

Domain knowledge-guided machine learning framework for state of health estimation in Lithium-ion batteries

*Original*

Domain knowledge-guided machine learning framework for state of health estimation in Lithium-ion batteries / Lanubile, Andrea; Bosoni, Pietro; Pozzato, Gabriele; Allam, Anirudh; Acquarone, Matteo; Onori, Simona. - In: COMMUNICATIONS ENGINEERING. - ISSN 2731-3395. - 3:1(2024). [10.1038/s44172-024-00304-2]

*Availability:*

This version is available at: 11583/2994713 since: 2024-11-22T16:51:02Z

*Publisher:*

Nature Portfolio

*Published*

DOI:10.1038/s44172-024-00304-2

*Terms of use:*

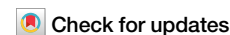
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

<https://doi.org/10.1038/s44172-024-00304-2>

# Domain knowledge-guided machine learning framework for state of health estimation in Lithium-ion batteries



Andrea Lanubile<sup>1,3</sup>, Pietro Bosoni<sup>1,3</sup>, Gabriele Pozzato<sup>1</sup>, Anirudh Allam<sup>1</sup>, Matteo Acquarone<sup>2</sup> & Simona Onori<sup>1</sup> ✉

Accurate estimation of battery state of health is crucial for effective electric vehicle battery management. Here, we propose five health indicators that can be extracted online from real-world electric vehicle operation and develop a machine learning-based method to estimate the battery state of health. The proposed indicators provide physical insights into the energy and power fade of the battery and enable accurate capacity estimation even with partially missing data. Moreover, they can be computed for portions of the charging profile and real-world driving discharging conditions, facilitating real-time battery degradation estimation. The indicators are computed using experimental data from five cells aged under electric vehicle conditions, and a linear regression model is used to estimate the state of health. The results show that models trained with power autocorrelation and energy-based features achieve capacity estimation with maximum absolute percentage error within 1.5% to 2.5%.

The pressing concern of global warming is driving a global shift towards electrified mobility. With the transportation sector contributing to approximately 12% of all global emissions<sup>1</sup>, adjustments are required in order to transition to a zero-emissions energy sector. Studies by the Intergovernmental Panel on Climate Change<sup>1</sup> and the International Energy Agency<sup>2</sup> emphasize the critical need for clean transportation solutions to address the urgent issue of climate change. This has driven governments and policymakers to innovate and collaborate in advancing electric vehicle (EV) technologies.

Lithium-ion batteries (LIBs) are the preferred energy storage technology for EVs due to their superior power and energy density, which enables longer driving ranges compared to other battery technologies<sup>3</sup>. For a compelling and sustainable EV mass market, accurate state of health (SOH) estimation<sup>4</sup> and remaining useful life (RUL)<sup>5</sup> prediction of LIB systems are essential. Existing methods for SOH estimation and RUL prediction can be broadly divided into model-based and data-driven approaches. Model-based estimation approaches rely on empirical or equivalent circuit models (ECMs), or electrochemical models, and formulate estimation algorithms around them. Various ECM-based filters for SOH estimation have been proposed in the literature, including Extended Kalman Filter<sup>6</sup>, dual and joint Extended Kalman Filter<sup>7</sup>, Unscented Kalman Filter<sup>8</sup>, Adaptive Extended Kalman Filter<sup>9</sup>, Particle Filter<sup>10</sup> and genetic algorithms<sup>11</sup>.

Other methods for SOH estimation and RUL prediction, utilizing empirical degradation models, include Unscented Kalman Filters<sup>12</sup> and Particle Filters<sup>13</sup>. In a Bayesian Monte Carlo approach<sup>14</sup>, the parameters of an empirical capacity model are updated to compute the posterior probability density function for capacity fade prediction. Despite their simplicity, these methods lack explicit physical understanding and require significant calibration effort. Also, electrochemical battery models<sup>15–17</sup> which demand increased computational power, have been employed, and adaptive observers based on the enhanced single particle model<sup>18,19</sup> have been tested in a battery-in-the-loop setup.

With the advancement in cloud computing technologies and Internet of things, data-driven methods for battery SOH estimation, such as linear regression, gaussian process regression, support vector machine, or artificial neural network, have gained traction in recent years<sup>20</sup>. For instance, multiple linear regression models have been trained using descriptive features of the voltage distribution<sup>21</sup> or incremental capacity curves<sup>22</sup> to predict capacity fade and resistance increase. Among the more sophisticated prediction models, Gaussian process regression models have been used for capacity estimation<sup>23</sup>, taking as inputs different statistical features extracted from the charging curves<sup>24</sup>. Furthermore, neural networks<sup>25</sup> are used to establish relationships between input features, such as equivalent circuit model parameters and state of charge (SOC), and battery capacity fade. Regarding

<sup>1</sup>Energy Science & Engineering, Stanford University, Stanford, CA, USA. <sup>2</sup>Energy Department, Politecnico di Torino, Torino, Italy. <sup>3</sup>These authors contributed equally: Andrea Lanubile, Pietro Bosoni. ✉e-mail: [sonori@stanford.edu](mailto:sonori@stanford.edu)

RUL prediction, support vector machines<sup>4</sup> and random forests<sup>26</sup> are utilized. These methods are effective in forecasting the remaining operational life-span of batteries based on historical data and operational conditions. Battery SOH estimation works can be classified into three primary categories based on the dataset used for development. The first category includes datasets acquired from field operations, which accurately reflect the aging phenomena affecting batteries in real-world EVs driving<sup>27</sup>. However, a challenge with these datasets is the absence of a baseline for evaluating SOH. To address this limitation, several studies<sup>28–30</sup> utilize internal resistance as a direct metric for assessing battery SOH. Alternatively, some studies<sup>31</sup> suggest using the peak values derived from incremental capacity curves to overcome this challenge. However, these metrics can be challenging to evaluate due to their strong dependence on operating conditions, such as temperature. Capacity fade is often measured using Coulomb counting<sup>32,33</sup>, which involves integrating battery management system (BMS) current over a limited SOC window. This method, however, may produce inaccurate results due to high sensor noise and quantization in-vehicle sensors. Conversely, tests are conducted in a temperature-controlled environment<sup>34</sup> to ensure consistent capacity measurements that serve as ground truth. Here, SOH is estimated using supervised learning models that directly utilize BMS signals—such as voltage, current, SOC, and pack temperature—as inputs. The second category of SOH estimation works relies on datasets collected in laboratory settings<sup>24,35–38</sup>. In these datasets, cells are cycled with current profiles that do not accurately represent actual EV battery operation. As a result, features computed using these datasets are not, in general, transferable nor generalizable to real-world applications. The third category of datasets utilizes data collected in laboratory settings aimed to mimic EV real-use case scenarios. Examples include ARTEMIS<sup>39</sup> and the Urban Dynamometer Driving Schedule (UDDS)<sup>40</sup>. These datasets provide more realistic conditions for testing and developing battery SOH estimation algorithms, still providing ground truth capacity through periodic reference performance tests (RPTs).

Using these datasets, different machine learning algorithms, such as support vector machine<sup>4</sup>, Gaussian process regression, or neural network<sup>41</sup>, have been used to estimate SOH for batteries undergoing EV driving cycles using statistical features from current, voltage, and temperature signals as input features. Other studies have proposed various physics-based health indicators to estimate battery capacity fade, e.g., features derived from the ECM<sup>42</sup> or the time taken for voltage to rise from a low to a high level during the charging process<sup>43</sup>. Another crucial physical quantity linked to battery aging is internal resistance, with several nuanced indicators proposed in the literature<sup>44,45</sup>. Most of these indicators are computed using ECMs and algorithms such as recursive least squares. However, these methods typically come with increased computational requirements. Another approach<sup>27,46</sup> involves evaluating the indicators during the vehicle acceleration (discharge) and braking (charge), requiring less computational power and facilitating its real-time implementation and integration within the vehicle BMS. Another physics-based SOH indicator is charging impedance<sup>27</sup>, which combines variations in electrolyte resistance, charge transfer resistance, and polarization due to aging. This feature can be extracted from the initial portion of the charging phase<sup>47</sup>. Additionally, the energy during charging and discharging can offer valuable insights into battery degradation. Energy metrics are typically calculated over extended portions of full charging profiles to effectively estimate capacity fade<sup>48–50</sup>.

The health features used in previous studies are typically based on idealized constant charging and discharging profiles. However, these profiles do not accurately reflect how electric vehicles are charged and discharged in real-world conditions. Most research has focused on extracting health indicators during complete, repetitive charging cycles, where the battery is charged from a low SOC to a high SOC. In reality, charging patterns are much more variable, and it's uncommon for batteries to go through full cycles or always follow the same charging profile. This discrepancy makes it difficult to apply findings from controlled experiments directly to real-world EV use.

The contributions of this work are the following. First, our work systematically formulates various SOH indicators based on domain knowledge and proposes a framework for their integration into BMS. The proposed SOH indicators include: power autocorrelation, resistance, charging impedance, energy during charging, and energy during discharging. Second, unlike previous research<sup>51</sup> that focused on voltage signal autocorrelation, here, the power autocorrelation is used to quantify the battery's power-delivery capability over time. Additionally, the proposed SOH indicators are derived from an experimental dataset<sup>40</sup> that replicates real-world EV battery operation. Unlike most prior studies<sup>48–50</sup> that rely on constant current discharging profiles, in this work, energy consumption is evaluated during discharging under realistic EV driving cycles. Moreover, a windowed approach is proposed to assess energy consumption during charging, thereby improving the effectiveness of the energy as an indicator of battery health, especially in scenarios involving partial charging. Furthermore, we modify the formulation of the charging impedance indicator<sup>27</sup> by calculating it over an optimized voltage window and averaging the values within this range to improve accuracy and reliability in assessing battery health during partial charging. Through correlation analysis, power autocorrelation, energy during charging, and energy during discharging emerge as the most effective indicators for capacity estimation. It is worth noting that the proposed SOH indicators are agnostic to specific battery chemistries. Moreover, they operate independently of cumulative data such as total aging cycles or ampere-hour throughput. This design choice helps mitigate inaccuracies that could arise from sensor errors or insufficient data. These indicators can be easily evaluated during EV operation. This makes them suitable for real-time deployment and integration into existing BMS strategies.

The battery capacity is estimated through the machine learning pipeline shown in Supplementary Note 1 where 1) SOH indicators are first extracted from the experimental dataset<sup>40</sup>, 2) the correlation between the indicators and SOH is analyzed through a regression analysis, and 3) a linear regression model (LRM) is trained to estimate capacity fade. The results show that models trained using power autocorrelation and energy-based features obtain capacity estimation with absolute percentage errors (APE) ranging between 1.5% and 2.5%.

Previous works that used linear regression models to estimate battery SOH<sup>21,22</sup>, are based on features selected over simplistic charge/discharge profile not representative of EV driving.

## Results

The SOH indicators are extracted using data from five Nickel Manganese Cobalt (NMC)/Graphite cells<sup>40</sup> (reported in Table 1 and detailed in Sec. Cell cycling and experimental dataset). In this section, a thorough analysis of each indicator and regression analysis is carried out, and the estimation results obtained by the linear regression model are shown.

### SOH indicators analysis

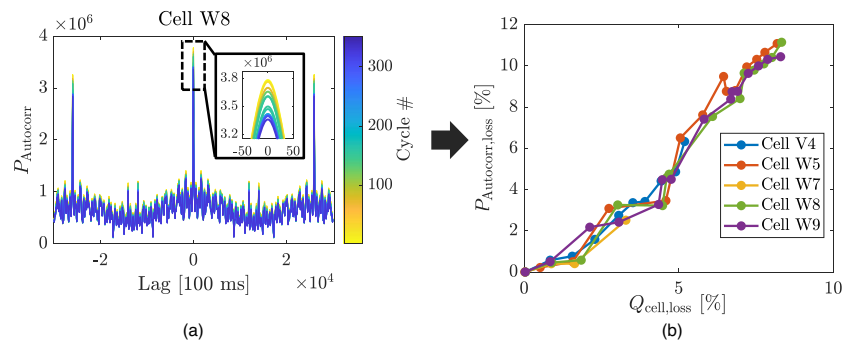
**Power autocorrelation function.** The autocorrelation function of the battery's power signal, evaluated during discharge, has shown to offer valuable insights into the battery SOH<sup>51</sup>. Assuming that discharging occurs periodically with an identical current profile, the power

**Table 1 | Battery cells**

Cell identifier	CC-a [1/h]	RPTs [#]	Cycles [#]
Cell V4	C/4	10	244
Cell W5	C/2	14	369
Cell W7	C/4	5	141
Cell W8	C/2	14	347
Cell W9	1C	14	341

Cells identifiers, charge C-rate, total number of RPTs and total number aging cycles of the battery cells<sup>40</sup> used in this work. Tests were conducted at a controlled temperature of 23 °C. See Supplementary Note 2 for further information on the dataset.

**Fig. 1 | Power autocorrelation.** **a** Power autocorrelation ( $P_{\text{Autocorr}}$ ) profiles calculated over discharge profiles for cell W8 throughout its lifetime. **b** Percentage decrease of the peak amplitude at null delay ( $P_{\text{Autocorr,loss}}$ ) is plotted against the percentage capacity loss ( $Q_{\text{cell,loss}}$ ).



autocorrelation  $P_{\text{Autocorr}}$  indicator varies as the battery ages, providing a method to monitor battery health.

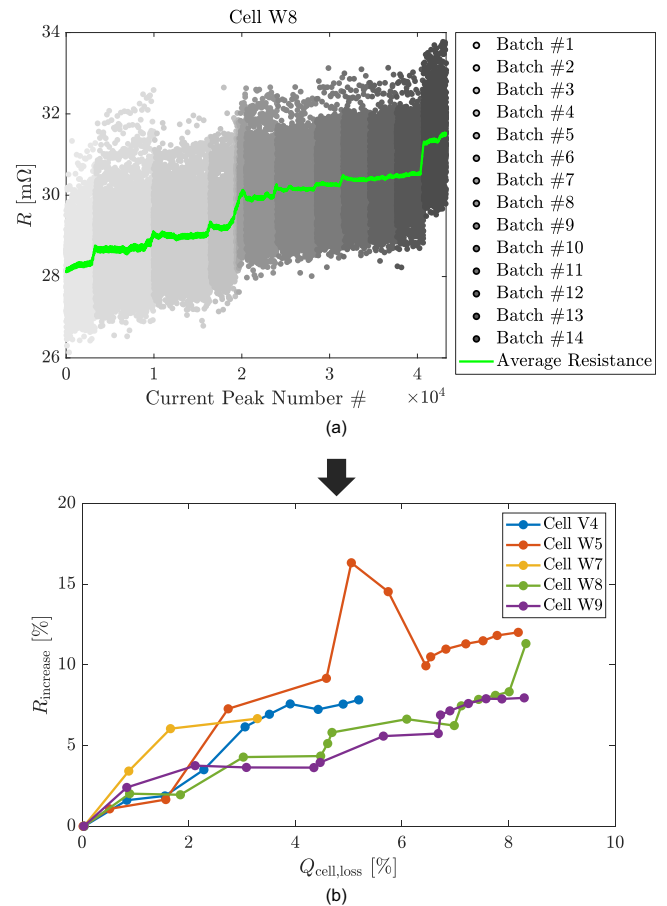
Of particular interest is the change in the central peak of the power autocorrelation function (see Fig. 1a). The reduction in this peak, defined as  $P_{\text{Autocorr,loss}} = (P_{\text{Autocorr,fresh}} - P_{\text{Autocorr,i}})/P_{\text{Autocorr,fresh}} \cdot 100$  where  $P_{\text{Autocorr,fresh}}$  is the central peak value for the fresh cell and  $P_{\text{Autocorr,i}}$  is the central peak value during cycle  $i$ , correlates with a decrease in capacity. This relationship is illustrated in Fig. 1b, where the power autocorrelation loss shows a strong linear relation with capacity loss. Capacity loss is calculated as  $Q_{\text{cell,loss}} = (Q_{\text{cell,fresh}} - Q_{\text{cell,i}})/Q_{\text{cell,fresh}}$ , with  $Q_{\text{cell,fresh}}$  and  $Q_{\text{cell,i}}$  representing the fresh cell capacity and the cell capacity at cycle  $i$ , respectively.

Despite the promising potential of this indicator for estimating capacity loss, it is important to highlight that the periodicity of the current profile may not hold true in real-driving conditions. Nevertheless, this study suggests that this indicator can be engineered in an offline setting, for example, as part of onboard diagnostics routines. In this work, the power autocorrelation function proves to be an effective SOH indicator given consistent usage of the UDDS discharge profile.

**Resistance.** Abrupt charge and discharge events, related to braking and acceleration maneuvers, respectively, offer the opportunity to evaluate the battery’s internal resistance<sup>27</sup>. As the battery ages, various factors such as electrode degradation, electrolyte breakdown, and formation of passivation layers contribute to an increase in its internal resistance,  $R$ . This increase limits the flow of ions within the battery, reducing conductivity and affecting the battery’s power output capability. As resistance increases, less power can be delivered to the motors due to higher Joule losses.

Demanding acceleration and braking events lead to changes in the battery current, referred to as current peaks. The resistance is calculated at each discharge current peak corresponding to an acceleration event over the discharging phase of the aging cycle, as described in Sec. Definition of SOH indicators. It is important to note that the battery’s internal resistance is influenced by factors such as SOC (see Supplementary Note 3),  $C$ -rate, and temperature. For accurate aging assessments in real-world scenarios, resistance should be measured under consistent conditions throughout the battery’s lifespan. In this work, temperature effects on this indicator are not studied since the cells are maintained in a controlled temperature environment.

A single resistance value is computed by averaging the resistances calculated during each discharge phase, which consists of multiple concatenated UDDS cycles between two charging phases, as further detailed in Sec. Cell cycling and experimental dataset, to minimize noise in the resistance, as shown in Fig. 2. This method effectively minimizes noise and variations in resistance measurements, offering a more consistent and representative value to assess battery health. Figure 2(a) highlights the importance of determining the average resistance. Despite the large standard deviation observed in the distribution of internal resistances for each discharge event, the average values, represented by green points, clearly exhibits an increasing trend as the battery ages.

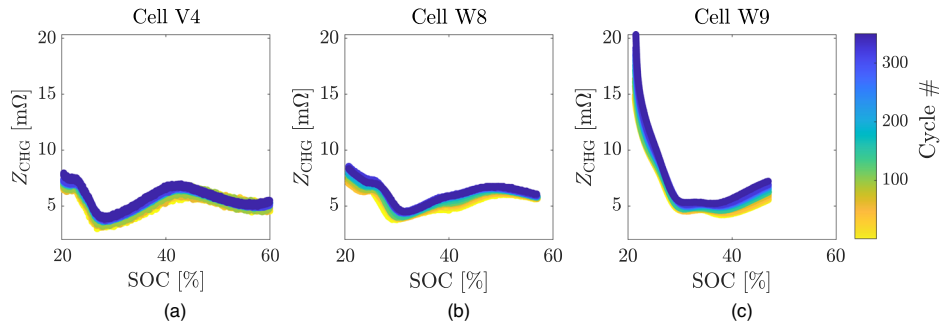


**Fig. 2 | Resistance.** **a** Internal resistance ( $R$ ) is plotted as a function of the current peak number during the discharge phases throughout the cell’s life span for cell W8. Shades of gray represent different batches of aging cycles. Batch  $j$  is defined as the period between the  $j$ th and the  $(j + 1)$ th reference performance test (RPT). Green points indicate the average internal resistance. **b** Percentage increase in internal resistance ( $R_{\text{increase}}$ ) relative to capacity loss ( $Q_{\text{cell,loss}}$ ) for all five cells.

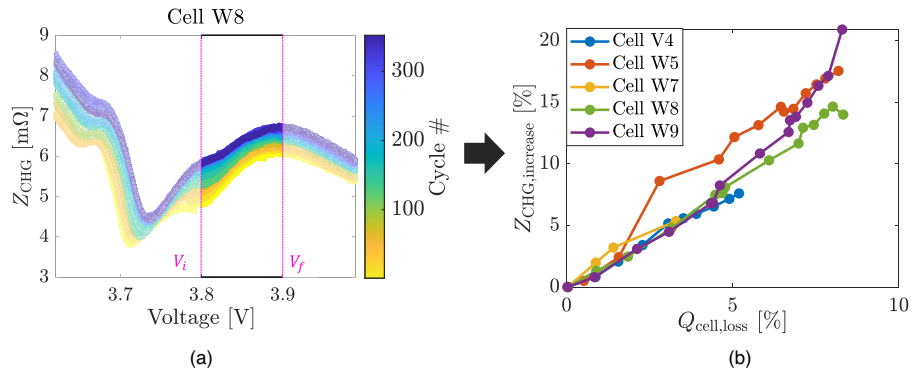
Additionally, Fig. 2b shows the percentage increase in average internal resistance for all five cells, correlated with their corresponding capacity losses. This increase is calculated as  $R_{\text{increase}} = (R_i - R_{\text{fresh}})/R_{\text{fresh}} \cdot 100$ , where  $R_{\text{fresh}}$  represents the average internal resistance measured during the first discharging phase of the cell, and  $R_i$  denotes the average internal resistance determined during the discharging at cycle  $i$ .

**Charging impedance.** The charging impedance<sup>27</sup>  $Z_{\text{CHG}}$  represents the battery’s resistance to the flow of electrons during charging. Variations in  $Z_{\text{CHG}}$  reflect how this resistance evolves as the battery ages. The  $Z_{\text{CHG}}$

**Fig. 3 | Charging impedance ( $Z_{CHG}$ ) as function of state of charge (SOC) and cycle number for cells V4 (a), W8 (b), and W9 (c). The yellow curve represents fresh cell conditions, while the dark blue curve denotes aged cell conditions.**



**Fig. 4 | Charging impedance. a** Charging Impedance ( $Z_{CHG}$ ) as a function of voltage for cell W8. Charging impedance is averaged over the voltage range [ $V_{in} = 3.8$  V,  $V_{fin} = 3.9$  V]. **b** Percentage variation of average  $Z_{CHG}$  ( $Z_{CHG, increase}$ ) as a function of capacity fade ( $Q_{cell, loss}$ ).



profiles for three cells (V4, W8, and W9), charged at different C-rates, are illustrated in Fig. 3 as a function of cell degradation and SOC.

The rising trend of  $Z_{CHG}$  over cells' lifetime aligns with the understanding that, as the battery ages, its overpotential increases due to factors such as the growth of the Solid-Electrolyte Interface, increased of contact resistance, and changes in reaction kinetics and transport dynamics<sup>6,52</sup>. Additionally, it is important to note that the  $Z_{CHG}$  profiles reach different SOC values at the end of charge (at 4 V). This phenomenon can be attributed to the varying polarization losses resulting from the different C-rates used during charging for cells V4, W8, and W9<sup>53</sup>. The charging impedance indicator is computed by averaging the impedance within the specific voltage range [ $V_{in} = 3.8$  V,  $V_{fin} = 3.9$  V], which is selected through the analysis reported in Supplementary Note 4. As shown in Fig. 4b, the increase in charging impedance ( $Z_{CHG, increase} = (Z_{CHG, i} - Z_{CHG, fresh}) / Z_{CHG, fresh} \cdot 100$ ) is highly correlated with capacity loss across all the battery cells. Therefore, the charging impedance  $Z_{CHG}$  can be used directly as a feature to correlate with capacity loss.

**Energy during charging.** The energy during charging indicator,  $E_{ch}$ , quantifies the energy stored in the battery during charging. This is computed by integrating the battery power within a specific voltage range [ $V_{in, ch}$ ,  $V_{fin, ch}$ ] (as detailed in Sec. Definition of SOH indicators). Figure 5a illustrates  $E_{ch}$  in relation to the charging duration required to reach  $V_{fin, ch}$  from  $V_{in, ch}$ . Figure 5b shows the energy during charging over the voltage range [ $V_{in, ch} = 3.6$  V,  $V_{fin, ch} = 3.9$  V] as a function of capacity loss. The y-axis of Fig. 5b quantifies the percentage energy loss during charging for each cell. Energy loss during charging for each cell is computed as  $E_{ch, loss} = (E_{ch, fresh} - E_{ch, i}) / E_{ch, fresh} \cdot 100$ , where  $E_{ch, fresh}$  is the energy for the fresh cell and  $E_{ch, i}$  is the amount of energy the battery is charged at during aging cycle  $i$  of the same cell. These results show that capacity loss is linearly correlated with energy loss during charging over the selected voltage range.

**Energy during discharging.** The energy during discharging indicator,  $E_{dis}$ , quantifies the energy delivered by the battery during its discharge

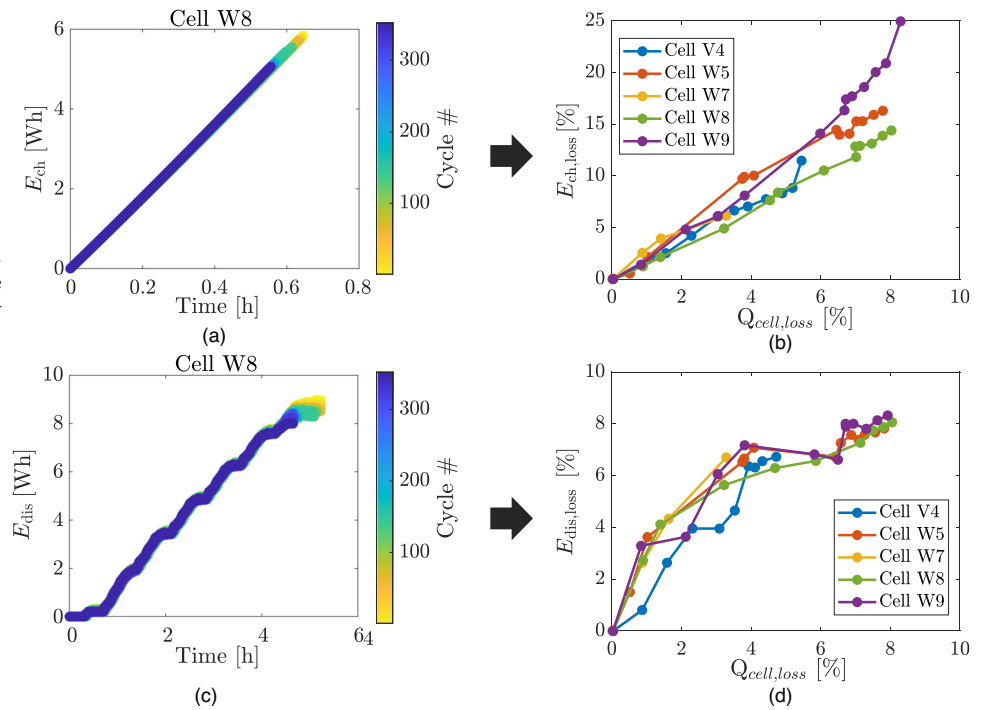
phase. This is computed by integrating the battery power over a specific voltage range [ $V_{in, dis}$ ,  $V_{fin, dis}$ ], as detailed in Sec. Definition of SOH indicators. Figure 5c illustrates  $E_{dis}$  in relation to the discharging duration needed to reach  $V_{fin, dis}$  from  $V_{in, dis}$ . Figure 5d displays the energy during discharging over the voltage range [ $V_{in, dis} = 3.85$  V,  $V_{fin, dis} = 3.4$  V] as a function of capacity loss. The y-axis of Fig. 5d quantifies the percentage energy loss during discharging for each cell. Energy loss during discharging for each cell is computed using  $E_{dis, loss} = (E_{dis, fresh} - E_{dis, i}) / E_{dis, fresh} \cdot 100$ , where  $E_{dis, fresh}$  represents the energy of a fresh cell, and  $E_{dis, i}$  is the energy charged during aging cycle  $i$  of the same cell. The results indicate a linear relationship between capacity loss and energy loss during discharging within the selected voltage range. In real EV scenarios, the variability in discharging rates complicates the consistent computation and monitoring of  $E_{dis}$ . A practical approach is to compare  $E_{dis}$  across driving scenarios with similar driving styles to account for this variability.

**SOH indicators regression analysis.** The health indicators are pre-processed according to the pipeline outlined in Supplementary Note 1. This process involves calculating incremental values for each feature:  $\Delta P_{Autocorr}$  (power autocorrelation),  $\Delta R_{ch}$  (resistance),  $\Delta Z_{CHG}^{NORM}$  (normalized charging impedance),  $\Delta E_{ch}$  (energy during charging), and  $\Delta E_{dis}$  (energy during discharging). These incremental values are derived by subtracting the initial feature value, measured during the first aging cycle, from the value at each subsequent aging cycle  $i$  throughout the cell's life cycle. Additional details are provided in Sec. Methods. In this work, we use features' incremental values to simplify the detection of aging trends. For each cell, we assess the correlation between its capacity loss and feature variations using Pearson's correlation coefficient  $r$ , defined as:

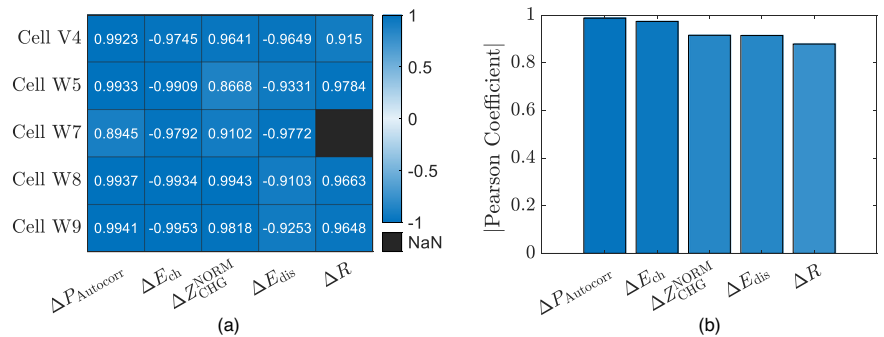
$$r = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Y_i - \bar{Y})^2}}, \quad (1)$$

where  $X_i$  represents the value of a specific incremental feature for a given cell at the  $i$ -th aging cycle,  $Y_i$  is the corresponding capacity loss value for

**Fig. 5 | Energy during charging and discharging.** **a** Energy during charging ( $E_{ch}$ ) as a function of charging time within the voltage range [ $V_{in,ch} = 3.6$  V,  $V_{fin,ch} = 3.9$  V] for cell W8 throughout its cycle life. **b** Energy loss during charging ( $E_{ch,loss}$ ) across all 5 cells shows a linear correlation with capacity loss ( $Q_{cell,loss}$ ). **c** Energy during discharging ( $E_{dis}$ ) as a function of discharging time within the voltage range [ $V_{in,dis} = 3.85$  V,  $V_{fin,dis} = 3.4$  V] for cell W8 throughout its cycle life. **d** Energy loss during discharging ( $E_{dis,loss}$ ) for all five cells demonstrates a linear correlation with capacity loss ( $Q_{cell,loss}$ ).



**Fig. 6 | Correlation analysis.** **a** Heatmap showing the Pearson's correlation coefficients between incremental power autocorrelation ( $\Delta P_{Autocorr}$ ), incremental energy during charging ( $\Delta E_{ch}$ ), incremental energy during discharging ( $\Delta E_{dis}$ ), normalized incremental charging impedance ( $\Delta Z_{CHG}^{NORM}$ ), and incremental resistance ( $\Delta R$ ) with capacity loss for each individual cell. **b** Histogram illustrating the Pearson's correlation coefficients between incremental features and capacity loss across all cell data. Note that the correlation between the capacity of cell W7 and  $\Delta R$  is not reported due to some computed resistances being deemed unreliable because of data acquisition issues (Sec. Cell cycling and experimental dataset and Supplementary Note 5), and thus interpreted as outliers during the pre-processing phase.



the same cell at that cycle,  $\bar{X}$  is the mean of the incremental feature values,  $\bar{Y}$  is the mean of the capacity loss values across all cycles, and  $N$  is the total number of data points (aging cycles analyzed). The results are shown in the heatmap of Fig. 6a. Each cell shows a high Pearson's correlation coefficient between capacity loss and each feature, underlying that the variations in these features are consistent indicators of aging across all the cells. However, since feature trends can vary across different cells, an additional analysis was conducted to identify features with more generalizable trends. We performed a correlation analysis between the extracted incremental features and capacity fade across all cells. This approach helps identify features that consistently reflect cell aging, regardless of individual cell differences. Figure 6b shows that some indicators generalize better across different cells.

We select features to train an estimation model according to two different cases. In the first case, Power autocorrelation ( $P_{Autocorr}$ ) is selected, as the sole feature, due to its superior overall performance. In the second case, we choose the best-performing feature for charging (energy during charging,  $E_{ch}$ ) and the best-performing feature for discharging (energy during discharging,  $E_{dis}$ ), excluding power autocorrelation.

The strong correlation between the extracted features and capacity fade can be attributed to the physical phenomena driving battery degradation. The linear relationship observed between charging impedance, resistance, and energy features with respect to charge throughput aligns with the linear trend of the capacity fade curve<sup>54</sup>. Given that the cells are cycled within a linear SOC window of 80% to 20% at ambient temperature, Solid-Electrolyte Interface layer growth is considered the dominant aging mechanism, leading to a linear capacity decrease trajectory. However, to thoroughly assess the aging modes present in the cells, a post-mortem analysis would be necessary.

**SOH estimation**

In this paper, we use capacity calculated at C/20 during RPTs as SOH metric. Additionally, for the purpose of training the machine learning models, the experimental C/20 capacity points are augmented using a linear data augmentation method as discussed in Sec. Data augmentation approach.

The features selected through the regression analysis are utilized to estimate capacity loss using a data-driven model. The performance of various models, namely, LRM, feed-forward neural networks, autoregressive

moving average with extra input, and recurrent neural networks, is compared using the same training and testing datasets, as detailed in Supplementary Note 6. Despite its simplicity, the LRM achieves estimation performance comparable to that of more complex models, owing to the strong linear correlation between the SOH indicators and capacity degradation. Therefore, the LRM is chosen for capacity loss estimation due to its lower computational time. Additionally, the LRM has the advantage of requiring fewer parameters to tune and fewer training samples compared to neural network-based models<sup>55</sup>. The LRM is trained using distinct sets of incremental SOH features: first with power autocorrelation, and then with energy during charging and energy during discharging (see Sec. SOH indicators regression analysis). Additionally, the estimation capabilities of the selected features are evaluated in two Scenarios. In Scenario 1 the LRM is trained exclusively on the data from cell W8 and tested on the other cells. In Scenario 2 the LRM is trained using data from all cells except the test cell. In the second Scenario, for cross-validation, the data is split into two subsets: one for the target cell and another for the remaining cells. The model is trained on the data from the remaining cells and tested on the data from the target cell.

Since the autocorrelation function of the power signal  $\Delta P_{\text{Autocorr}}$  exhibits the highest correlation with capacity fade, the data-driven model is initially trained using  $\Delta P_{\text{Autocorr}}$  as input. Figure 7 displays the capacity estimation results for both  $\Delta P_{\text{Autocorr}}$  and the energy-based features. In Scenario 1, the training dataset consists solely of data from cell W8, while in Scenario 2, it includes data from all cells except the test cell. The absolute percentage error (as defined in Sec. Methods) remains consistently below 1.5%, underscoring the relevant information provided by this individual feature. Moreover, using a more extensive set of training data from multiple cells (Scenario 2) does not improve estimation accuracy, leading to conclude that  $\Delta P_{\text{Autocorr}}$  is effective even with limited data. However, this feature has limitations in real-world scenarios and is better suited for offline diagnostics rather than online applications. It is also important to note that gaps in the observed capacity curves are due to voltage measurements anomalies, which resulted in unreliable feature values. This irregularity is attributed to unidentified equipment issues, as discussed in Sec. Cell cycling and experimental dataset and detailed further in Supplementary Note 5.

The LRM is subsequently trained using features that can be calculated during vehicle operation, specifically during driving and charging. The features selected for their high linear correlation with capacity during charging and discharging are energy during charging ( $\Delta E_{\text{ch}}$ ) and energy during discharging ( $\Delta E_{\text{dis}}$ ), respectively. As illustrated in Fig. 7, accurate capacity fade estimation is achieved with these features. Notably, when the LRM is trained using data from only cell W8 (Scenario 1), it achieves an absolute percentage error below 2.5% when tested on data from the other four cells. This result highlights the strong estimation capability of these features even with a limited dataset. For a more comprehensive analysis, the same estimation model is trained using data from multiple cells, leading to improved performance with the larger dataset. When using data from four cells for training (Scenario 2) and testing on the remaining cell, the absolute percentage error is below 1.6%. Notably, the estimation models perform well even for cells like W7 and W5, where some data is missing. This adaptability of the features and estimation models to partially available data is particularly advantageous in real-world scenarios, where acquiring complete EV battery data may not always be feasible. Moreover, to evaluate if adding extra features alongside the energy-based indicators could enhance model estimation capabilities, the LRM was also trained with incremental resistance and charging impedance included as additional inputs. However, the performance of the model with these additional features was worse than when using only energy during charging and discharging, as shown in Fig. 8. This indicates that the inclusion of resistance and charging impedance may introduce more noise than valuable information. It should be noted that for cell W7, only charging impedance is used as an additional feature, as the resistance data was compromised due to acquisition issues discussed in Sec. Methods. The superior performance of energy during charging and discharging as SOH indicators, compared to the increase in resistance or

charging impedance, can be attributed to several factors. Energy loss reflects not only resistance increases but also other factors such as heat generation, electrode degradation, and Solid-Electrolyte Interface formation, which impact overall energy efficiency. Additionally, the integration of the power signal offers a comprehensive measure of battery energy dynamics throughout an entire cycle, whereas resistance and charging impedance are computed over shorter time periods, making them more sensitive to short-term fluctuations.

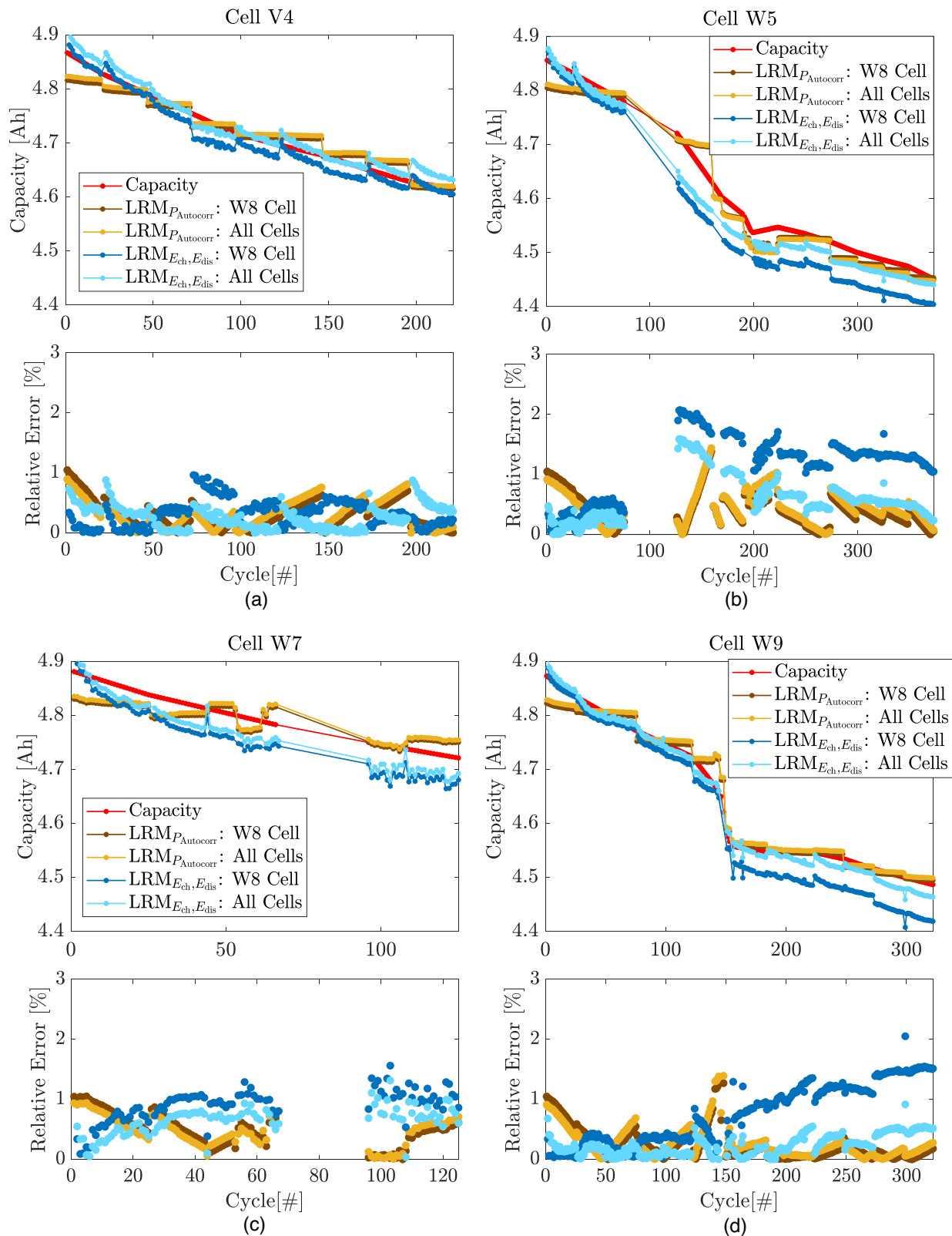
## Conclusions

This work extracts and evaluates five knowledge based SOH indicators, demonstrating their effectiveness as inputs to ML models for estimating capacity fade. The formulation of these indicators is guided by battery domain knowledge, allowing for the quantification of internal state variability due to battery degradation. Since none of the indicators rely on cumulative information (such as cycle number or Ah-throughput), they are suitable for real-world applications even with partial battery history. The high correlation between the indicators and capacity indicates that battery aging mechanisms leading to capacity fade are directly related to energy decrease and impedance rise. Two subsets of the engineered indicators, i.e., power autocorrelation, energy during charging, and energy during discharging, were utilized to train the estimation model for accurate cell capacity estimation. Due to their high correlation with capacity fade, combining energy during charging and energy during discharging as inputs results in accurate SOH estimation, with an absolute percentage error consistently below 2.5%. Conversely, power autocorrelation is the most informative feature, enabling precise capacity fade estimation with an absolute percentage error below 1.5%, even with limited training data. However, its effectiveness is influenced by the periodicity of discharging events. Consequently, power autocorrelation cannot be directly used as an SOH indicator in real-world driving scenarios but could be incorporated into a diagnostic tool by applying a periodic current signal to the battery when it is not in use. These findings suggest that domain knowledge-based features have the potential to be used as online tools for real-time capacity estimation. However, the model's effectiveness may be limited in practical applications. The dataset used in this study does not account for temperature variations or practical discharge events typical in real-world battery usage. Additionally, the current and voltage signals used to extract features have a high signal-to-noise ratio, which may not always be present in EV batteries. Having demonstrated the potential of these features on the studied dataset<sup>40</sup>, further investigations will be conducted using field data as future work. While this study primarily focuses on capacity estimation, utilizing a larger dataset could allow for the application of these indicators in RUL prediction. Extending the method proposed in this paper, these indicators could be integrated into forecasting models, enabling the BMS to anticipate and effectively manage battery capacity degradation.

## Methods

### Cell cycling and experimental dataset

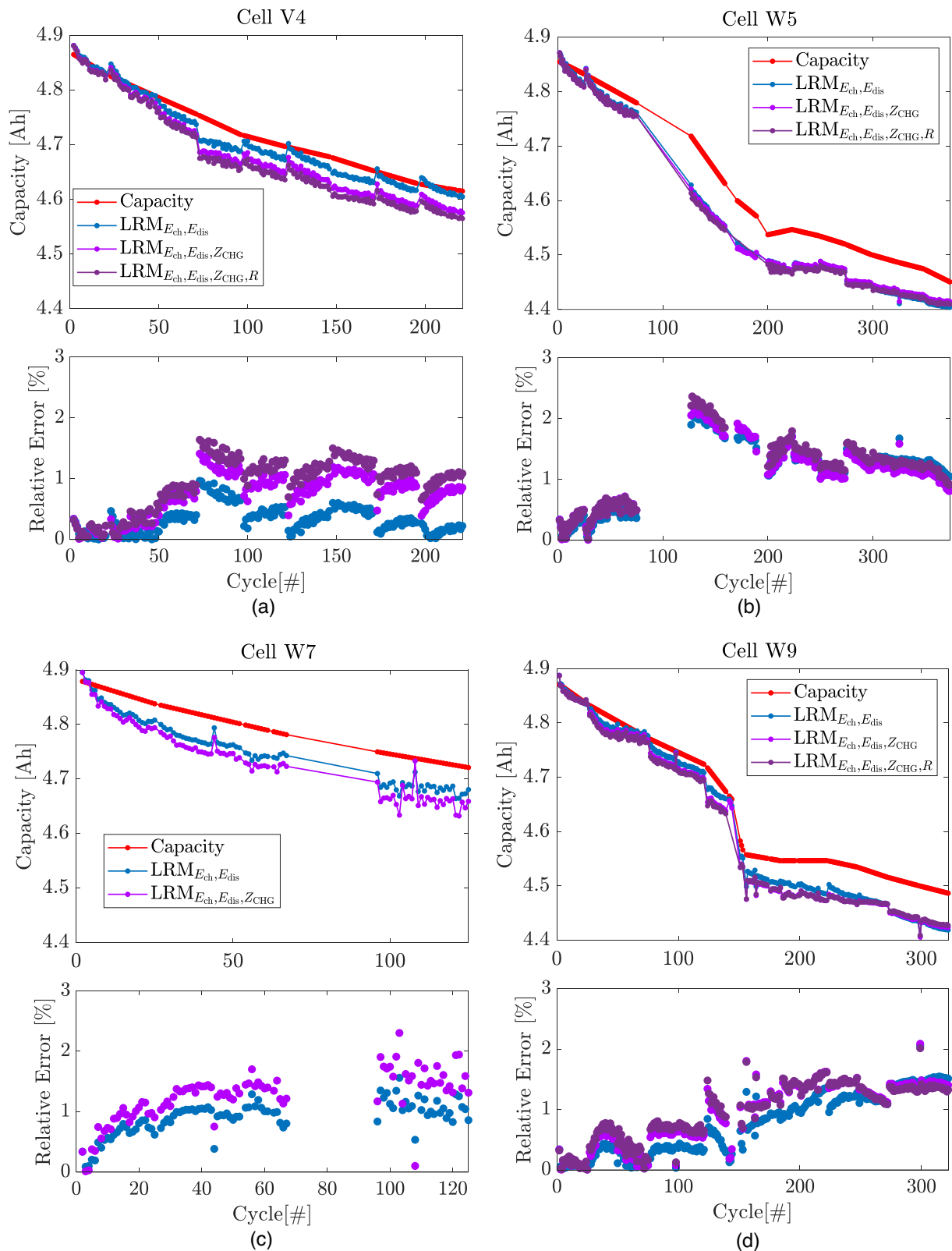
The experimental dataset<sup>40</sup> used in this work involves INR21700-M50T battery cells with graphite/silicon anode and nickel manganese cobalt oxides (NMC) cathode tested over a period of 30 months. For each cell, periodic RPTs, including C/20 capacity tests, Hybrid Pulse Power Characterization, and Electrochemical Impedance Spectroscopy, were conducted to assess the battery aging from fresh conditions. The cells underwent aging cycles as described in Supplementary Note 2. Each cycle includes a Constant Current-Constant Voltage (CC-CV) charge phase followed by a discharge phase. Specifically, there are two charge phases. Once the batteries reach 20% SOC (from the discharge phase), they are charged through the CC-a phase (at different C-rates) until reaching 4 V. They then continue charging at C/4 until 4.2 V, followed by the CV phase until the current drops below 50mA. The discharge phase, using concatenated UDDS driving profiles, simulates EV battery discharging, reducing the cell's SOC from 80% to 20%. Aging cycles conducted between the  $j^{\text{th}}$  and  $(j+1)^{\text{th}}$  RPTs for each cell are grouped into the  $j^{\text{th}}$  batch of aging cycles. Supplementary Note 2 details the



**Fig. 7 | SOH estimation results from the linear regression model (LRM) using power autocorrelation ( $LRM_{P_{Autocorr}}$ ), and energy during charging and discharging ( $LRM_{E_{ch},E_{dis}}$ ) as input features versus the aging cycle number (*Cycle*). Profiles of capacity loss and estimation error for cells V4 (a), W5 (b), W7 (c), and W9 (d). Augmented capacity points (obtained as discussed in Sec. Data augmentation approach) are shown in red. SOH estimation using power autocorrelation as input is shown in brown (with training data from cell W8) and yellow (with training data**

from all cells except the test cell). The dark blue and light blue lines show SOH estimation using energy features as input, with training data from cell W8 (dark blue) and from all available cells except the test cell (light blue). Gaps in the capacity curves for cells W5 (b) and W7 (c) are due to voltage measurements anomalies affecting the reliability of feature values (see Sec. Cell cycling and experimental dataset and Supplementary Note 5). The capacity drop for cell W8 (d) results from issues with the aging protocol implementation.





**Fig. 8 | SOH estimation results from the linear regression model (LRM) using charge and discharge energies, charging impedance, and resistance as features versus the aging cycle number (Cycle).** The capacity loss and estimation error profiles for cells V4 (a), W5 (b), W7 (c), and W9 (d) are shown. Augmented Capacity points (obtained as discussed in Sec. Data augmentation approach) are shown in red. Three scenarios are displayed: blue represents the LRM output trained solely with

energy during charging and energy during discharging ( $LRM_{E_{ch},E_{dis}}$ ); the light purple represents the LRM output trained with energy during charging, energy during discharging along with charging impedance ( $LRM_{E_{ch},E_{dis},Z_{CHG}}$ ); the dark purple represents the LRM output trained with energy during charging, energy during discharging, charging impedance, and resistance ( $LRM_{E_{ch},E_{dis},Z_{CHG},R}$ ). All models are trained exclusively using data from cell W8.

number of aging cycles in each batch for all cells used in this study. Among the ten cells (G1, V4, V5, W3, W4, W5, W7, W8, W9, W10) in the dataset, five (V4, W5, W7, W8, W9) are used in this study, as detailed in Table 1. The remaining cells were excluded for the following reasons. Cells W3, W10, and G1 were charged using a fast-charging 3C current profile during the CC-a phase, resulting in a very short charging duration interval that hindered feature extraction. Cell V5 was excluded due to insufficient aging, having undergone only 59 cycles with less than a 3% capacity decrease from the beginning of life. Cells W4, W5, and W7 were reported to have voltage measurements anomalies due to experimental issues, as noted in the “README” file of the dataset<sup>40</sup> and detailed in Supplementary Note 5. Specifically, cell W4 was affected for 310 cycles out of the total 760.

### Data augmentation approach

This work uses capacity to describe battery SOH. Given the limited number of RPTs, we have adopted an approach that uses data augmentation with linear interpolation for training purposes. For each cell, to assign a capacity value at every aging cycle  $i$  contained in batch  $j$ , we use the capacity values measured at the  $j$ -th and  $(j + 1)$ -th RPTs and estimate the capacity for cycle  $i$ ,  $Q_i$  as follows:

$$Q_i = \frac{i - \text{cycle}_j^{\text{RPT}}}{\text{cycle}_{j+1}^{\text{RPT}} - \text{cycle}_j^{\text{RPT}}} \times (Q_{j+1}^{\text{RPT}} - Q_j^{\text{RPT}}) + Q_j^{\text{RPT}} \quad (2)$$

where  $\text{cycle}_j^{\text{RPT}}$  and  $\text{cycle}_{j+1}^{\text{RPT}}$  denote the numbers of the aging cycle preceding the  $j$ -th and  $(j + 1)$ -th RPTs, respectively, while  $Q_j^{\text{RPT}}$  and  $Q_{j+1}^{\text{RPT}}$  represent the capacity values measured during these tests for the considered cell. Index  $i$  ranges from 1 to the number of aging cycles a cell has undergone (Table 1, fourth column), while index  $j$  ranges from 1 to the number of times the cell has been tested (Table 1, third column).

For example, capacity for cell V4 at aging cycle #30, namely  $Q_{30}^{V4}$ , is defined as:

$$Q_{30}^{V4} = \frac{30 - \text{cycle}_2^{\text{RPT},V4}}{\text{cycle}_3^{\text{RPT},V4} - \text{cycle}_2^{\text{RPT},V4}} \times (Q_3^{\text{RPT},V4} - Q_2^{\text{RPT},V4}) + Q_2^{\text{RPT},V4} \quad (3)$$

where  $\text{cycle}_2^{\text{RPT}} = 20$  and  $\text{cycle}_3^{\text{RPT}} = 45$ , since cell V4 has undergone 20 aging cycles before RPT #2 and 45 aging cycles before RPT #3.

### Definition of SOH indicators

$V_{\text{ch}}$  and  $V_{\text{dis}}$  represent the voltage profiles during charging and discharging, respectively.  $I_{\text{ch}}$  and  $I_{\text{dis}}$  are the current profiles during charging and discharging, respectively. Voltage variations due to acceleration peaks during discharging are indicated with  $\Delta V_{\text{acc}}$ , and the corresponding current variations with  $\Delta I_{\text{acc}}$ . The autocorrelation function measures the linear relationship between a signal  $x(t)$  and its time-delayed version  $x(t + \tau)$ , where  $\tau$  is the time delay. In this work, power autocorrelation during the discharge phase is quantified by correlating the power signal with its delayed copies. First, cell power is calculated from the voltage and current signals as follows:

$$P(t) = V_{\text{dis}}(t) \cdot I_{\text{dis}}(t) \quad (4)$$

The autocorrelation function of the power signal  $\hat{\rho}_\tau$  is computed with delays  $\tau$  limited to a range  $[-\tau_{\text{max}}, \tau_{\text{max}}]$ . In our study,  $\tau_{\text{max}}$  is set to 3000 s. For each value within this range,  $\hat{\rho}_\tau$  is computed as follows:

$$\hat{\rho}_\tau = \frac{1}{T} \sum_{t=\tau+1}^T (P(t) - \bar{P})(P(t - \tau) - \bar{P}) \quad (5)$$

where  $T$  is the duration of the discharging phase,  $P(t)$  is the power at time  $t$ ,  $\bar{P}$  is the average of the power over the time window  $T$ , and  $P(t - \tau)$  is the power

at instant  $t - \tau$ . The power autocorrelation indicator  $P_{\text{Autocorr}}$  is defined as the autocorrelation with null delay:  $P_{\text{Autocorr}} = \hat{\rho}_{\tau=0}$ .

The resistance  $R$  indicator is extracted for each aging cycle during the discharging phase using the following procedure. First, acceleration peaks are identified during the discharge<sup>27</sup> as explained in Supplementary Note 7. Then, the resistance  $R_{\text{peak}}$  corresponding to the  $l^{\text{th}}$  current peak within the  $i^{\text{th}}$  aging cycle is computed as follows:

$$R_{\text{peak},l}^i = \frac{\Delta V_{\text{acc},l}^i}{\Delta I_{\text{acc},l}^i} \quad (6)$$

where  $\Delta V_{\text{acc},l}^i$  and  $\Delta I_{\text{acc},l}^i$  are the voltage and current variations at the peak occurrence, respectively, as shown in Supplementary Note 7. Thus,  $P$  resistances  $R_{\text{peak},1}^i, R_{\text{peak},2}^i, \dots, R_{\text{peak},P}^i$  are computed for each  $i^{\text{th}}$  aging cycle, with  $i = 1, \dots, N$ , where  $N$  represents the number of aging cycles during the cell’s life and  $P$  is the total number of acceleration peaks within each cycle. Note that the number of total acceleration peaks,  $P$ , varies with the aging cycle. Subsequently, a single resistance value for each aging cycle is obtained by averaging the  $P$  resistances extracted from all acceleration peaks within that cycle:

$$R^i = \frac{\sum_{l=1}^P R_{\text{peak},l}^i}{P} \quad i = 1, 2, \dots, N \quad (7)$$

The instantaneous battery charging impedance  $Z_{\text{CHG}_{\text{ist}}}$  is computed over the CC-a phase<sup>27</sup> as follows:

$$Z_{\text{CHG}_{\text{ist}}}(t_k) = -\frac{V_{\text{ch}}(t_k) - V_{\text{ch}}(t_{k-1})}{I_{\text{ch}}} \quad (8)$$

where  $V_{\text{ch}}(t_k) - V_{\text{ch}}(t_{k-1})$  is the voltage difference over the interval  $\Delta t = t_k - t_{k-1}$ , and  $I_{\text{ch}}$  is the constant charging current during the CC-a phase.

The choice of the time window  $\Delta t$  is crucial. Increasing  $\Delta t$  helps filter out noise from the voltage difference in the numerator of Equation (9) and reduces current quantization effects. However, too large a window can excessively filter and result in information loss. Therefore,  $\Delta t$  is tuned to balance noise reduction while preserving the information content of  $Z_{\text{CHG}_{\text{ist}}}$ . The time intervals  $\Delta t$  are selected based on the C-rate:  $\Delta t = 60$  s for C/4,  $\Delta t = 30$  s for C/2, and  $\Delta t = 1$  s for 1C charging events.

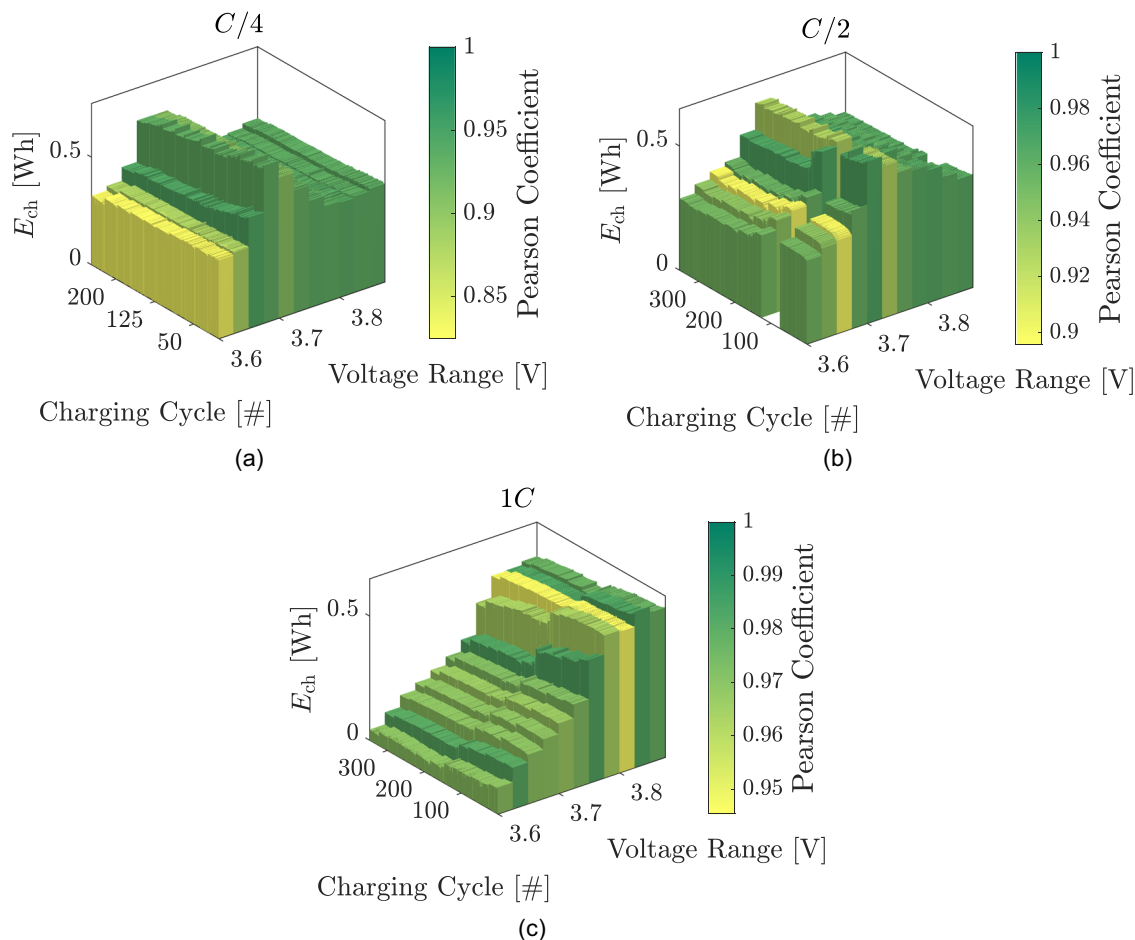
After extracting the instantaneous battery impedance for all the time intervals of the charging phase, the  $Z_{\text{CHG}}$  indicator is computed for each charging phase by averaging the  $Z_{\text{CHG}_{\text{ist}}}$  within a specific voltage range  $[V_{\text{in}}, V_{\text{fin}}]$ :

$$Z_{\text{CHG}} = \frac{1}{M} \sum_{t_k=t_{\text{in}}}^{t_{\text{fin}}} Z_{\text{CHG}_{\text{ist}}}(t_k) \quad (9)$$

where  $M$  is the number of  $Z_{\text{CHG}_{\text{ist}}}$  measurements within the considered voltage range, and  $t_{\text{in}}$  and  $t_{\text{fin}}$  are the initial and final time instants such that  $V(t_{\text{in}}) = V_{\text{in}}$  and  $V(t_{\text{fin}}) = V_{\text{fin}}$ , respectively. The voltages  $V_{\text{in}}$  and  $V_{\text{fin}}$  were set to 3.8 V and 3.9 V, respectively, based on the sensitivity analysis presented in Supplementary Note 4.

An alternative formulation would be to compute the average of  $Z_{\text{CHG}_{\text{ist}}}$  within a SOC range instead of a voltage range. However, we opted for the voltage-based formulation to avoid estimation errors affecting the SOC, which is a non-measurable quantity generally estimated by the BMS. Additionally, a different definition of charging impedance, discussed in Supplementary Note 8, has been excluded in the present work due to its lower correlation with capacity fade.

Finally, the energy during charging and discharging is computed on the CC-a charging segment (see Supplementary Note 2) and driving UDDS profile, respectively, by integrating the electrical power within a fixed voltage



**Fig. 9 | Impact of voltage range, C-rate and charging cycle number on energy during charging.** The amount of energy during charging ( $E_{ch}$ ) depends on the voltage range and the charging rate. As the charging rate increases (from  $C/4$  for cell

V4 (a), to  $C/2$  for cell W8 (b), to  $1C$  for W9 (c), the peak in charging energy shifts towards higher voltage ranges, i.e., [3.675 V - 3.7 V] at  $C/4$ , [3.7 V - 3.725 V] at  $C/2$  and [3.825 V - 3.85 V] at  $1C$ .

window, specifically  $[V_{in,ch}, V_{fin,ch}]$  and  $[V_{in,dis}, V_{fin,dis}]$ :

$$E_{ch} = \int_{t_{in}}^{t_{fin}} V_{ch}(t) \cdot I_{ch}(t) dt \tag{10}$$

$$E_{dis} = \int_{t_{in}}^{t_{fin}} V_{dis}(t) \cdot I_{dis}(t) dt \tag{11}$$

where  $V_{ch}$  is the cell voltage during charging,  $I_{ch}$  is the cell current during charging and  $t_{in}$  and  $t_{fin}$  are the initial and final time instants such that  $V_{ch}(t_{in}) = V_{in,ch}$  and  $V_{ch}(t_{fin}) = V_{fin,ch}$ . Similarly,  $V_{dis}$  is the cell voltage during discharging,  $I_{dis}$  is the cell current during discharging and  $t_{in}$  and  $t_{fin}$  are the initial and final time instants such that  $V_{dis}(t_{in}) = V_{in,dis}$  and  $V_{dis}(t_{fin}) = V_{fin,dis}$ . Thus, energy is not only a function of the C-rate but also depends on the voltage window over which it is calculated.

We selected the fixed voltage windows  $[V_{in,ch} = 3.6 \text{ V}, V_{fin,ch} = 3.9 \text{ V}]$  and  $[V_{in,dis} = 3.85 \text{ V}, V_{fin,dis} = 3.4 \text{ V}]$  for computing  $E_{ch}$  and  $E_{dis}$ , respectively, to bypass the initial and final stages of charging and discharging, which are potentially prone to noise.

**Sensitivity of charging energy to voltage window**

To assess the feasibility of using energy during charging for partial charging profiles, the correlation between  $E_{ch}$  and capacity loss was quantified across different voltage ranges. First, the interval  $[V_{in,ch}, V_{f,ch}]$  was divided into sub-intervals of 0.25 V amplitude, and the energy was computed for each sub-interval.

As shown in Fig. 9, there is a strong correlation between energy during charging and capacity loss across all voltage sub-intervals. These results show that energy can be effectively used to estimate the SOH for partial and narrow charging periods. The analysis indicates that the voltage interval with the highest correlation also depends on the charging rate. This insight facilitates straightforward integration into the BMS.

**Pre-processing and incremental indicators**

Data pre-processing is essential for effectively using SOH indicators in data-driven algorithms. A critical step is removing outliers—data points that deviate from the majority. Outliers can affect feature extraction and machine learning model performance. Therefore, a careful approach is used to remove outlier-containing data, ensuring more robust and reliable feature representation. A second step of the pre-processing phase is the computation of the incremental features, denoted by  $\Delta$ . This subsection explains how to obtain these features, using incremental resistances as an example.

For each cell in the dataset, the vector of incremental resistances  $\Delta \mathbf{R}$  is calculated as follows:

1. For each aging cycle  $i^{th}$ ,  $i = 1, \dots, N$ , the resistance during the discharge phase over acceleration peaks is calculated as a function of SOC.  $R^1$  represents the average resistance over the SOC range of 80% and 20% during discharge. The resulting resistance vector is:

$$\mathbf{R} = [R^1, R^2, \dots, R^N] \tag{12}$$

where  $R^1$  is the average fresh cell resistance and  $R^N$  is the average resistance at the last cycle.

2. Obtain the incremental resistance vector by subtracting  $R^1$  from each value in  $R$ .

$$\Delta R = R - R^1 \quad (13)$$

This approach ensures that the first element of the incremental vector for each feature is zero, facilitating the comparison of aging trends across cells. Additionally, the charging impedance vector  $Z_{\text{CHG}}$  requires further pre-processing due to its dependency on the C-rate (see, Fig. 3. this feature strongly depends on the C-rate at which it is computed. To standardize across different C-rates, the incremental vector  $\Delta Z_{\text{CHG}}$  is normalized using the fresh cell impedance value:

$$\Delta Z_{\text{CHG}}^{\text{NORM}} = \frac{\Delta Z_{\text{CHG}}}{Z_{\text{CHG}}^1} \quad (14)$$

where  $Z_{\text{CHG}}^1$  is the charging impedance calculated over the first aging cycle in Batch #1 in the voltage range [3.8 V - 3.9 V] as described in Sec. Definition of SOH indicators. Normalization reduces variations from different charging rates, providing a consistent feature representation. This pre-processing step is crucial for evaluating the ML model across cells cycled at various rates, effectively excluding C-rate as a training feature. It ensures a more refined data representation for machine learning algorithms.

### Estimation model

In this work, the LRM estimates capacity fade due to its strong linear correlation with SOH indicators. The LRM relates the response variable  $y$  to the input vector  $u$  as follows<sup>56</sup>:

$$y(t) = \beta_0 + \beta u(t) + \epsilon(t) \quad (15)$$

where  $\epsilon$  represents model error, capturing deviations between the model and observed data. Coefficients  $\beta$  are determined using the least-squares method, which minimizes the model error on the training dataset. To evaluate the accuracy of the estimation models, the root mean square error (RMSE) is calculated as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N e_i^2}{N}} \quad (16)$$

$$e_i = \frac{Q_{\text{cell},i} - Q_{\text{est},i}}{Q_{\text{cell},i}} \quad (17)$$

where  $e_i$  is the relative error, with  $Q_{\text{cell},i}$  and  $Q_{\text{est},i}$  representing the actual and estimated capacities at the cycle  $i$ , respectively. Additionally, the absolute percentage error is given by  $\text{APE}(\%) = |e_i| \cdot 100$ .

### Data availability

All experimental data<sup>40</sup> are available online at the following Open Science Framework repository: [OSF](#).

### Code availability

The code supporting the findings of this study is available at the following Open Science Framework repository: [OSF](#).

Received: 2 December 2023; Accepted: 23 October 2024;

Published online: 12 November 2024

### References

1. On Climate Change, I. P. Ipcc sixth assessment report <https://www.ipcc.ch/report/ar6/wg1/> (accessed August 2021).
2. Agency, I. E. Global ev outlook 2022 <https://www.iea.org/reports/global-ev-outlook-2022> (May 2022).
3. Li, M., Lu, J., Chen, Z. & Amine, K. 30 years of lithium-ion batteries. *Advanced Materials* **30**, 1800561 (2018).
4. Nuhic, A., Terzimehic, T., Soczka-Guth, T., Buchholz, M. & Dietmayer, K. Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods. *J. Power Sources* **239**, 680–688 (2013).
5. Rezvanizani, S. M., Liu, Z., Chen, Y. & Lee, J. Review and recent advances in battery health monitoring and prognostics technologies for electric vehicle (EV) safety and mobility. *J. Power Sources* **256**, 110–124 (2014).
6. Plett, G. L (2015) Battery management systems, Volume II: Equivalent-circuit methods. Artech House, Boston.
7. Plett, G. L. Dual and joint EKF for simultaneous SOC and SOH estimation. *21st Electric Vehicle Symposium (EVS21)* 1–12 (2005).
8. Zhang, F., Liu, G. & Fang, L. Battery state estimation using unscented kalman filter. In *2009 IEEE International Conference on Robotics and Automation*, 1863–1868 (IEEE, Kobe, Japan, 2009).
9. Taborelli, C. et al. Advanced battery management system design for soc/soh estimation for e-bikes applications. *Int. J. Powertrains* **5**, 325 (2016).
10. Chu, A., Allam, A., Cordoba Arenas, A., Rizzoni, G. & Onori, S. Stochastic capacity loss and remaining useful life models for lithium-ion batteries in plug-in hybrid electric vehicles. *J. Power Sources* **478**, 228991 (2020).
11. Chen, Z., Mi, C. C., Fu, Y., Xu, J. & Gong, X. Online battery state of health estimation based on genetic algorithm for electric and hybrid vehicle applications. *J. Power Sources* **240**, 184–192 (2013).
12. Miao, Q., Xie, L., Cui, H., Liang, W. & Pecht, M. Remaining useful life prediction of lithium-ion battery with unscented particle filter technique. *Microelectron. Reliab.* **53**, 805–810 (2013).
13. Xing, Y., Ma, E. W., Tsui, K.-L. & Pecht, M. An ensemble model for predicting the remaining useful performance of lithium-ion batteries. *Microelectron. Reliab.* **53**, 811–820 (2013).
14. He, W., Williard, N., Osterman, M. & Pecht, M. Prognostics of lithium-ion batteries based on dempster-shafer theory and the bayesian monte carlo method. *J. Power Sources* **196**, 10314–10321 (2011).
15. Li, J., Adewuyi, K., Lofti, N., Landers, R. G. & Park, J. A single particle model with chemical/mechanical degradation physics for lithium ion battery state of health (soh) estimation. *Appl. Energy* **212**, 1178–1190 (2018).
16. Moura, S. J., Chaturvedi, N. A. & Krstić, M. Adaptive partial differential equation observer for battery state-of-charge/state-of-health estimation via an electrochemical model. *J. Dyn. Syst. Meas. Control* **136**, 011015 (2014).
17. Santhanagopalan, S., Zhang, Q., Kumaresan, K. & White, R. E. Parameter estimation and life modeling of lithium-ion cells. *J. Electrochem. Soc.* **155**, A345 (2008).
18. Prada, E. et al. A simplified electrochemical and thermal aging model of lifepo4-graphite li-ion batteries: Power and capacity fade simulations. *J. Electrochem. Soc.* **160**, A616 (2013).
19. Allam, A. & Onori, S. Online capacity estimation for lithium-ion battery cells via an electrochemical model-based adaptive interconnected observer. *IEEE Trans. Control Syst. Technol.* **29**, 1636–1651 (2021).
20. Nagulapati, V. M. et al. Capacity estimation of batteries: Influence of training dataset size and diversity on data driven prognostic models. *Reliab. Eng. Syst. Saf.* **216**, 108048 (2021).
21. Vilsen, S. B. & Stroe, D.-I. Battery state-of-health modelling by multiple linear regression. *J. Clean. Prod.* **290**, 125700 (2021).
22. Lin, C. P., Cabrera, J., Yu, D. Y. W., Yang, F. & Tsui, K. L. SoH estimation and soc recalibration of lithium-ion battery with incremental capacity analysis; cubic smoothing spline. *J. Electrochem. Soc.* **167**, 090537 (2020).

23. Liu, D., Pang, J., Zhou, J., Peng, Y. & Pecht, M. Prognostics for state of health estimation of lithium-ion batteries based on combination gaussian process functional regression. *Microelectron. Reliab.* **53**, 832–839 (2013).
24. Yang, D., Zhang, X., Pan, R., Wang, Y. & Chen, Z. A novel gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. *J. Power Sources* **384**, 387–395 (2018).
25. Yang, D., Wang, Y., Pan, R., Chen, R. & Chen, Z. A neural network based state-of-health estimation of lithium-ion battery in electric vehicles. *Energy Procedia* **105**, 2059–2064 (2017).
26. Mansouri, S. S., Karvelis, P., Georgoulas, G. & Nikolakopoulos, G. Remaining useful battery life prediction for uavs based on machine learning. *IFAC-PapersOnLine* **50**, 4727–4732 (2017).
27. Pozzato, G. et al. Analysis and key findings from real-world electric vehicle field data. *Joule* **7**, 1–19 (2023).
28. Yang, H., Hong, J., Liang, F. & Xu, X. Machine learning-based state of health prediction for battery systems in real-world electric vehicles. *J. Energy Storage* **66**, 107426 (2023).
29. Hou, Y., Zhang, Z., Liu, P., Song, C. & Wang, Z. Research on a novel data-driven aging estimation method for battery systems in real-world electric vehicles. *Adv. Mech. Eng.* **13**, 16878140211027735 (2021).
30. Hong, J. et al. Online accurate state of health estimation for battery systems on real-world electric vehicles with variable driving conditions considered. *J. Clean. Prod.* **294**, 125814 (2021).
31. She, C., Wang, Z., Sun, F. & Zhang, L. Battery aging assessment for real-world electric buses based on incremental capacity analysis and radial basis function neural network. *IEEE Trans. Ind. Inform.* **16.5**, 3345–3354 (2019).
32. Song, L., Zhang, K., Liang, T., Han, X. & Zhang, Y. Intelligent state of health estimation for lithium-ion battery pack based on big data analysis. *J. Energy Storage* **32**, 101836 (2020).
33. He, Z. et al. State-of-health estimation based on real data of electric vehicles concerning user behavior. *J. Energy Storage* **41**, 102867 (2021).
34. Huo, Q., Ma, Z., Zhao, X., Zhang, T. & Zhang, Y. Bayesian network based state-of-health estimation for battery on electric vehicle application and its validation through real-world data. *IEEE Access* **9**, 11328–11341 (2021).
35. Zhou, Y., Huang, M., Chen, Y. & Tao, Y. A novel health indicator for on-line lithium-ion batteries remaining useful life prediction. *J. Power Sources* **321**, 1–10 (2016).
36. Wu, J., Wang, Y., Zhang, X. & Chen, Z. A novel state of health estimation method of li-ion battery using group method of data handling. *J. Power Sources* **327**, 457–464 (2016).
37. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4.5**, 383–391 (2019).
38. Bole, B., Kulkarni, C. S. & Daigle, M. Adaptation of an electrochemistry-based Li-ion battery model to account for deterioration observed under randomized use. *Annu. Conf. PHM Soc.* **6**, 1 (2014).
39. Birkel, C. Oxford battery degradation dataset 1 (2017).
40. Pozzato, G., Allam, A. & Onori, S. Lithium-ion battery aging dataset based on electric vehicle real-driving profiles. *Data Brief.* **41**, 107995 (2022).
41. Zhang, Y., Wik, T., Bergström, J., Pecht, M. & Zou, C. A machine learning-based framework for online prediction of battery ageing trajectory and lifetime using histogram data. *J. Power Sources* **526**, 231110 (2022).
42. Lyu, Z., Wang, G. & Tan, C. A novel bayesian multivariate linear regression model for online state-of-health estimation of lithium-ion battery using multiple health indicators. *Microelectron. Reliab.* **131**, 114500 (2022).
43. Shi, M., Xu, J., Lin, C. & Mei, X. A fast state-of-health estimation method using single linear feature for lithium-ion batteries. *Energy* **256**, 124652 (2022).
44. Chen, L., Lü, Z., Lin, W., Li, J. & Pan, H. A new state-of-health estimation method for lithium-ion batteries through the intrinsic relationship between ohmic internal resistance and capacity. *Measurement* **116**, 586–595 (2018).
45. Lin, M. et al. A data-driven approach for estimating state-of-health of lithium-ion batteries considering internal resistance. *Energy* **277**, 127675 (2023).
46. Paxton, W. et al. Battery management system for determining a health of a power source based on driving events (2024). US patent filed under application number US17/975.
47. Cui, Y. et al. State of health diagnosis model for lithium ion batteries based on real-time impedance and open circuit voltage parameters identification method. *Energy* **144**, 647–656 (2018).
48. Cai, L., Lin, J. & Liao, X. An estimation model for state of health of lithium-ion batteries using energy-based features. *J. Energy Storage* **46**, 103846 (2022).
49. Gong, D., Gao, Y., Kou, Y. & Wang, Y. State of health estimation for lithium-ion battery based on energy features. *Energy* **257**, 124812 (2022).
50. Peng, S. et al. State of health estimation of lithium-ion batteries based on multi-health features extraction and improved long short-term memory neural network. *Energy* **282**, 128956 (2023).
51. Khaleghi, S., Firouz, Y., Van Mierlo, J. & Van den Bossche, P. Developing a real-time data-driven battery health diagnosis method, using time and frequency domain condition indicators. *Appl. Energy* **255**, 113813 (2019).
52. Ovejas, V. J. & Cuadras, A. Effects of cycling on lithium-ion battery hysteresis and overvoltage. *Sci. Rep.* **9**, 14875 (2019).
53. Fly, A. & Chen, R. Rate dependency of incremental capacity analysis (dq/dv) as a diagnostic tool for lithium-ion batteries. *J. Energy Storage* **29**, 101329 (2020).
54. Yang, X.-G., Leng, Y., Zhang, G., Ge, S. & Wang, C.-Y. Modeling of lithium plating induced aging of lithium-ion batteries: Transition from linear to nonlinear aging. *J. Power Sources* **360**, 28–40 (2017).
55. Jiao, S., Gao, Y., Feng, J., Lei, T. & Yuan, X. Does deep learning always outperform simple linear regression in optical imaging? *Opt. Express* **28**, 3717 (2020).
56. Khuri, A. I. Introduction to linear regression analysis, fifth edition by douglas c. montgomery, elizabeth a. peck, g. geoffrey vining. *Int. Stat. Rev.* **81**, 318–319 (2013).

## Acknowledgements

The authors thank the members and alumni of the Stanford Energy Control Lab: Luca Pulvirenti, Sara Ha, Sai Thatipamula, Muhammad Aadil Khan and, in particular, Le Xu for their feedback on the manuscript. This work was partially funded by the Precourt Institute of Energy, Stanford University. This research is enabled in part through computational resources and support provided by Sherlock compute cluster of Stanford University.

## Author contributions

Conceptualization: A.A., G.P., and S.O.; data curation: A.L., G.P., and P.B.; formal analysis: A.L., M.A., and P.B.; funding acquisition: S.O.; investigation: A.L., G.P., P.B., and S.O.; methodology: A.L., G.P., M.A., and P.B., and S.O.; project administration: S.O.; resources: S.O.; supervision: S.O.; visualization: A.L., G.P., M.A., and P.B.; validation: A.L., M.A., and P.B.; writing—original draft: A.L., G.P., M.A., P.B., S.O.; writing—review and editing: A.L., M.A., P.B., S.O.

## Competing interests

The authors declare the following competing interests: A.A. is with Archer Aviation and G.P. is with Form Energy. They were both affiliated with Stanford University at the time of the research. All other authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s44172-024-00304-2>.

**Correspondence** and requests for materials should be addressed to Simona Onori.

**Peer review information** *Communications Engineering* thanks Yongzhi Zhang, Zhibin Zhao, and the other, anonymous, reviewer for their contribution to the peer review of this work. Primary Handling Editors: [Jiangong Zhu] and [Rosamund Daw and Saleem Denholme].

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024