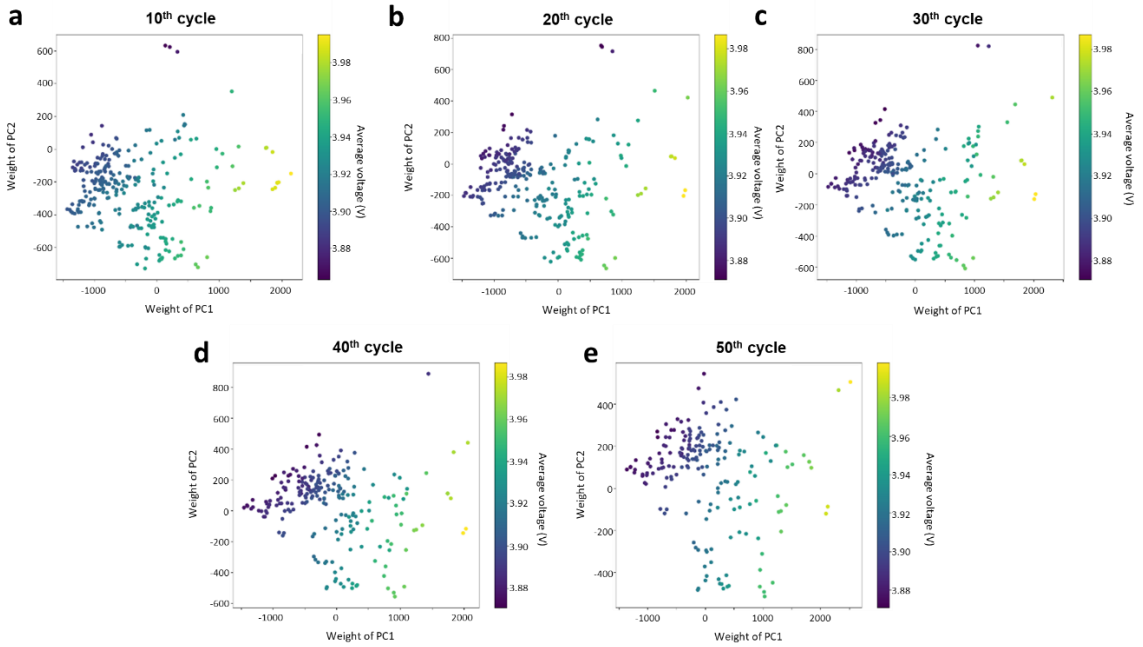


**Figure S13.** Relationship between the principal components and average voltage for **(a)** 10th, **(b)** 20th, **(c)** 30th, **(d)** 40th and **(e)** 50th cycle.

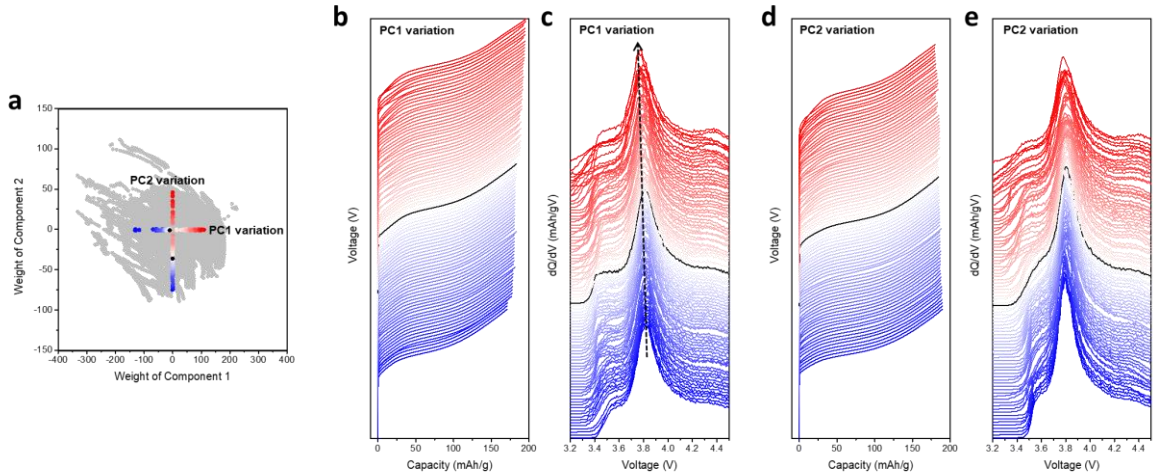


**Figure S14.** Effect of PC1 and PC2 on electrochemical behavior from the actual dataset. **(a)** Weights of the PC1 and PC2 across all data points, with the regions of PC1 and PC2 variation indicated. Variation in the **(b)** charge curves and **(c)** dQ/dV curves as influenced by the PC1. Variation in the **(d)** charge curves and **(e)** dQ/dV curves as influenced by the PC2.

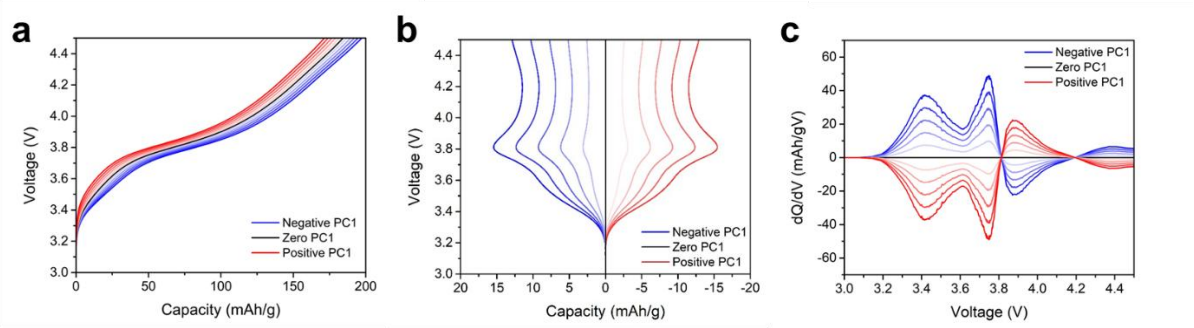


**Figure S15. (a)** Charge curves and changes in **(b)** charge curves and **(c)** dQ/dV curves as influenced by the PC1.

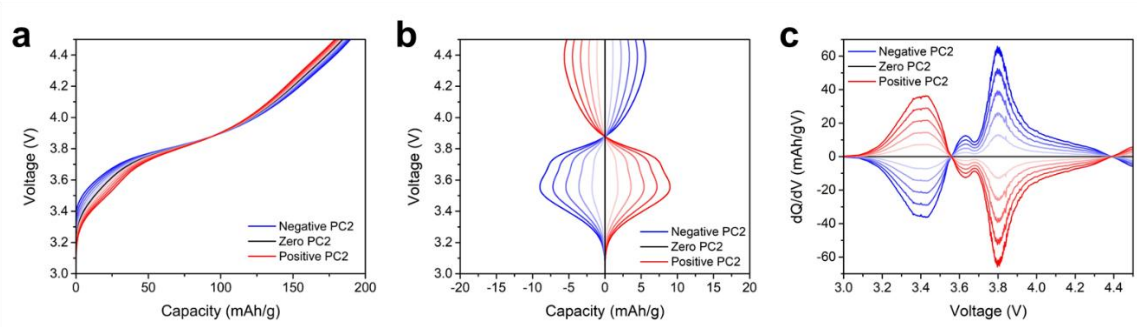


**Figure S16. (a)** Charge curves and changes in **(b)** charge curves and **(c)** dQ/dV curves as influenced by the PC2.

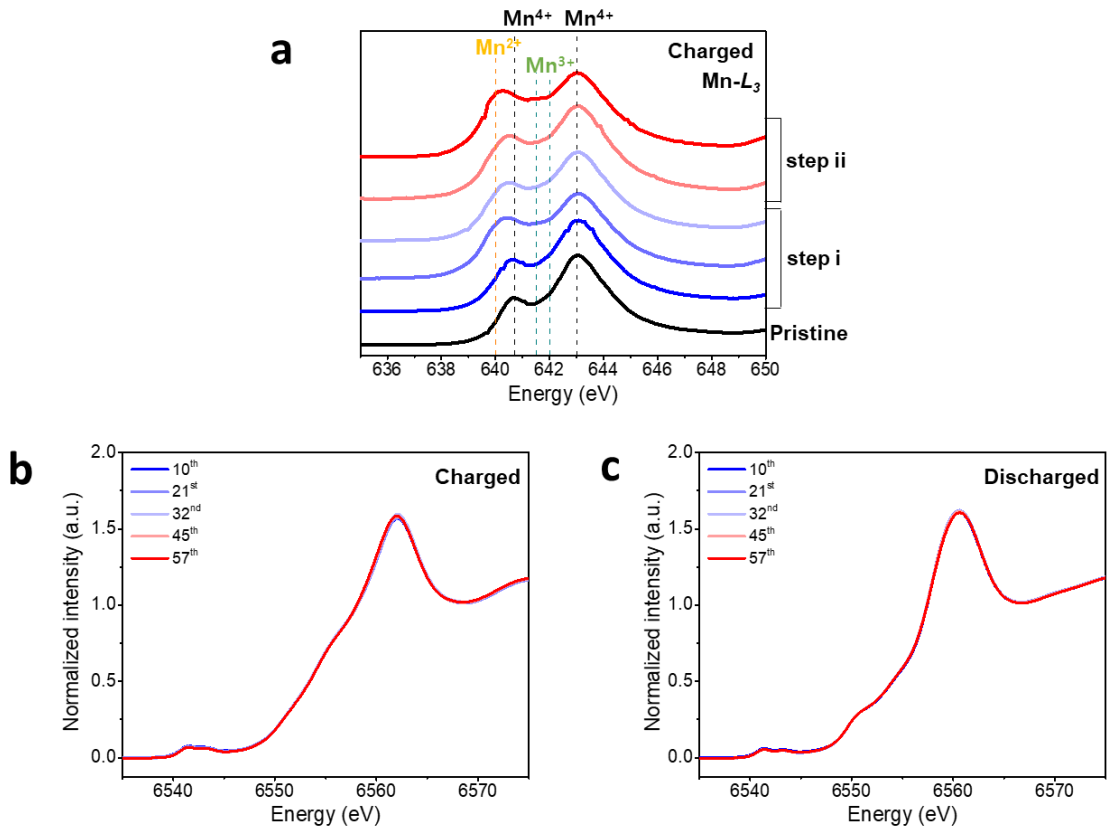


**Figure S17. (a)** Mn L-edge sXAS spectra of charged electrodes. Mn K-edge XANES spectra of **(b)** charged and **(c)** discharged electrodes.

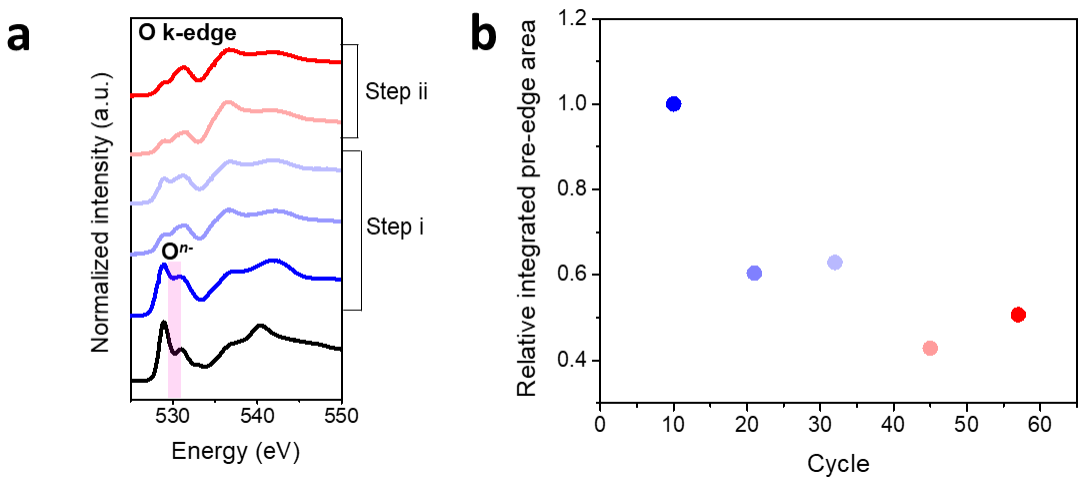


**Figure S18. (a)** O K-edge sXAS spectra of charged electrodes. **(b)** The variation of the integrated area under the pre-edge in the low-energy region (below 533 eV) of O K-edge sXAS during cycling.

**Table S6.** Fitted resistance of the charged electrode.

| Cycle | $R_s$ (ohm) | $R_{sf}$ (ohm) | $R_{ct}$ (ohm) |
|---|---|---|---|
| $10^{th}$ | 2.218 | 2.874 | 20.74 |
| $21^{st}$ | 2.072 | 3.982 | 21.31 |
| $32^{nd}$ | 2.914 | 5.018 | 24.71 |
| $45^{th}$ | 3.519 | 7.384 | 43.97 |
| $57^{th}$ | 3.954 | 11.01 | 45.89 |

## Supplementary Section 6. Extension of PCA to Include Test Dataset and Component Consistency Analysis (Table S7, Figure S19 – 21)

We extended Principal Component Analysis (PCA) to include the test dataset in addition to the training dataset. We used the same approach for PCA as described in the Methods section, with the only difference being the addition of the test dataset to the original training set.

The new PCA components derived from the combined dataset of the training and test sets are shown in **Figure S19**. These components are essentially identical to those obtained from the training set alone, as shown in **Figure 2a**.

**Figure S20** illustrates the relationship between the new 1st Principal Components from the PCA (which now includes both training and test sets) and two key variables: Capacity and Average Voltage. This relationship shows the same trends as observed in **Figure S8**, indicating consistent correlations and patterns even when the test dataset is included in the PCA.

**Table S7** and **Figure S21** demonstrate that the reconstruction accuracy of the test set is similar to that of the training set (**Table S5**).

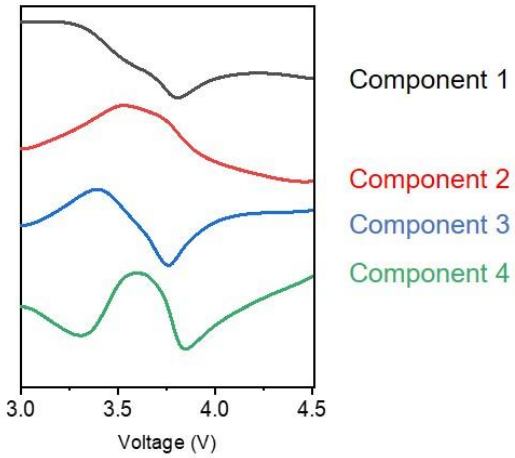


**Figure S19.** New PCA components derived from the combined dataset of training and test sets.

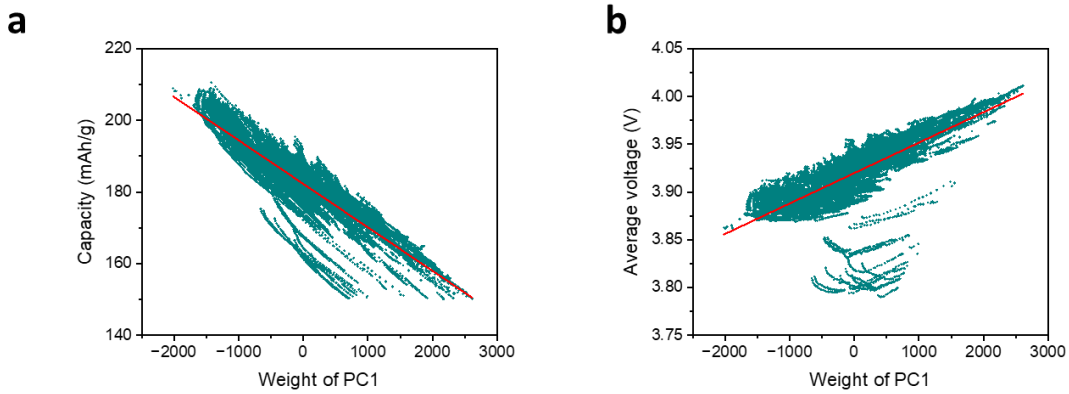


**Figure S20. (a)** PC1 – capacity curve and **(b)** PC1 – average voltage including test set that contains data outside of AD.

**Table S7.** Accuracy of reconstructed curve of test set based on PCA components obtained from train set

| | mean | median | 25% | 75% | 90% | 95% |
|---|---|---|---|---|---|---|
| RMSE | 0.3836 | 0.2472 | 0.1891 | 0.3367 | 1.1388 | 1.4650 |

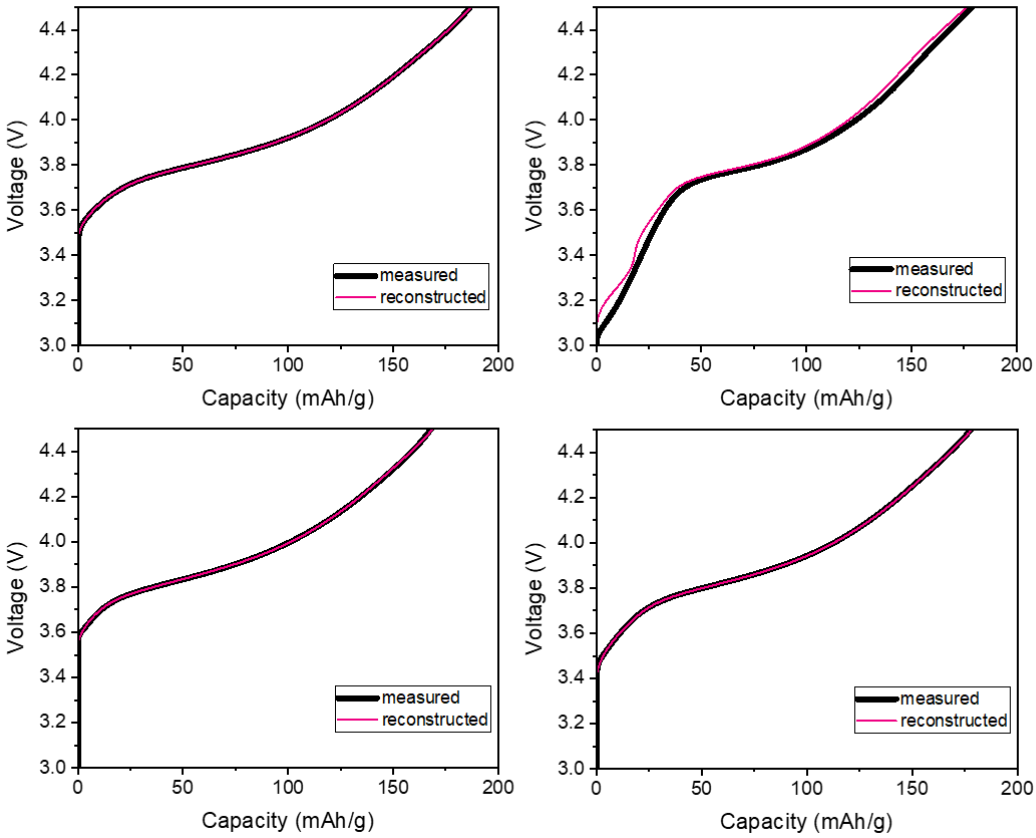


**Figure S21.** Reconstructed charge curves of test set based on PCA components obtained from train set.

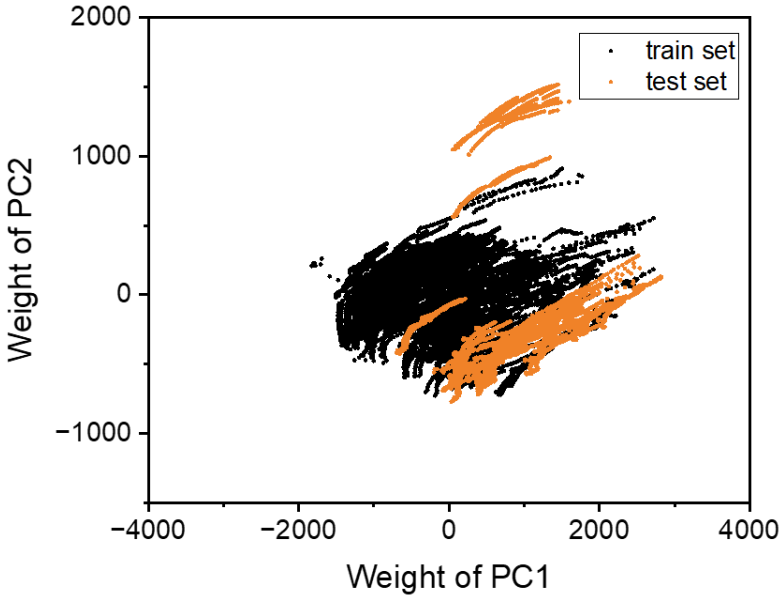


**Figure S22.** Charge curve prediction outside applicability domain for all cells.

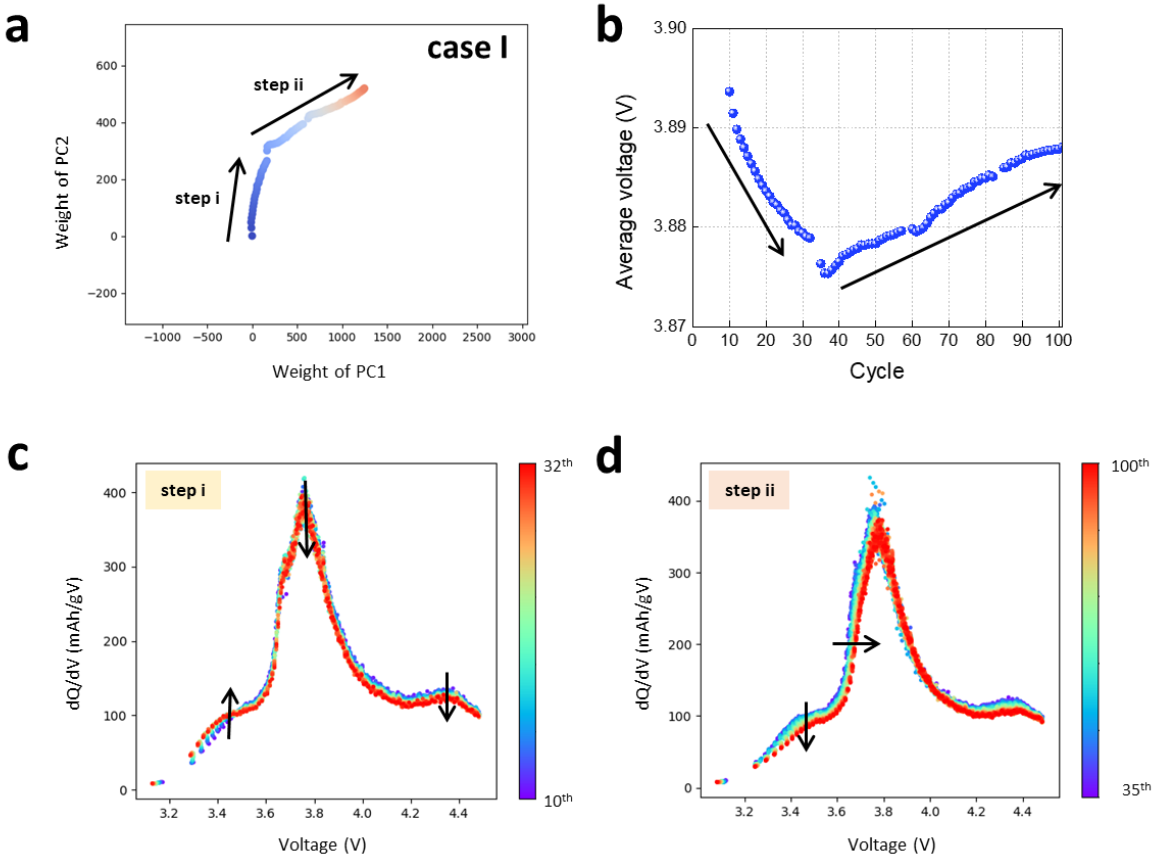


**Figure S23**. **(a)** Degradation behavior of LRLO (Case I). **(b)** Voltage retention with degradation. **(c)** and **(d)** Differential charge curve with degradation.

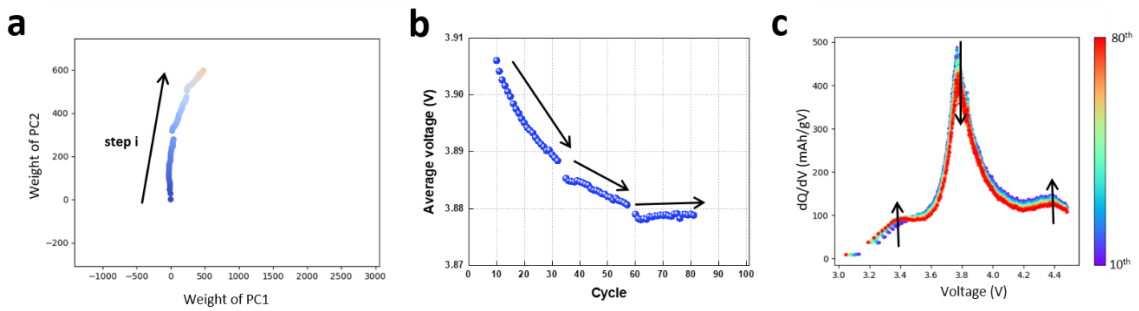


**Figure S24**. **(a)** Degradation behavior of LRLO (Case II). **(b)** Voltage retention with degradation. **(c)** Differential charge curve with degradation.

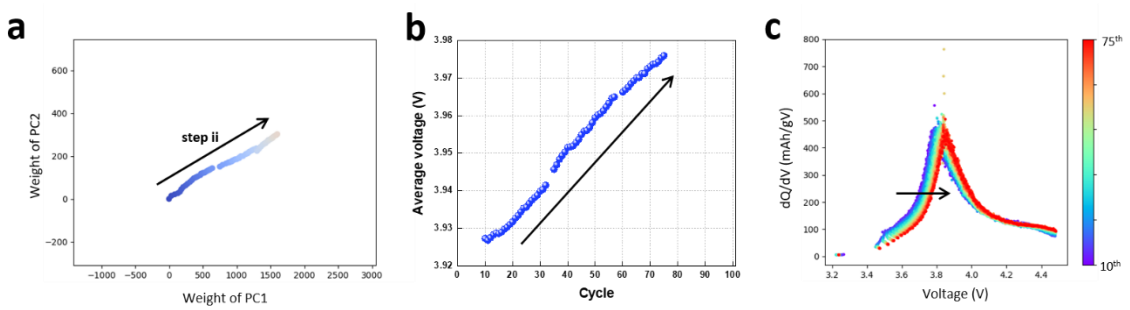


**Figure S25. (a)** Degradation behavior of LRLO (Case III). **(b)** Voltage retention with degradation. **(c)** Differential charge curve with degradation.

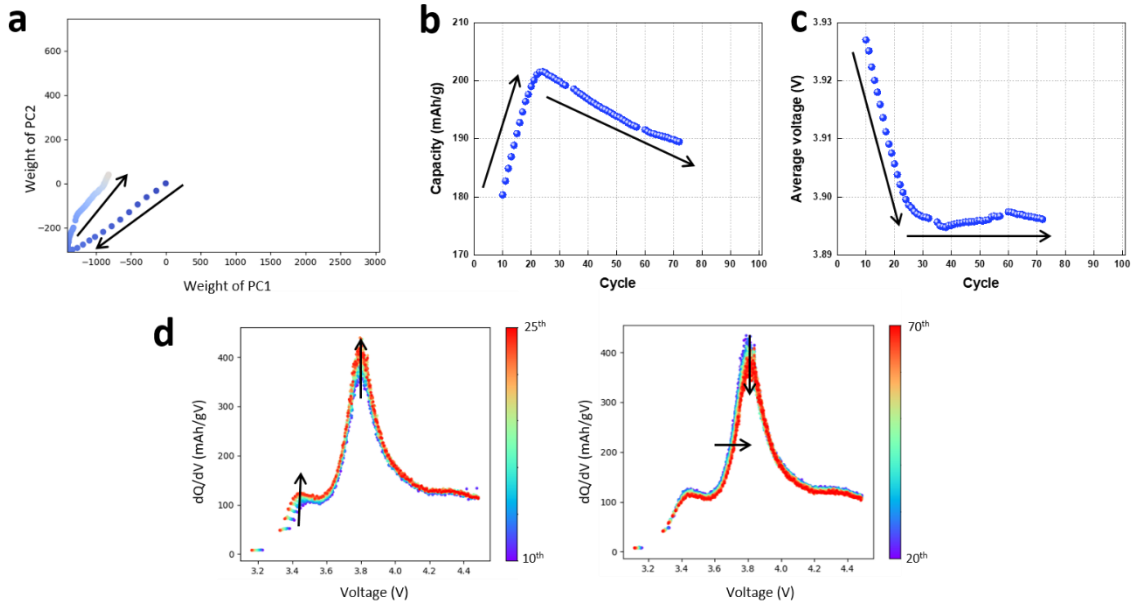


**Figure S26. (a)** Degradation behavior of LRLO (Case IV). **(b)** Capacity retention and **(c)** Voltage retention with degradation. **(d)** Differential charge curve with degradation.

**Table S8.** Accuracy of predicted curve by the number of PCA components via linear regression.

| Number of components | RMSE | | | | | |
|---|---|---|---|---|---|---|
| | mean | median | 25% | 75% | 90% | 95% |
| 1 | 6.4055 | 5.2005 | 3.5205 | 7.4131 | 12.4517 | 18.3536 |
| 2 | 3.2406 | 2.2046 | 1.2691 | 3.4555 | 9.4899 | 11.1451 |
| 3 | 2.9558 | 1.5360 | 0.9043 | 2.6254 | 10.6347 | 12.6155 |
| 4 | 2.7954 | 2.1429 | 1.0992 | 4.1307 | 6.1842 | 7.2780 |
| 5 | 2.7943 | 2.1235 | 1.1597 | 4.3097 | 5.5248 | 6.8018 |
| 6 | 2.5411 | 1.6633 | 0.9691 | 3.4374 | 5.8958 | 7.4763 |
| 7 | 2.3895 | 1.6813 | 1.0225 | 3.2946 | 5.1526 | 6.3071 |
| 8 | 3.1350 | 2.7585 | 1.6229 | 4.1508 | 6.0397 | 7.0650 |
| 9 | 3.5915 | 3.2580 | 2.0245 | 4.9959 | 6.3477 | 7.1838 |
| 10 | 3.5231 | 2.9068 | 1.9773 | 4.7201 | 6.8535 | 7.9810 |

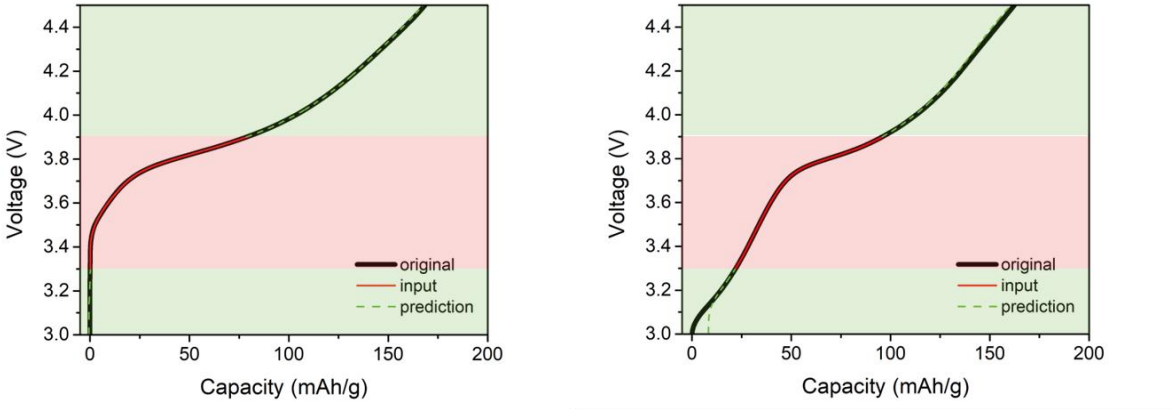


**Figure S27.** Predicted charge curves with 3.3 V - 3.9 V as input. (left) good case and (right) bad case.

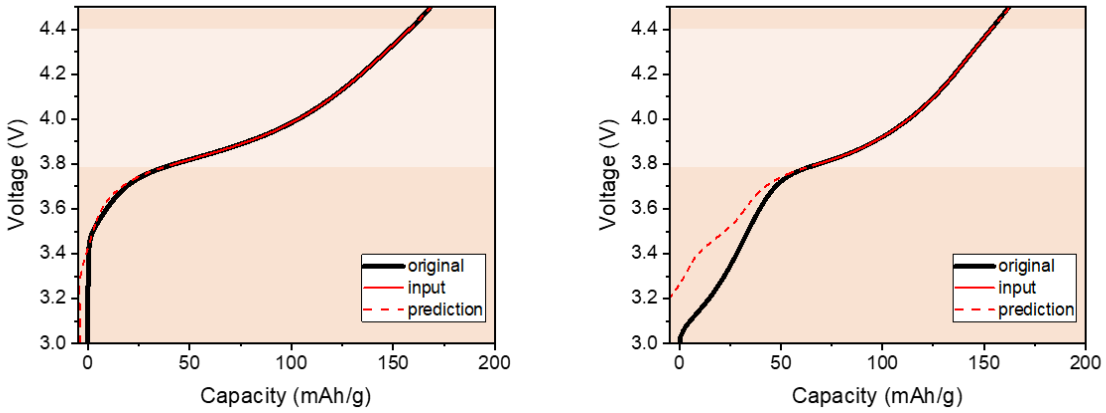


**Figure S28.** Predicted charge curves with 3.8 V – 4.4 V as input. (left) good case and (right) bad case.

## Supplementary Section 7. Application of Principal Component Analysis (PCA) on Capacity Data: Example (Figure S29 - 31, Table S9)

In this supplementary section, we present a simple example to illustrate the application of Principal Component Analysis (PCA) on cumulative capacity data measured at various voltage points. As shown in the capacity-voltage curve in **Figure S29**, we have prepared five virtual training datasets. In this example, capacities are recorded at six voltage points, which are tabulated in **Table S9**.

The data can be represented as the data matrix $X_{train}$:

$$X_{train} = \begin{bmatrix} 0 & 50 & 50 & 100 & 100 & 150 \\ 0 & 75 & 75 & 150 & 150 & 150 \\ 0 & 0 & 0 & 75 & 75 & 150 \\ 0 & 49 & 49 & 98 & 98 & 147 \\ 0 & 51 & 51 & 102 & 102 & 153 \end{bmatrix}$$

Each row of $X_{train}$ corresponds to one of the five datasets, and each column corresponds to one of the six voltage points.

To perform PCA, we first compute the mean of each column (voltage point):

$$\mu = \frac{1}{n}\sum_{i=1} X_{train,i} = [0, 45, 45, 105, 105, 150]$$

where *n* is the number of datasets.

Next, we center the data by subtracting the mean vector $\mu$ from each row of $X_{train}$:

$$X' = X_{train} - \mu = \begin{bmatrix} 0 & 5 & 5 & -5 & -5 & 0 \\ 0 & 30 & 30 & 45 & 45 & 0 \\ 0 & -45 & -45 & -30 & -30 & 0 \\ 0 & 4 & 4 & -7 & -7 & 3 \\ 0 & 6 & 6 & -3 & -3 & 3 \end{bmatrix}$$

The centered data matrix $X'$ has zero mean in each column.

We then compute the covariance matrix $C$ of the centered data:

$$C = \frac{1}{n-1} X'^T X'$$

From the covariance matrix *C*, we perform eigenvalue decomposition to obtain the principal components. The eigenvectors of *C* correspond to the principal components, and the associated eigenvalues indicate the amount of variance explained by each component.

The principal components can be expressed using 3 components, yielding the approximation:

$$X' = \begin{bmatrix} 0 & 5 & 5 & -5 & -5 & 0 \\ 0 & 30 & 30 & 45 & 45 & 0 \\ 0 & -45 & -45 & -30 & -30 & 0 \\ 0 & 4 & 4 & -7 & -7 & 3 \\ 0 & 6 & 6 & -3 & -3 & 3 \end{bmatrix}$$

$$\approx \begin{bmatrix} 0 & -10 & -0.08 \\ -75 & 15 & 0.27 \\ -75 & 15 & -0.03 \\ 3.006 & -11.024 & 2.906 \\ -3.006 & -8.976 & -3.066 \end{bmatrix} \begin{bmatrix} 0 & 5 & 5 & -5 & -5 & 0 \\ 0 & 30 & 30 & 45 & 45 & 0 \\ 0 & -45 & -45 & -30 & -30 & 0 \end{bmatrix}$$

$$= UV^T$$

where *U* is the matrix of scores (projections of the data onto the principal components), and $V^T$ is the transpose of the matrix of principal component loadings (eigenvectors).

Each row of U represents the weights for each sample, and each row of $V^T$ corresponds to the values of each principal component at each voltage point. By using these matrices, we can plot the weights of Component 1 and Component 2, resulting in **Figure S30**. Additionally, plotting the values of the principal components with respect to voltage yields **Figure S31**.

**Note**: To reproduce the PCA calculations and explore the data further, please refer to the provided Jupyter notebooks available on GitHub. The actual calculations using Python's scikit-learn package and the detailed calculation process performed in this study can be found at https://github.com/LRLO-PCA/Data-driven-insights-into-reaction-mechanism-of-Li-rich-cathodes.

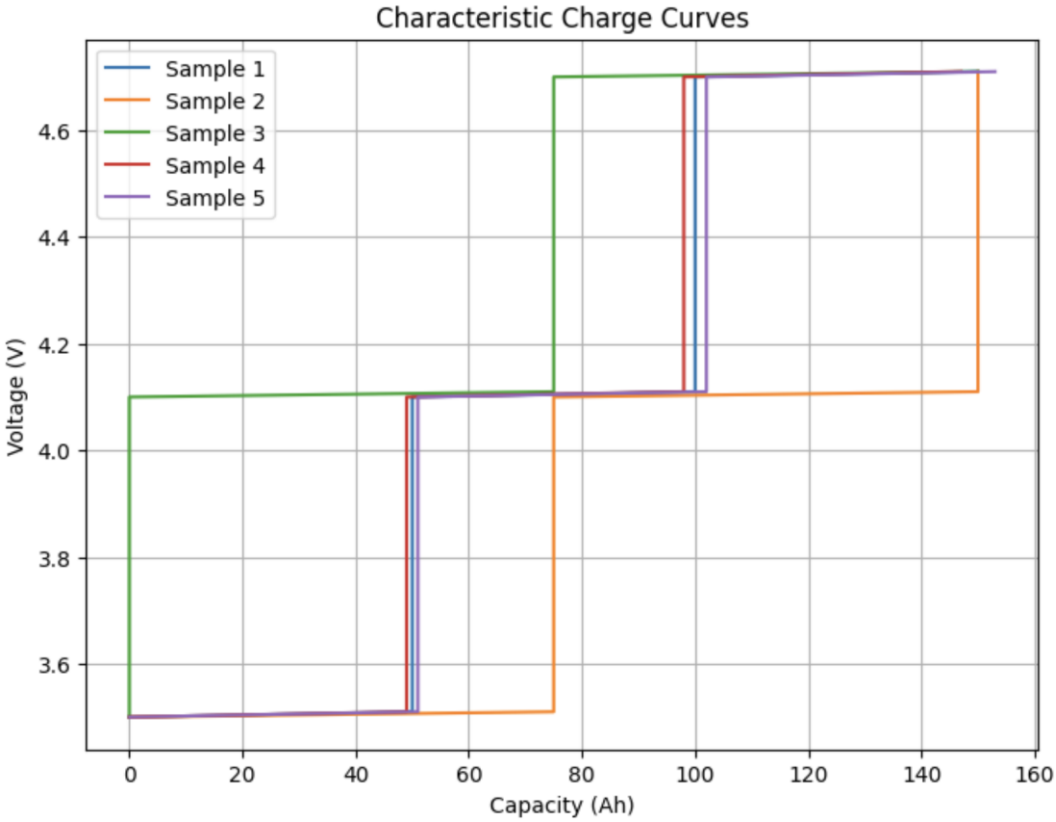


**Figure S29.** Capacity-voltage curve of the virtual training datasets.

**Table S9.** Capacities at six voltage points for five virtual training datasets.

| Voltage (V) | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| 3.50 | 0 | 0 | 0 | 0 | 0 |
| 3.51 | 50 | 75 | 0 | 49 | 51 |
| 4.10 | 50 | 75 | 0 | 49 | 51 |
| 4.11 | 100 | 150 | 75 | 98 | 102 |
| 4.70 | 100 | 150 | 75 | 98 | 102 |
| 4.71 | 150 | 150 | 150 | 147 | 153 |

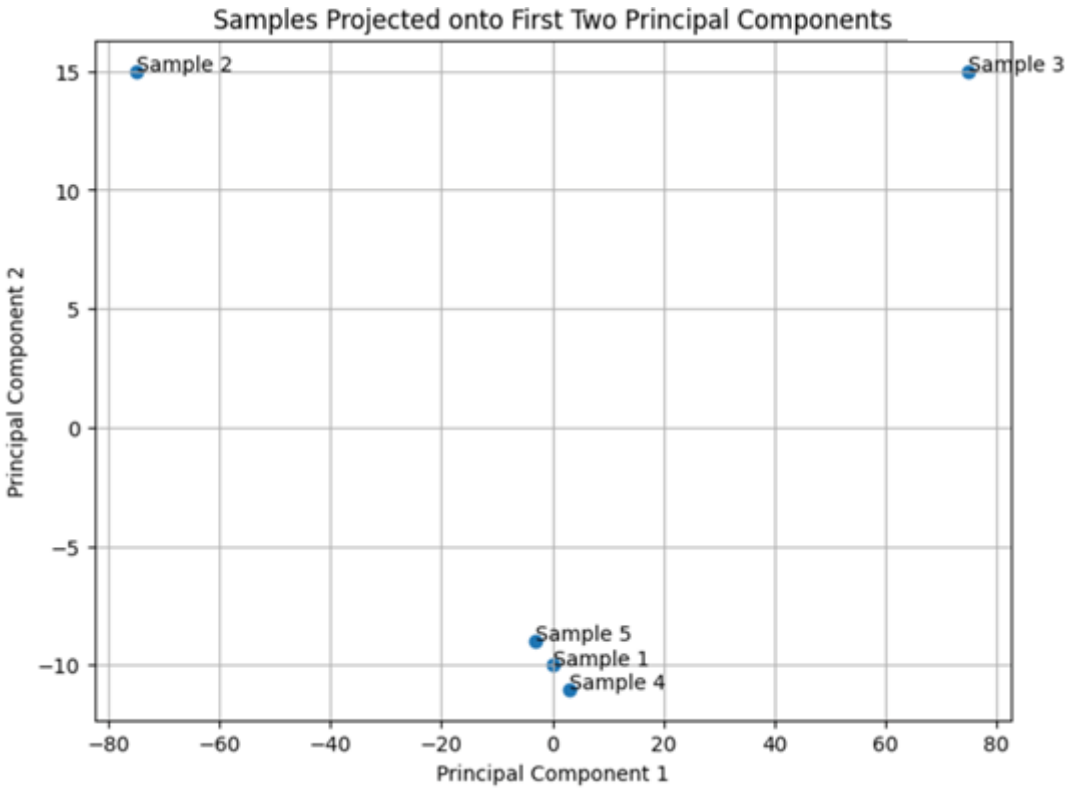


**Figure S30**. Scatter plot of the weights (scores) for Principal Component 1 and Principal Component 2 for each sample, derived from the scores matrix U.

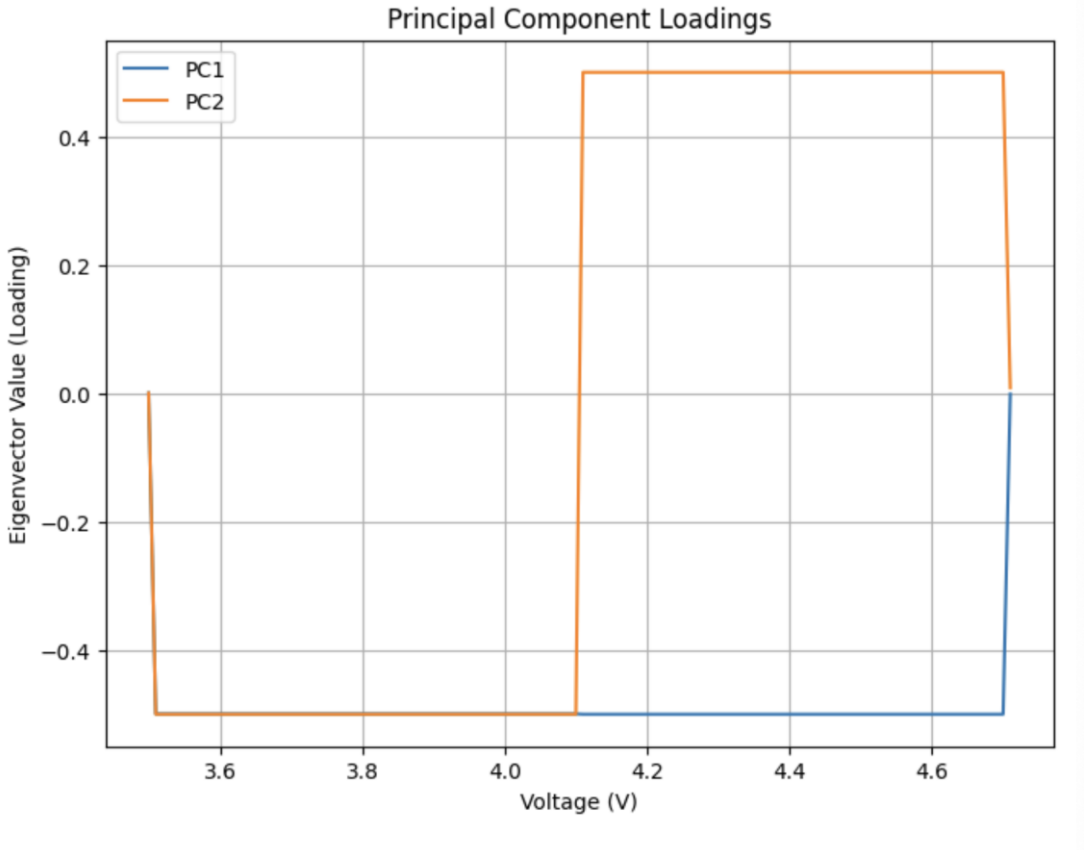


**Figure S31.** Visualization of the principal component for each voltage point, derived from the principal component loading matrix $V^T$.