Supplementary Information (SI) for Energy & Environmental Science.
This journal is © The Royal Society of Chemistry 2026

Supplementary Information

Inline quality grading of commercial lithium-ion battery manufacturing via data-efficient learning and transferable assessment

Chen Liang [1, 2, 8], Shengyu Tao [2, 3, 8, *], Chunqiu Xia [3], Xinghao Huang [2], Hang Hu [4], Rui Wang [5], Daoyi Dong [6], Ziyang Lyu [1, *], Guangmin Zhou [2, *], Huadong Mo [7, *]

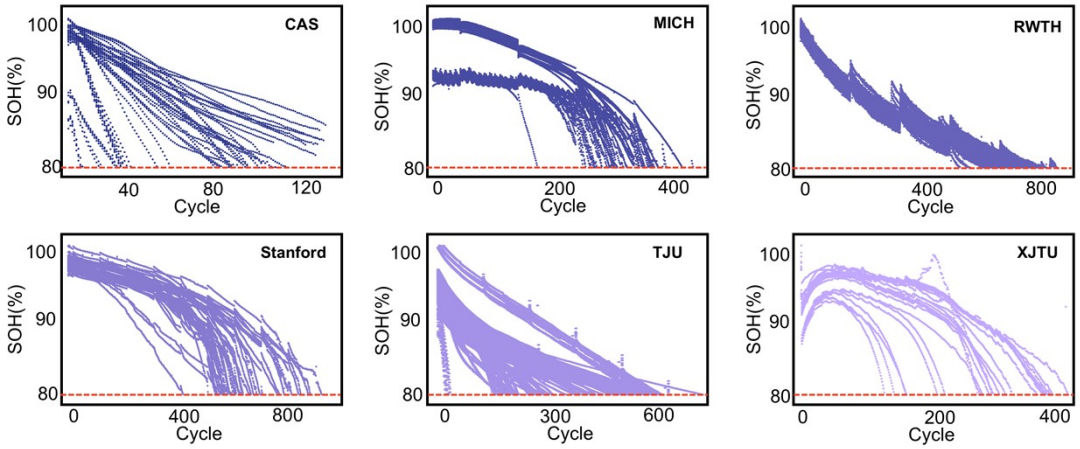
[1] School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China
[3] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, 41296, Sweden
[4] Department of Mechanical Engineering, Tsinghua University, Beijing, 100084, China
[5] Faculty of Mechanical Engineering and Mechanics, Ningbo University, Ningbo, 315211, China
[6] Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
[7] School of Systems and Computing, University of New South Wales, Canberra, ACT 2610, Australia
[8] These authors contributed equally to this article

* Corresponding authors: shengyu.tao@chalmers.se (S.T.), ziyang.lyu@unsw.edu.au (Z.L.), huadong.mo@unsw.edu.au (H.M.), guangminzhou@sz.tsinghua.edu.cn (G.Z.).

Contents

Supplementary Fig. 1. Degradation trajectories of the batch-based datasets(*b-datasets*).

This figure shows the SOH degradation trajectories of cells from different *b_datasets*. Each subplot corresponds to a specific dataset. The dashed red line indicates the EoL threshold at 80% SOH. Despite being tested under nominally similar operating conditions, the cells exhibit noticeable variability in degradation patterns and cycle lifetimes, reflecting the influence of batch-to-batch variations and stochastic ageing factors.

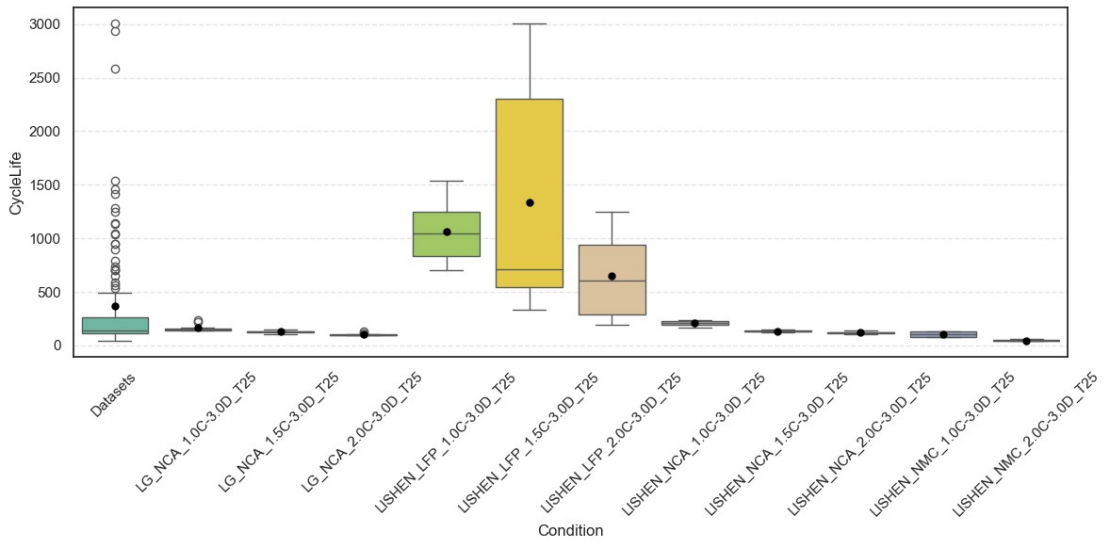


Supplementary Fig. 2. Degradation trajectories of the material-based datasets(*material_datasets*).

This figure presents the SOH degradation trajectories of cells grouped according to their material. Different colors represent datasets with distinct cathode materials and operating conditions. The curves demonstrate the diversity of degradation behaviors across different materials, including variations in early-stage capacity fade and long-term ageing rates.

The legend indicates the specific dataset corresponding to each ageing trajectory. Each label follows the format *Manufacturer_Material_Chargerate–Dischargerate_Temperature*, describing the testing conditions of the cells. For example, LG_NCA_1.0C-3.0D_T25 denotes cells manufactured by LG with an NCA cathode chemistry, tested under a 1C charge rate and 3C discharge rate at 25 °C. Similarly, LISHEN_LFP_1.5C-3.0D_T25 represents Lishen cells with an LFP cathode chemistry cycled under a 1.5C charge and 3C discharge rate at 25 °C.

Supplementary Fig. 3. Distribution of cycle life across the *material_dataset*



Supplementary Fig. 3 illustrates the distribution of cycle life for different materials and testing conditions in the *material_dataset*. Each boxplot represents the statistical distribution of EOL cycle life under a specific material-operating-condition combination, including variations in cathode chemistry (e.g., NCA, LFP and NMC) and charge/discharge rates. The aggregated dataset exhibits a wide range of cycle life values, spanning from fewer than 100 cycles to nearly 3000 cycles, indicating substantial heterogeneity across different materials and experimental conditions.

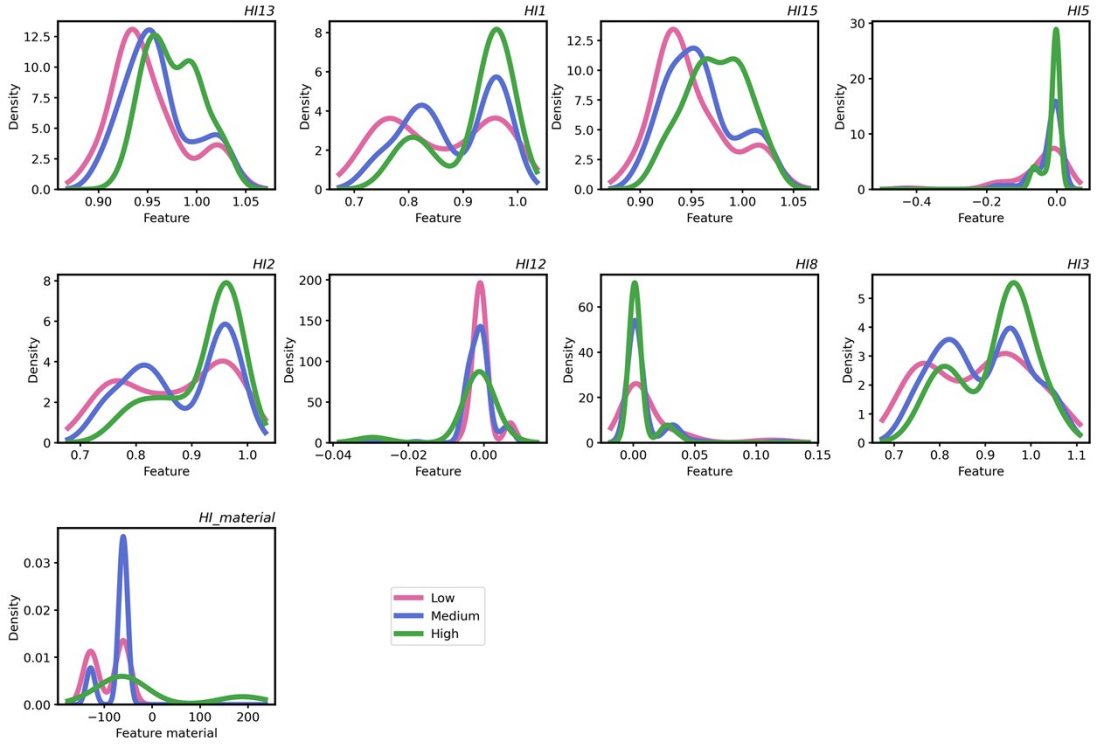
Significant differences can be observed among material groups. For example, cells based on LFP chemistry generally exhibit higher cycle life and larger variance, reflecting their well-known long-cycle stability but also indicating strong variability under different operating conditions. In contrast, NCA- and NMC-based cells tend to show comparatively shorter cycle life distributions, with narrower interquartile ranges in most conditions. These discrepancies highlight the strong influence of material and testing protocols on battery ageing behavior.

Moreover, even within the same material, noticeable dispersion exists across different C-rate combinations. This observation suggests that battery degradation is affected not only by material chemistry but also by operational factors such as charging rate, which can significantly alter electrochemical stress and ageing dynamics.

Overall, the pronounced heterogeneity observed in the *material_dataset* further motivates the use of the Linear Mixture Model (LME) modelling strategy within the Data-Efficient Learning and Transferable Assessment (DELTA) framework. By explicitly modelling material-dependent random effects, the framework is able to capture latent variability across different materials and operating conditions, thereby improving robustness and generalization when applied to heterogeneous datasets.

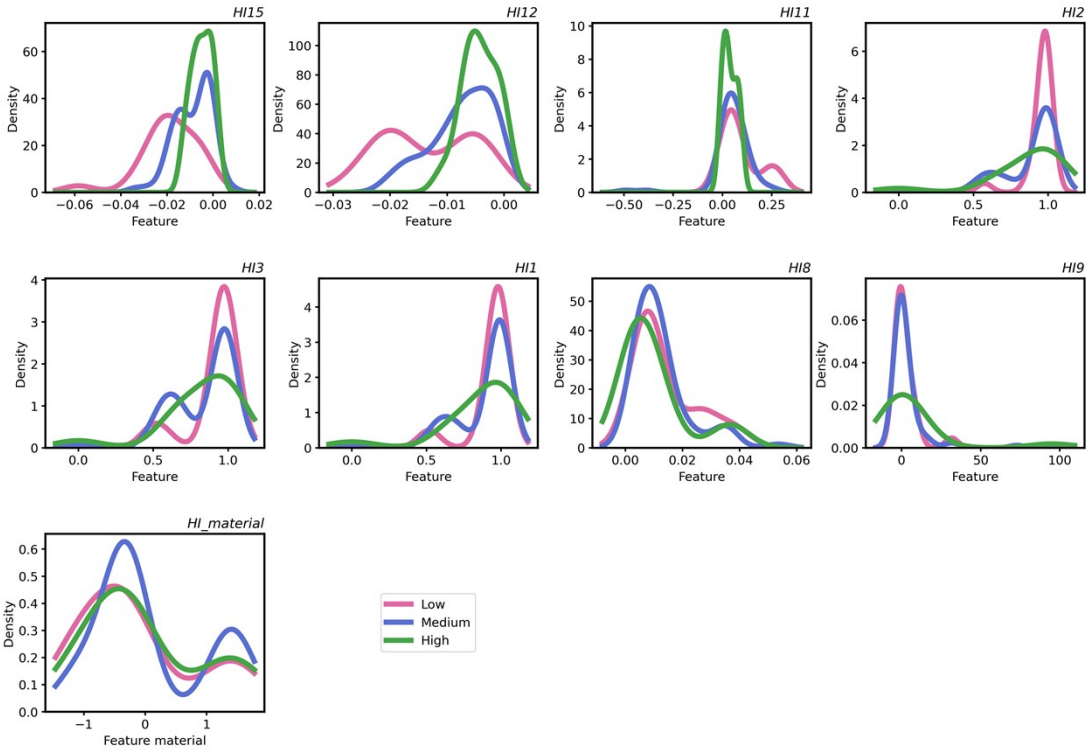
Supplementary Fig. 4. Bimodal distributions of extracted health indicators in *batch_dataset*.

This figure shows the health indicators (HIs) of *batch_dataset* we select from 16 HIs candidates using the coefficient of association and their Kernel density estimates. The curves correspond to 3 cycle-life categories (low, normal, and high). Several HIs exhibit bimodal or multimodal distributions.

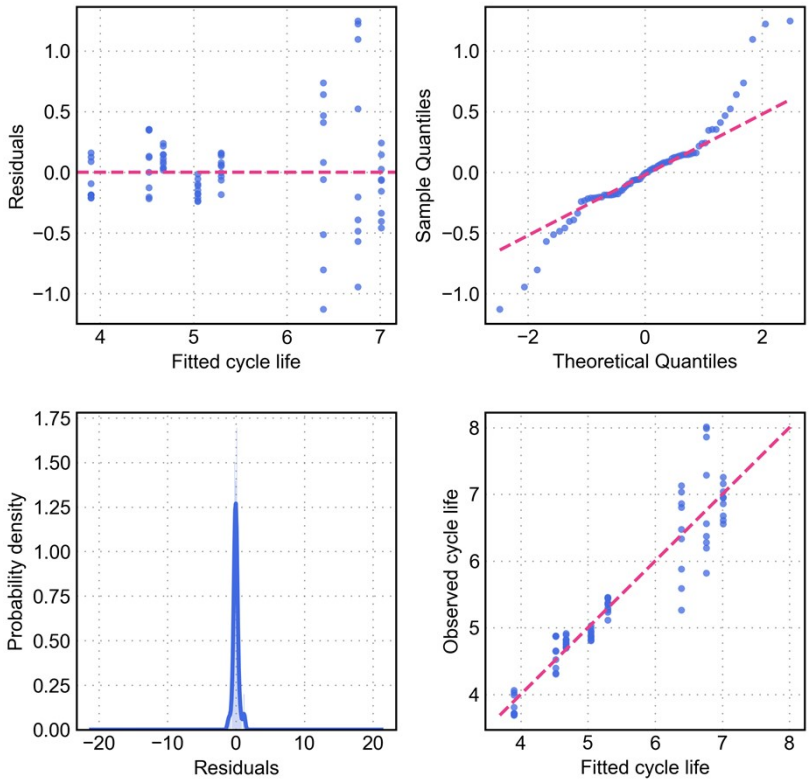


Supplementary Fig. 5. Bimodal distributions of extracted health indicators in *material_dataset*.

This figure shows the health indicators (HIs) of *material_dataset* we select from 16 HIs candidates using coefficient of association and their Kernel density estimates. The curves correspond to 3 cycle-life categories (low, normal, and high). Several HIs exhibit bimodal or multimodal distributions. The presence of bimodal feature distributions further justifies the use of K-means clustering before Gaussian mixture model classification, which helps separate underlying feature modes and enhances the robustness of the subsequent probabilistic classification.

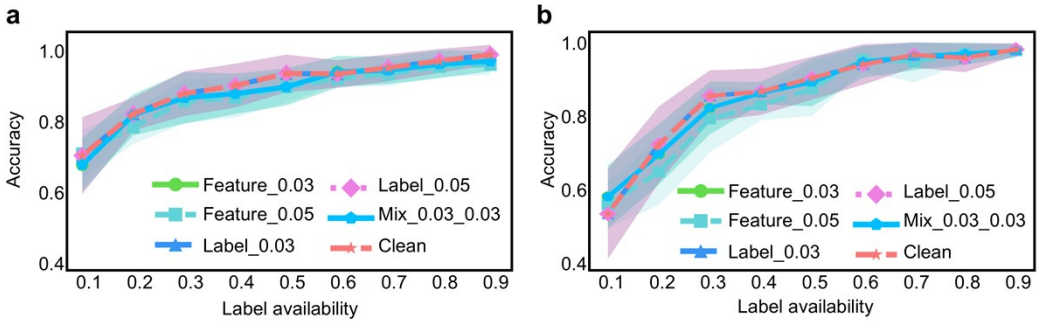


Supplementary Fig. 6. The diagnostic plots of the LME model for the *material_dataset*



Residual diagnostic plots are presented to assess the validity of the LME model. Residuals are randomly distributed around zero without clear patterns, indicating good model fit. The Q-Q plot shows that residuals approximately follow a normal distribution, with only minor deviations at the tails. The residual density is centered around zero, suggesting limited systematic bias. In addition, the observed versus fitted values exhibit a strong linear relationship close to the identity line, confirming that the model captures the major variation in cycle life. These results support the suitability of the LME model for extracting material-related features in the DELTA framework.

Supplementary Fig. 7. Accuracy under various noise conditions



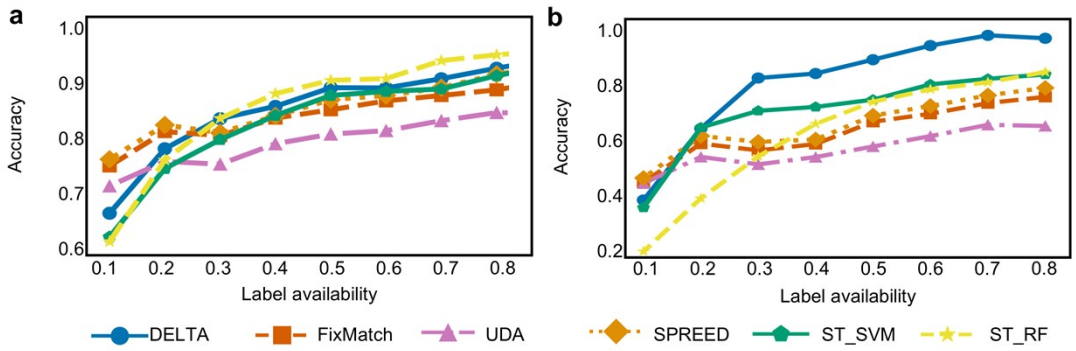
Supplementary Fig. 7 illustrates the classification accuracy of DELTA under different noise conditions as a function of label availability. Panel (a) corresponds to the *batch_dataset*, while panel (b) corresponds to the *material_dataset*. The evaluated noise settings include feature noise (Gaussian perturbations with standard deviations of 0.03 and 0.05), label noise (random label flipping with ratios of 0.03 and 0.05), a mixed noise scenario combining both feature and label perturbations, and a clean baseline without noise.

Overall, the classification accuracy consistently increases as label availability grows, indicating that the model effectively benefits from additional labeled information. Even when only a small proportion of labels are available (20%), the model already achieves moderate accuracy levels, demonstrating the capability of DELTA to exploit structural information in unlabeled data.

The results further show that feature noise has a slightly larger impact on performance than label noise, particularly at very low label availability levels. Nevertheless, the performance degradation remains relatively small across all noise configurations. As label availability increases beyond approximately 0.5, the accuracy curves under different noise settings gradually converge toward the clean baseline, indicating that the model becomes increasingly robust to noise when more labeled data are available.

A similar trend can be observed for the *material_dataset* in panel (b). Although the initial accuracy values are slightly lower due to higher dataset variability, the classification performance quickly improves as label availability increases. Under higher label availability levels (70%-90%), all noise scenarios achieve accuracy values close to the clean baseline, further demonstrating the robustness of DELTA under realistic noisy conditions.

Supplementary Fig. 8. Accuracy comparison with baseline methods



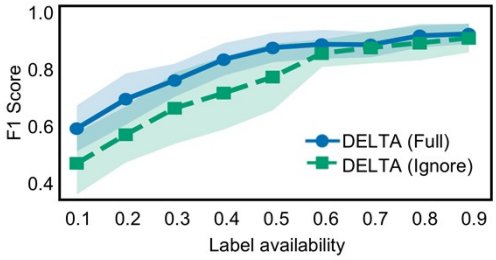
Supplementary Fig. 8 compares the classification accuracy of DELTA with several baseline methods under different label availability ratios. Panel (a) presents the results on the *batch_dataset*, while panel (b) shows the results on the *material_dataset*. The baseline methods include FixMatch, UDA, SPRED, ST_SVM, and ST_RF.

On the *batch_dataset* (panel a), all methods exhibit improved performance as label availability increases, reflecting the importance of labeled data in semi-supervised classification. Among the baseline methods, ST_RF and SPRED achieve relatively strong performance at moderate label availability levels. However, DELTA shows the fastest performance growth as label availability increases and consistently achieves higher accuracy once sufficient labeled samples are available.

On the *material_dataset* (panel b), the performance gap between DELTA and the baseline methods becomes even more pronounced. While several baseline methods exhibit unstable performance or relatively slow improvement as label availability increases, DELTA achieves significantly higher accuracy across most label availability levels. In particular, when label availability exceeds approximately 0.6, DELTA approaches near-perfect classification accuracy, whereas the best baseline methods remain substantially lower.

These results further confirm that DELTA provides superior scalability with respect to label availability and stronger robustness across different datasets, highlighting the effectiveness of the proposed framework in handling partially labeled battery datasets.

Supplementary Fig. 9. Ablation experiments on label missing mechanisms



Supplementary Fig. 9 presents the accuracy comparison between the complete DELTA framework ("Full") and the variant that ignores the label-missing mechanism ("Ignore") under different levels of label availability. As the proportion of available labels increases from 0.1 to 0.9, the prediction accuracy of both models improves steadily, indicating that additional labeled data consistently enhances model performance.

Across the entire range of label availability, the full DELTA framework consistently achieves higher accuracy than the variant that ignores the label-missing mechanism. The performance gap is particularly pronounced under extremely limited labeling conditions (for example, label availability below 0.3), where the full model maintains noticeably higher accuracy and exhibits reduced performance variability. These results demonstrate that explicitly modelling the label-missing mechanism enables more effective utilization of partially labeled data, thereby improving robustness and prediction accuracy under label-scarce conditions.

Supplementary Fig. 10. Ablation experiments on material features HI_m

Supplementary Fig. 10 shows the accuracy comparison between the DELTA framework incorporating the LME-derived material feature HI_m and a variant in which this feature is removed. As label availability increases, the prediction accuracy of both models improves gradually, reflecting the positive impact of increased supervision. However, the model incorporating HI_m consistently outperforms the variant without this feature across all levels of label availability. The improvement becomes more evident in the low-label regime, where the inclusion of HI_m leads to higher accuracy and more stable performance. This observation indicates that the random-effects modelling provided by the LME effectively captures latent inter-cell variability and material-induced heterogeneity in early-cycle data. By incorporating this uncertainty-aware feature, the DELTA framework achieves improved classification reliability and robustness.

Supplementary Table 1. Description of the *batch_dataset*.

This table summarizes the detailed descriptions of the *batch_dataset* used in this study. For each dataset, the corresponding cathode material, chemical formula, rated capacity, cut-off voltage range, operating temperature, charge/discharge rate, and the number of cells are provided. These datasets originate from multiple public sources, including CAS[1], MICH[2], RWTH[3], Stanford[4], TJU[5], and XJTU[6], covering different material systems and diverse operating conditions, most of which can be processed by BatteryLife[7].

Datasets	Material	Chemical Formula	Qn (Ah)	Cut-off Voltage (V)	Temperature (°C)	Charge/discharge rate (C)	N
CAS	NCA	-(LG)	2.35	2.65-4.2	25	1/3	10
		-(LISHEN)			25	1/3	10
		-(LISHEN)			0	0.5/3	10
	NCM	-(LISHEN)	2.6	2.75-4.2	25	1/3	9
		-(LISHEN)			25	2/3	10
MICH	NCM	$LiNi_{1/3}Co_{1/3}Mn_{1/3}O_2$	2.36	3.0-4.2	25	1/1	10
					45		10
					25		10
					45		10
RWTH	NCM	-	1.11	3.5-3.9	25	2/2	48
Stanford	NCM	$LiNi_{0.5}Mn_{0.3}Co_{0.2}O_2$	0.24	3-4.4	30	1/0.75	41
TJU	NCA	$Li_{0.86}Ni_{0.86}Co_{0.11}Al_{0.03}O_2$	3.5	2.65-4.2	25	0.25/1	7
					25	0.5/1	19
					25	1/1	9
					35	0.5/1	3
					45	0.5/1	28
	NCM	$Li_{0.86}Ni_{0.86}Co_{0.11}Mn_{0.07}O_2$	3.5	2.5-4.2	25	0.5/1	23
					35	0.5/1	4
					45	0.5/1	28
	NCA+NCM	-	-	-	25	0.5/1	3
					25	0.5/2	3
					25	0.5/4	3
XJTU	NCM	$LiNi_{0.5}Co_{0.2}Mn_{0.3}O_2$	2.0	2.5-4.2	20	2/1	8
						3/1	15

Supplementary Table 2. Description of the *material_dataset*.

This table summarizes the detailed descriptions of the *material_dataset* used in this study. Cells sharing the same material are grouped together and assigned the corresponding *m_label*. The material_dataset was constructed from the CAS dataset because it contains multiple battery chemistries, enabling material-level grouping and cross-material comparisons. For each material group, the material type, chemical formula, rated capacity, cut-off voltage range, operating temperature, charge/discharge rate, and number of cells are provided.

Datasets	Material	Chemical Formula	Qn (Ah)	Cut-off Voltage (V)	Temperature (°C)	Charge/discharge rate (C)	N
CAS	NCA	-(LG)	3.35	2.65-4.2	25	1C/3C	10
						1.5C/3C	10
						2C/3C	10
	NCA	-(LISHEN)	3.35	2.65-4.2		1C/3C	10
						1.5C/3C	10
						2C/3C	10
	NCM	-(LISHEN)	2.6	2.75-4.2		1C/3C	9
						2C/3C	10
	LFP	-	1.5	2.0-4.0		1C/3C	15
						1.5C/3C	30
						2C/3C	15

Supplementary Table 3. Description of the base and extended sub-datasets of the *batch_dataset*.

This table summarizes the detailed descriptions of the two sub-datasets constructed from the *batch_dataset*, including the base dataset and the extended dataset. The batch_dataset was assembled from all eligible publicly available datasets and subsequently divided into the base and extended datasets. For each dataset, the material system, chemical formula, rated capacity, cut-off voltage range, operating temperature, charge/discharge rate, and the number of cells are provided.

Type	Datasets	Material	Chemical Formula	Qn (Ah)	Cut-off Voltage (V)	Temperature (°C)	Charge/discharge rate (C)	Number
Base dataset	MICH	NCM	$LiNi_{1/3}Co_{1/3}Mn_{1/3}O_2$	2.36	3.0-4.2	25	1/1	10
						45		10
						25		10
						45		10
	RWTH	NCM	-	1.11	3.5-3.9	25	2/2	48
	Stanford	NCM	$LiNi_{0.5}Mn_{0.3}Co_{0.2}O_2$	0.24	3-4.4	30	1/0.75	41
	TJU	NCA	$Li_{0.86}Ni_{0.86}Co_{0.11}Al_{0.03}O_2$	3.5	2.65-4.2	25	0.25/1	7
						25	0.5/1	19
						25	1/1	9
						35	0.5/1	3
						45	0.5/1	28
		NCM	$Li_{0.86}Ni_{0.86}Co_{0.11}Mn_{0.07}O_2$	3.5	2.5-4.2	25	0.5/1	23
						35	0.5/1	4
						45	0.5/1	28
		NCA+ NCM	-	-	-	25	0.5/1	3
						25	0.5/2	3
						25	0.5/4	3
	XJTU	NCM	$LiNi_{0.5}Co_{0.2}Mn_{0.3}O_2$	2.0	2.5-4.2	20	2/1	8
							3/1	15
Extended dataset	CAS	NCA	-(LG)	2.35	2.65-4.2	25	1/3	10
			-(LISHEN)		2.65-4.2	25	1/3	10
			-(LISHEN)			0	0.5/3	10
		NCM	-(LISHEN)	2.6	2.75-4.2	25	1/3	9
			-(LISHEN)			25	2/3	10

Supplementary Table 4. Fifteen candidate HIs

Feature	Definition	Formula	Source cycles	Physical interpretation
HI1	Early capacity ratio	$HI_1 = \frac{Q_2}{Q_{nom}}$	Cycle 2	Early-stage capacity health
HI2	Mid-cycle capacity ratio	$HI_2 = \frac{Q_{\frac{2+n}{2}}}{Q_{nom}}$	Cycle (2+n)/2	Intermediate degradation level
HI3	Late early-cycle capacity ratio	$HI_3 = \frac{Q_n}{Q_{nom}}$	Cycle n	Early degradation magnitude
HI4	Maximum ΔQ(V)	$HI_4 = \frac{max(Q_{2(V)} - Q_{n(V)})}{Q_{nom}}$	Cycle 2 vs n	Maximum capacity difference along voltage curve
HI5	Minimum ΔQ(V)	$HI_5 = \frac{min(Q_{2(V)} - Q_{n(V)})}{Q_{nom}}$	Cycle 2 vs n	Minimum capacity difference
HI6	Mean ΔQ(V)	$HI_6 = \frac{mean(Q_{2(V)} - Q_{n(V)})}{Q_{nom}}$	Cycle 2 vs n	Average degradation across voltage
HI7	Variance of ΔQ(V)	$HI_7 = \frac{Var(Q_{2(V)} - Q_{n(V)})}{Q_{nom}^2}$	Cycle 2 vs n	Variability of degradation
HI8	Standard deviation of ΔQ(V)	$HI_8 = \frac{Std(Q_{2(V)} - Q_{n(V)})}{Q_{nom}}$	Cycle 2 vs n	Dispersion of capacity difference
HI9	Kurtosis of ΔQ(V)	$HI_9 = Kurtosis(Q_{2(V)} - Q_{n(V)})$	Cycle 2 vs n	Shape changes of Q-V curve
HI10	Skewness of ΔQ(V)	$HI_{10} = Skew(Q_{2(V)} - Q_{n(V)})$	Cycle 2 vs n	Asymmetry of degradation
HI11	Voltage slope of ΔQ(V)	$HI_{11} = \frac{d(Q_{2(V)} - Q_{n(V)})}{dV}$	Cycle 2 vs n	Polarization evolution
HI12	Capacity fade slope	$HI_{12} = \frac{dQ}{dN}$	All cycles	Rate of capacity degradation
HI13	Capacity intercept	$Q = aN + b$	All cycles	Estimated initial capacity
HI14	Regression goodness-of-fit	$HI_{14} = R^2$	All cycles	Linearity of degradation trend
HI15	Average capacity fade per cycle	$HI_{15} = mean(\Delta Q)$	All cycles	Mean degradation per cycle

This table summarizes the definitions of 15 candidate HIs extracted from early-cycle battery data. The indicators are derived from three main aspects of degradation behavior: capacity-based metrics, differential capacity–voltage curve features, and capacity fade trend characteristics. The table lists the definition, mathematical formulation, source cycles used for calculation, and the corresponding physical interpretation of each indicator.

Specifically, HI1–HI3 describe early-stage capacity retention across different cycle stages, reflecting the initial health condition and early degradation level of the cells. HI4–HI11 are extracted from the differences between capacity-voltage (Q-V) curves at early and later cycles, capturing structural changes in electrochemical behaviour, such as capacity loss distribution along the voltage range, polarization evolution, and statistical characteristics of degradation patterns. HI12–HI15 characterize the overall capacity fade dynamics across the cycling process through regression-based metrics and statistical measures, including degradation rate, linearity of the degradation trend, and average capacity loss per cycle.

Together, these candidate indicators provide complementary information about battery ageing behaviour and form the basis for subsequent feature selection and health indicator construction.

Supplementary Table 5. Material grouping of the *batch_dataset*

Material	Number	Group	Num_group
NCM	206	NCA	206
NCA	56	NCM	56
NCANCM	9	NCANCM	9

Supplementary Table 6. Charging-rate grouping of the *batch_dataset*

Charging rate	Number	Group	Num_group
0.25	7	verylow	7
0.5	112	low	112
1	81	medium	81
2	56	high	56
3	15	veryhigh	15

Supplementary Table 7. Temperature grouping of the *batch_dataset*

Temperature	Number	Group	Num_group
20	23	low	23
25	145	midium	145
30	41	high	47
35	6		
45	56	Veryhigh	56

Material, temperature, and charge rate are widely recognized as the three dominant factors governing battery ageing behavior, among which material exerts the most pronounced influence. Accordingly, in the LME model, material is treated as a random effect to capture material-dependent variability, while temperature and charge rate are considered fixed effects to quantify their systematic influence on degradation behavior.

Because these factors are discrete experimental variables, they were grouped into representative levels to facilitate statistical modelling. The grouping definitions for the *batch_dataset* are summarized in Supplementary Tables 5-7, corresponding to material type, charging rate, and temperature, respectively.

For the *material_dataset*, only cells cycled under room-temperature conditions were included. Therefore, temperature variation is not considered in this dataset. Consequently, the grouping scheme focuses on material type and charging rate, as summarized in Supplementary Tables 8-9.

Supplementary Table 8. Material grouping of the *material_dataset*

Material	Number	Group	Num_group
NCM	19	NCM	19
NCA	60	NCA	60
LFP	60	LFP	60

Supplementary Table 9. Charging-rate grouping of the *material_dataset*

Charging rate	Number	Group	Num_group
1	44	low	44
1.5	50	medium	50
2	35	high	35

Supplementary Table 10. Fixed-effects of the LME model for the *batch_dataset*.

This table summarizes the estimated fixed effects of the LME model for the *batch_dataset*. In this model, temperature and charging rate are treated as fixed effects, while the material system is modelled as a random effect. The intercept represents the expected EOL cycle life under the reference operating conditions, which correspond to the high temperature (highT) and high charging rate (highC) categories. The results show that both temperature and charging rate have statistically significant effects on cycle life ($p < 0.001$ for all estimated coefficients). Among the temperature-related effects, both lowT and midT show large negative coefficients relative to the reference category, indicating that these temperature groups are associated with shorter predicted cycle life under the adopted grouping scheme. Similarly, several charging-rate levels exhibit significant deviations from the reference condition. In particular, the lowC and verylowC groups show substantial negative coefficients, suggesting pronounced differences in degradation behaviour across charging-rate regimes. Overall, these results confirm that operating conditions strongly influence battery ageing, and that temperature and charging rate introduce systematic variations in cycle life across the dataset.

	Estimate	Std. Error	t-value	p-value
(Intercept)	1054.364551	82.13205133	12.83743112	***
lowT	-593.741737	32.1670511	-18.45807174	***
midT	-336.2625703	14.81582101	-22.69618201	***
veryhighT	-160.9742509	19.66361168	-8.186403065	***
highT	0	0	0	
lowC	-451.3203597	16.80768646	-26.85202159	***
mediumC	-307.4562084	15.5470799	-19.77581709	***
veryhighC	-152.9083333	32.73254768	-4.671446134	***
verylowC	-363.6384853	33.64721292	-10.80738801	***
highC	0	0	0	

Note: highT and highC are the reference categories. Standard errors are reported in the second column. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Supplementary Table 11. Random-effects of LME model for the *batch_dataset*.

This table presents the estimated random effects associated with different material systems in the LME model for the batch_dataset. The reported values correspond to the estimated best linear unbiased predictors (EBLUPs) of the random-effect terms. These coefficients represent the material-dependent deviations from the global mean cycle life after accounting for the fixed effects of temperature and charging rate. The results show that the NCANCM material system exhibits a large positive deviation, indicating that its baseline cycle life is substantially higher than the dataset average under comparable operating conditions. In contrast, both NCA and NCM show negative deviations, implying relatively shorter baseline cycle lifetimes compared with the global mean. These differences highlight the systematic influence of cathode material chemistry on battery degradation behaviour. The estimated EBLUPs are subsequently used as a static feature (HI_m) to capture the material-dependent ageing contribution in the downstream classification framework.

Material	EBLUPs
NCA	-128.1777815
NCANCM	188.425596
NCM	-60.24781448

Supplementary Table 12. Fixed-effects of the LME model for the *material_dataset*

This table reports the fixed-effect estimates of the LME model for the *material_dataset*. Because this dataset only contains cells tested under room-temperature conditions, temperature is not included as a factor in the model. Consequently, the fixed effects only account for charging-rate variations. The intercept represents the expected cycle life under the reference charging-rate category (highC). The coefficients for lowC and mediumC are positive and statistically significant, indicating that these charging-rate levels are associated with longer predicted cycle life compared with the high charging-rate condition. This result is consistent with well-established electrochemical ageing mechanisms, as higher charging rates typically accelerate degradation processes such as polarization increase, lithium plating, and structural stress in electrode materials.

	Estimate	Std. Error	t-value	p-value
(Intercept)	4.990191199	0.607357333	8.216236026	***
lowC	0.621381485	0.105885844	5.868409408	***
mediumC	0.370743396	0.124069312	2.988195801	**
highC	0	0	0	

Note: highC is the reference category. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Supplementary Table 13. Random effects of the LME model for the *material_dataset*

This table summarizes the random-effect estimates (EBLUPs) associated with different material systems in the material_dataset. These coefficients quantify the material-dependent deviations in cycle life after controlling for charging-rate effects. The results indicate that the LFP material system exhibits a positive deviation, suggesting a longer baseline cycle life relative to the dataset average. In contrast, NCM shows a negative deviation, indicating relatively shorter lifetimes under similar charging conditions. The NCA system exhibits a smaller negative deviation compared with NCM. These results further confirm that material chemistry introduces intrinsic variability in battery ageing behaviour, even when operating conditions are controlled. Incorporating these material-dependent effects as static features enables the subsequent modelling framework to better capture systematic ageing differences across material systems.

Material	EBLUPs
LFP	1.402677323
NCA	-0.314301511
NCM	-1.088375812

Supplementary Table 14. Performance of DELTA for *batch_dataset*.

This table summarizes the quantitative results corresponding to **Fig. 3a**, reporting the classification accuracy and F1 score of the DELTA framework under different label availability ratios for the *batch_dataset*.

As the proportion of labeled samples increases from 0.1 to 1.0, both accuracy and F1 score exhibit a clear upward trend. When only 10% of the labels are available, the classification accuracy is 0.69, indicating that the model still maintains a reasonable predictive capability even under extremely limited supervision. As the label availability increases to 0.3-0.4, the accuracy improves substantially to approximately 0.83-0.85, reflecting the increasing effectiveness of the semi-supervised learning mechanism.

When the label availability reaches 0.5, the accuracy rises to 0.87, demonstrating that the model can achieve relatively reliable classification performance once half of the samples are labeled. Beyond this point, the improvement becomes more gradual, with accuracy stabilizing between 0.87 and 0.92 for label availability levels between 0.6 and 1.0. The close agreement between accuracy and F1 score across all label visibility levels further indicates balanced classification performance across different classes without significant class imbalance effects.

Overall, these results demonstrate that DELTA exhibits strong robustness to limited label availability, maintaining stable and high performance once a moderate proportion of labeled data is available.

Label availability	Accuracy	F1 score
1.0	0.923898	0.923461
0.9	0.913931	0.913734
0.8	0.900311	0.900092
0.7	0.884889	0.885137
0.6	0.871334	0.871287
0.5	0.872260	0.872289
0.4	0.845151	0.843830
0.3	0.827076	0.826221
0.2	0.781662	0.782647
0.1	0.689156	0.687769

Supplementary Table 15. Performance of DELTA for *material_dataset*.

reports the classification performance of DELTA under varying label availability ratios for the *material_dataset*, corresponding to **Fig. 3b**.

A similar monotonic improvement trend can be observed as the label availability increases. When only 10% of labels are available, the classification accuracy is 0.59, which is slightly lower than that of the batch_dataset but still demonstrates meaningful predictive capability under scarce supervision. As the label availability increases to 0.2 and 0.3, the accuracy improves rapidly to 0.74 and 0.83, respectively, indicating that additional labeled samples significantly enhance the model's ability to distinguish between cycle-life classes.

Once label availability reaches 0.5, the classification accuracy already exceeds 0.92 and continues to improve steadily as label visibility increases. Under near fully labeled conditions (0.9-1.0), the accuracy approaches 0.99, suggesting that the DELTA framework is capable of achieving near-perfect classification performance when sufficient labeled data are available.

The F1scores closely follow the same trend as the accuracy values, confirming consistent predictive performance across all classes. Compared with the *batch_dataset*, the *material_dataset* exhibits slightly higher classification accuracy across most label availability levels, which may be attributed to reduced variability in experimental conditions when grouped by material systems.

These results further confirm the scalability and robustness of the DELTA framework, demonstrating that its classification performance improves consistently with increasing label availability while remaining effective even in low-label scenarios.

Label availability	accuracy	F1 score
1.0	0.985582	0.985559
0.9	0.989153	0.989131
0.8	0.974339	0.974293
0.7	0.963624	0.962929
0.6	0.956349	0.956151
0.5	0.923942	0.924655
0.4	0.888228	0.88662
0.3	0.833069	0.832253
0.2	0.739683	0.73894
0.1	0.586376	0.57583

Supplementary Table 16. Robustness of DELTA under noisy conditions for the *batch_dataset*

Supplementary Table 16-1. Accuracy

Noise\Label availability	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
clean	0.914 ± 0.022	0.900 ± 0.026	0.885 ± 0.030	0.871 ± 0.031	0.872 ± 0.042	0.845 ± 0.049	0.827 ± 0.050	0.782 ± 0.037	0.689 ± 0.086
feat_0.03	0.899 ± 0.022	0.893 ± 0.031	0.879 ± 0.030	0.876 ± 0.035	0.842 ± 0.041	0.827 ± 0.045	0.818 ± 0.058	0.780 ± 0.046	0.669 ± 0.059
feat_0.05	0.892 ± 0.023	0.888 ± 0.023	0.871 ± 0.027	0.861 ± 0.013	0.842 ± 0.036	0.818 ± 0.043	0.810 ± 0.049	0.753 ± 0.037	0.693 ± 0.055
label_0.03	0.914 ± 0.022	0.900 ± 0.026	0.885 ± 0.030	0.871 ± 0.031	0.872 ± 0.042	0.845 ± 0.049	0.827 ± 0.050	0.782 ± 0.037	0.689 ± 0.086
label_0.05	0.914 ± 0.022	0.900 ± 0.026	0.885 ± 0.030	0.871 ± 0.031	0.872 ± 0.042	0.845 ± 0.049	0.827 ± 0.050	0.782 ± 0.037	0.689 ± 0.086
mix_0.03_0.03	0.899 ± 0.022	0.893 ± 0.031	0.879 ± 0.030	0.876 ± 0.035	0.842 ± 0.041	0.827 ± 0.045	0.818 ± 0.058	0.780 ± 0.046	0.669 ± 0.059

Supplementary Table 16-2. F1 Score

Noise\Label availability	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
clean	0.914 ± 0.022	0.900 ± 0.026	0.885 ± 0.029	0.871 ± 0.031	0.872 ± 0.042	0.844 ± 0.050	0.826 ± 0.053	0.783 ± 0.038	0.688 ± 0.088
feat_0.03	0.899 ± 0.023	0.893 ± 0.031	0.879 ± 0.030	0.875 ± 0.036	0.840 ± 0.043	0.825 ± 0.047	0.816 ± 0.060	0.781 ± 0.047	0.669 ± 0.058
feat_0.05	0.891 ± 0.023	0.887 ± 0.023	0.871 ± 0.028	0.861 ± 0.013	0.841 ± 0.037	0.818 ± 0.045	0.808 ± 0.051	0.754 ± 0.036	0.693 ± 0.055
label_0.03	0.914 ± 0.022	0.900 ± 0.026	0.885 ± 0.029	0.871 ± 0.031	0.872 ± 0.042	0.844 ± 0.050	0.826 ± 0.053	0.783 ± 0.038	0.688 ± 0.088
label_0.05	0.914 ± 0.022	0.900 ± 0.026	0.885 ± 0.029	0.871 ± 0.031	0.872 ± 0.042	0.844 ± 0.050	0.826 ± 0.053	0.783 ± 0.038	0.688 ± 0.088
mix_0.03_0.03	0.899 ± 0.023	0.893 ± 0.031	0.879 ± 0.030	0.875 ± 0.036	0.840 ± 0.043	0.825 ± 0.047	0.816 ± 0.060	0.781 ± 0.047	0.669 ± 0.058

This table summarizes the classification performance of the DELTA framework on the *batch_dataset* under different noise settings and label availability levels. Both accuracy and F1 score are reported as mean ± standard deviation over repeated experiments. The table includes several scenarios: a clean baseline without noise, feature noise with Gaussian perturbations (standard deviation = 0.03 or 0.05), label noise where a proportion of training labels is randomly flipped (0.03 or 0.05), and a mixed setting combining feature and label noise.

Under the clean setting, the model achieves high performance across all label availability levels. When the label availability is 0.1, the accuracy already reaches 0.914, indicating strong predictive capability even under extremely limited supervision. As the label availability decreases toward 0.9 (i.e., fewer labeled samples), the performance gradually declines, reaching 0.689, reflecting the increasing difficulty of classification when labeled information becomes scarce.

When feature noise is introduced, the classification performance shows only a moderate decrease compared with the clean baseline. For instance, under feat_0.03, the accuracy at label availability 0.1 remains close to 0.90, demonstrating that the model is relatively insensitive to small perturbations in input features. Even with stronger feature noise (feat_0.05), the performance degradation remains limited, indicating that the feature extraction and clustering steps effectively mitigate noise influence.

In contrast, label noise produces almost identical results to the clean setting in this experiment. This behaviour suggests that the semi-supervised structure of DELTA can effectively tolerate small levels of label corruption by leveraging the structure of the unlabeled data.

The mixed-noise scenario, which combines feature and label perturbations, yields results similar to those of the feature-noise-only case. This indicates that the model's performance degradation is primarily driven by feature perturbations rather than label noise. Overall, the results confirm that the DELTA framework maintains stable classification accuracy and F1 score across a range of noisy conditions, demonstrating strong robustness to both measurement noise and annotation errors.

Supplementary Table 17. Robustness of DELTA under noisy conditions for the *material_dataset*

Supplementary Table 17-1. Accuracy

Noise\Label availability	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
clean	0.989 ± 0.017	0.967 ± 0.038	0.975 ± 0.033	0.949 ± 0.054	0.913 ± 0.057	0.877 ± 0.062	0.866 ± 0.068	0.737 ± 0.101	0.552 ± 0.120
feat_0.03	0.986 ± 0.018	0.978 ± 0.029	0.971 ± 0.036	0.956 ± 0.042	0.902 ± 0.064	0.873 ± 0.030	0.834 ± 0.070	0.710 ± 0.080	0.597 ± 0.084
feat_0.05	0.986 ± 0.018	0.978 ± 0.024	0.956 ± 0.054	0.964 ± 0.033	0.884 ± 0.072	0.841 ± 0.040	0.804 ± 0.087	0.663 ± 0.087	0.576 ± 0.062
label_0.03	0.989 ± 0.017	0.967 ± 0.038	0.975 ± 0.033	0.949 ± 0.054	0.913 ± 0.057	0.877 ± 0.062	0.866 ± 0.068	0.737 ± 0.101	0.552 ± 0.120
label_0.05	0.989 ± 0.017	0.967 ± 0.038	0.975 ± 0.033	0.949 ± 0.054	0.913 ± 0.057	0.877 ± 0.062	0.866 ± 0.068	0.737 ± 0.101	0.552 ± 0.120
mix_0.03_0.03	0.986 ± 0.018	0.978 ± 0.029	0.971 ± 0.036	0.956 ± 0.042	0.902 ± 0.064	0.873 ± 0.030	0.834 ± 0.070	0.710 ± 0.080	0.597 ± 0.084

Supplementary Table 17-2. F1 Score

Noise\Label availability	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
clean	0.989 ± 0.017	0.967 ± 0.038	0.975 ± 0.033	0.949 ± 0.054	0.913 ± 0.056	0.876 ± 0.061	0.864 ± 0.069	0.732 ± 0.104	0.534 ± 0.124
feat_0.03	0.986 ± 0.018	0.978 ± 0.029	0.971 ± 0.037	0.956 ± 0.042	0.903 ± 0.062	0.873 ± 0.030	0.830 ± 0.072	0.709 ± 0.082	0.595 ± 0.087
feat_0.05	0.986 ± 0.018	0.978 ± 0.024	0.957 ± 0.054	0.964 ± 0.033	0.885 ± 0.072	0.841 ± 0.039	0.800 ± 0.091	0.659 ± 0.093	0.561 ± 0.074
label_0.03	0.989 ± 0.017	0.967 ± 0.038	0.975 ± 0.033	0.949 ± 0.054	0.913 ± 0.056	0.876 ± 0.061	0.864 ± 0.069	0.732 ± 0.104	0.534 ± 0.124
label_0.05	0.989 ± 0.017	0.967 ± 0.038	0.975 ± 0.033	0.949 ± 0.054	0.913 ± 0.056	0.876 ± 0.061	0.864 ± 0.069	0.732 ± 0.104	0.534 ± 0.124
mix_0.03_0.03	0.986 ± 0.018	0.978 ± 0.029	0.971 ± 0.037	0.956 ± 0.042	0.903 ± 0.062	0.873 ± 0.030	0.830 ± 0.072	0.709 ± 0.082	0.595 ± 0.087

This table reports the robustness evaluation of DELTA for the *material_dataset* under the same set of noise conditions. Similar to Table 16, the results include classification accuracy and F1 score across different label availability ratios and noise configurations.

Under the clean setting, the model achieves very high classification accuracy, reaching approximately 0.99 when label availability is 0.1, and gradually decreasing as the proportion of labeled samples becomes

smaller. This indicates that the DELTA framework can effectively exploit both labeled and unlabeled information to maintain high classification performance.

When feature noise is introduced, the classification accuracy decreases slightly but remains relatively high across most label availability levels. For example, under feat_0.03, the accuracy at label availability 0.1 remains around 0.986, and even with stronger perturbations (feat_0.05), the performance reduction remains moderate. These results suggest that the framework is capable of handling moderate feature perturbations without significant loss of predictive power.

Label noise scenarios again yield results almost identical to the clean baseline, indicating that small levels of label corruption have a limited impact on overall classification performance. This further demonstrates the robustness of the semi-supervised learning strategy used in DELTA.

Under the mixed-noise setting, where both feature and label perturbations are introduced simultaneously, the performance trends remain consistent with those in the feature-noise scenarios. Although accuracy and F1 score decrease slightly at higher noise levels or lower label availability, the model maintains a relatively strong classification capability.

Overall, the results presented in Supplementary Table 17 confirm that the DELTA framework remains robust and stable under realistic noisy conditions, even when both feature perturbations and label corruption are present. This robustness is particularly important for practical battery datasets, where measurement noise and imperfect labeling are unavoidable.

Supplementary Table 18 Comparing with other methods of the *batch_dataset*

Supplementary Table 18-1 Accuracy

Label availability	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
DELTA	0.9003 ± 0.0265	0.8849 ± 0.0301	0.8713 ± 0.0307	0.8723 ± 0.0419	0.8452 ± 0.0488	0.8271 ± 0.0497	0.7817 ± 0.0371	0.6892 ± 0.0855
FixMatch	0.8684 ± 0.0285	0.8602 ± 0.0322	0.8521 ± 0.0291	0.8391 ± 0.0328	0.8271 ± 0.0365	0.8020 ± 0.0493	0.8074 ± 0.0428	0.7558 ± 0.0285
SPRED	0.8916 ± 0.0324	0.8720 ± 0.0347	0.8602 ± 0.0343	0.8542 ± 0.0326	0.8304 ± 0.0393	0.8059 ± 0.0451	0.8171 ± 0.0398	0.7658 ± 0.0317
ST_RF	0.9194 ± 0.0149	0.9113 ± 0.0230	0.8850 ± 0.0292	0.8823 ± 0.0255	0.8633 ± 0.0372	0.8271 ± 0.0437	0.7664 ± 0.0560	0.6459 ± 0.0634
ST_SVM	0.8886 ± 0.0158	0.8695 ± 0.0268	0.8660 ± 0.0275	0.8596 ± 0.0205	0.8306 ± 0.0310	0.7953 ± 0.0370	0.7527 ± 0.0411	0.6521 ± 0.0533
UDA	0.8354 ± 0.0392	0.8231 ± 0.0346	0.8086 ± 0.0385	0.8038 ± 0.0435	0.7896 ± 0.0453	0.7604 ± 0.0635	0.7640 ± 0.0537	0.7262 ± 0.0286

Supplementary Table 18-2 F1 score

Label availability	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
DELTA	0.9001 ± 0.0264	0.8851 ± 0.0293	0.8713 ± 0.0307	0.8723 ± 0.0416	0.8438 ± 0.0503	0.8262 ± 0.0526	0.7826 ± 0.0378	0.6878 ± 0.0878
FixMatch	0.8669 ± 0.0291	0.8591 ± 0.0326	0.8512 ± 0.0294	0.8379 ± 0.0334	0.8259 ± 0.0370	0.8006 ± 0.0499	0.8056 ± 0.0446	0.7548 ± 0.0280
SPRED	0.8908 ± 0.0327	0.8713 ± 0.0351	0.8597 ± 0.0346	0.8534 ± 0.0329	0.8292 ± 0.0397	0.8050 ± 0.0454	0.8164 ± 0.0400	0.7650 ± 0.0313
ST_RF	0.9187 ± 0.0149	0.9104 ± 0.0233	0.8837 ± 0.0294	0.8814 ± 0.0257	0.8624 ± 0.0374	0.8262 ± 0.0438	0.7660 ± 0.0585	0.6398 ± 0.0683
ST_SVM	0.8886 ± 0.0156	0.8691 ± 0.0275	0.8654 ± 0.0276	0.8592 ± 0.0213	0.8309 ± 0.0306	0.7960 ± 0.0375	0.7534 ± 0.0405	0.6442 ± 0.0581
UDA	0.8318 ± 0.0407	0.8203 ± 0.0358	0.8065 ± 0.0394	0.8004 ± 0.0465	0.7862 ± 0.0485	0.7557 ± 0.0673	0.7602 ± 0.0576	0.7233 ± 0.0295

Supplementary Table 18 reports the detailed accuracy and F1 score results for DELTA and several baseline semi-supervised learning methods on the *batch_dataset* across different label-hiding ratios. The evaluated methods include FixMatch, SPRED, ST_RF, ST_SVM, and UDA, and the results are presented as mean ± standard deviation over repeated runs.

Overall, all methods show a gradual performance decline as the hide ratio increases, reflecting the increasing difficulty of the classification task when fewer labeled samples are available. Among the baseline methods, ST_RF achieves the highest performance under very low hide ratios, with an F1 score of 0.9254 at a hide ratio of 0.1, followed closely by GELTA and SPRED. However, as the hide ratio increases, the performance of most baseline methods deteriorates more rapidly. For example, when the hide ratio reaches 0.8, the F1 score of ST_RF decreases to approximately 0.766, while FixMatch and UDA show even lower values.

Compared with these baseline approaches, DELTA demonstrates better scalability with respect to label availability, as illustrated in Fig. 4a. Although some traditional semi-supervised methods achieve relatively strong performance when labeled data are abundant, their performance degrades significantly when labels become scarce. In contrast, DELTA maintains more stable classification performance across a wide range of label availability conditions, highlighting its ability to effectively exploit both labeled and unlabeled data.

Another notable observation from Table 18 is the close agreement between accuracy and F1 score across all methods, indicating that the classification results are relatively balanced across different classes without severe class imbalance effects. The reported standard deviations are also relatively small, suggesting that the results are stable across repeated experiments.

Overall, the results summarized in Supplementary Table 18 provide quantitative evidence supporting the superior scalability and stability of the DELTA framework on the batch-based dataset.

Supplementary Table 19 Comparing with other methods of the *material_dataset*

Supplementary Table 19-1 Accuracy

Label availability	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
DEALTA	0.9672 ± 0.0384	0.9745 ± 0.0332	0.9492 ± 0.0540	0.9128 ± 0.0569	0.8769 ± 0.0617	0.8657 ± 0.0679	0.7365 ± 0.1013	0.5517 ± 0.1200
FixMatch	0.8150 ± 0.0406	0.8003 ± 0.0657	0.7718 ± 0.0525	0.7518 ± 0.0639	0.6936 ± 0.0597	0.6769 ± 0.0589	0.6956 ± 0.0740	0.5955 ± 0.0837
SPRED	0.8379 ± 0.0345	0.8197 ± 0.0521	0.7915 ± 0.0434	0.7675 ± 0.0602	0.7056 ± 0.0635	0.6987 ± 0.0522	0.7153 ± 0.0719	0.6064 ± 0.0778
ST_RF	0.8806 ± 0.0598	0.8548 ± 0.0724	0.8368 ± 0.0783	0.8048 ± 0.0831	0.7464 ± 0.0664	0.6636 ± 0.0947	0.5550 ± 0.1016	0.4169 ± 0.0978
ST_SVM	0.8733 ± 0.0399	0.8623 ± 0.0480	0.8478 ± 0.0455	0.8082 ± 0.0476	0.7898 ± 0.0560	0.7796 ± 0.0525	0.7360 ± 0.0515	0.5321 ± 0.1171
UDA	0.7399 ± 0.0504	0.7436 ± 0.0622	0.7136 ± 0.0517	0.6873 ± 0.0624	0.6596 ± 0.0578	0.6404 ± 0.0552	0.6608 ± 0.0725	0.5929 ± 0.0816

Supplementary Table 19-2 F1 score

Label availability	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
DELTA	0.9672 ± 0.0382	0.9747 ± 0.0326	0.9492 ± 0.0541	0.9134 ± 0.0557	0.8757 ± 0.0614	0.8639 ± 0.0686	0.7315 ± 0.1039	0.5342 ± 0.1239
FixMatch	0.8051 ± 0.0449	0.7838 ± 0.0784	0.7562 ± 0.0681	0.7405 ± 0.0715	0.6779 ± 0.0626	0.6586 ± 0.0647	0.6781 ± 0.0786	0.5750 ± 0.1001
SPRED	0.8306 ± 0.0387	0.8097 ± 0.0589	0.7799 ± 0.0514	0.7598 ± 0.0661	0.6946 ± 0.0660	0.6848 ± 0.0592	0.7026 ± 0.0757	0.5891 ± 0.0899
ST_RF	0.8790 ± 0.0631	0.8549 ± 0.0736	0.8379 ± 0.0786	0.8041 ± 0.0852	0.7462 ± 0.0670	0.6625 ± 0.0955	0.5132 ± 0.1360	0.3134 ± 0.1427
ST_SVM	0.8710 ± 0.0406	0.8613 ± 0.0482	0.8476 ± 0.0462	0.8076 ± 0.0474	0.7879 ± 0.0556	0.7776 ± 0.0546	0.7152 ± 0.0621	0.4617 ± 0.1631
UDA	0.7216 ± 0.0530	0.7220 ± 0.0744	0.6963 ± 0.0596	0.6727 ± 0.0665	0.6428 ± 0.0593	0.6182 ± 0.0637	0.6415 ± 0.0748	0.5697 ± 0.0936

Supplementary Table 19 presents the detailed accuracy and F1 score comparisons for the *material_dataset*, corresponding to the results shown in Fig. 4b. Similar to Table 18, the evaluated methods include FixMatch, SPRED, ST_RF, ST_SVM, and UDA.

In general, the performance of all methods decreases as the hide ratio increases, which corresponds to a reduction in the proportion of labeled training samples. Among the baseline methods, GMMSSLM

achieves the highest performance under very low hide ratios, with an accuracy and F1 score close to 0.97 when the hide ratio is 0.1–0.2. However, its performance declines rapidly as the hide ratio increases. For example, when the hide ratio reaches 0.8, its F1 score drops to approximately 0.53, indicating a significant degradation when labeled data become scarce.

Other baseline methods show similar trends. ST_RF and ST_SVM initially achieve moderate performance but experience substantial performance loss at higher hide ratios. FixMatch and UDA exhibit relatively lower performance across most label availability levels, reflecting the challenges of applying general-purpose semi-supervised learning methods to battery degradation data.

Compared with these baseline methods, DELTA consistently achieves higher F1 scores across nearly all label availability levels, as shown in Fig. 4b. In particular, at moderate hide ratios (e.g., 0.3–0.5), where the classification problem becomes more challenging, DELTA maintains significantly higher performance than the baseline methods. This improvement demonstrates the advantage of incorporating domain-aware feature extraction and statistical modelling in the DELTA framework.

Additionally, the relatively small standard deviations reported in Table 19 indicate that the performance improvements are statistically stable across repeated experiments. These results further confirm that DELTA provides more robust and reliable classification performance for partially labeled battery datasets.

Supplementary Table 20. Computational efficiency comparison on the *batch_dataset*

Method	Accuracy	F1 score	Training time(s)	Predict time(s)	Total time(s)
DELTA	0.8849 ± 0.0301	0.8851 ± 0.0293	0.0747 ± 0.0083	0.0001 ± 0.0002	0.0748 ± 0.0083
FixMatch	0.8602 ± 0.0322	0.8591 ± 0.0326	8.2112 ± 0.4373	0.0007 ± 0.0001	8.2119 ± 0.4374
SPRED	0.8720 ± 0.0347	0.8713 ± 0.0351	7.2506 ± 0.5109	0.0007 ± 0.0001	7.2513 ± 0.5110
ST_RF	0.9113 ± 0.0230	0.9104 ± 0.0233	0.2260 ± 0.0691	0.0000 ± 0.0000	0.2260 ± 0.0691
ST_SVM	0.8695 ± 0.0268	0.8691 ± 0.0275	0.4097 ± 0.1297	0.0000 ± 0.0000	0.4097 ± 0.1297
UDA	0.8231 ± 0.0346	0.8203 ± 0.0358	8.0285 ± 0.4345	0.0007 ± 0.0001	8.0291 ± 0.4345

Supplementary Table 20 summarizes the computational efficiency and classification performance of different methods on the *batch_dataset*, including accuracy, F1 score, training time per run, inference time per run, and total runtime.

Among the baseline methods, ST_RF achieves the highest predictive performance with an accuracy of 0.9113 ± 0.0230 and an F1 score of 0.9104 ± 0.0233, while maintaining a relatively short training time of 0.2260 s. DELTA also demonstrates strong performance, reaching an accuracy of 0.8849 ± 0.0301 with a very short training time of 0.0747 s, making it one of the most computationally efficient baseline methods.

In contrast, deep semi-supervised learning approaches such as FixMatch, SPRED, and UDA require substantially longer training times, typically around 7-8 s per run, which is one to two orders of magnitude higher than those of traditional machine-learning–based approaches. Despite the higher computational cost, their classification performance remains lower than that of the best-performing methods on this dataset.

Across all methods, the inference time per sample is extremely small (on the order of 10^{-4}-10^{-3} s), indicating that prediction itself is computationally inexpensive once the models are trained. However, the substantial differences in training time highlight the importance of algorithmic efficiency when models need to be retrained repeatedly or applied to large-scale datasets.

Overall, these results demonstrate clear trade-offs between predictive performance and computational cost among different semi-supervised learning strategies on the *batch_dataset*, with simpler models achieving significantly faster training while maintaining competitive classification accuracy.

Supplementary Table 21. Computational efficiency comparison on the *material_dataset*

Method	Accuracy	F1 score	Training time(s)	Predict time(s)	Total time(s)
GMMSSLM	0.8657 ± 0.0679	0.8639 ± 0.0686	0.2288 ± 0.1225	0.0001 ± 0.0002	0.2290 ± 0.1224
DELTA	0.6769 ± 0.0589	0.6586 ± 0.0647	0.9952 ± 0.0133	0.0002 ± 0.0001	0.9955 ± 0.0133
SPRED	0.6987 ± 0.0522	0.6848 ± 0.0592	0.8708 ± 0.0097	0.0002 ± 0.0001	0.8711 ± 0.0097
ST_RF	0.6636 ± 0.0947	0.6625 ± 0.0955	0.0939 ± 0.0139	0.0000 ± 0.0000	0.0939 ± 0.0139
ST_SVM	0.7796 ± 0.0525	0.7776 ± 0.0546	0.0139 ± 0.0060	0.0000 ± 0.0000	0.0139 ± 0.0060
UDA	0.6404 ± 0.0552	0.6182 ± 0.0637	0.9686 ± 0.0178	0.0003 ± 0.0001	0.9689 ± 0.0179

Supplementary Table 21 presents the computational efficiency and classification performance of different methods on the *material_dataset*. Compared with the *batch_dataset*, the overall computational cost on this dataset is substantially lower, with the training time of all baseline methods remaining below 1s per run, reflecting the smaller dataset size and reduced computational complexity.

Among the evaluated methods, DELTA achieves the best predictive performance, with an accuracy of 0.8657 ± 0.0679 and an F1 score of 0.8639 ± 0.0686, while requiring a moderate training time of 0.2288 s. ST_SVM also shows relatively strong performance, reaching an accuracy of 0.7796 ± 0.0525 with an extremely short training time of only 0.0139 s, highlighting its high computational efficiency.

Deep semi-supervised approaches such as FixMatch, SPRED, and UDA exhibit longer training times (approximately 0.87-1.00 s) while delivering comparatively lower predictive performance on this dataset. In contrast, ST_RF achieves a very short training time (0.0939 s) but shows relatively lower classification accuracy.

Similar to the observations on the batch_dataset, the inference time for all methods remains extremely small, typically on the order of 10^{-4} s per sample, indicating that prediction is computationally inexpensive once the models are trained.

Overall, these results suggest that while computational costs are generally lower on the material_dataset, there remain clear differences in the balance between training efficiency and predictive performance among different algorithms.

Supplementary Table 22. Benchmark settings used for fair comparison of DELTA and baseline methods

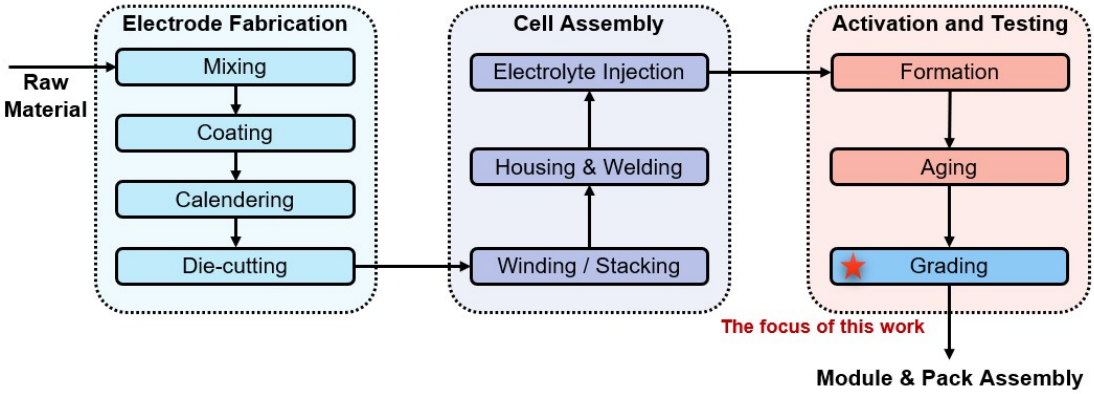
Category	Parameter	Value
Data partition	Cross-validation	10-fold stratified cross-validation
Data partition	Repeated runs	3
Data partition	Random seed	42
Label availability	Hidden-label ratio	0.7
Label availability	Available-label ratio	0.3
Training	Epochs	100
Training	Evaluation interval	Every 10 epochs
Batch size	Labeled samples	16
Batch size	Unlabeled samples	32
Batch size	Test samples	64
Network architecture	Backbone	Three-layer MLP
Network architecture	Hidden dimension	64
Network architecture	Dropout	0.3
Optimization	Optimizer	Adam
Optimization	Learning rate	1×10^{-3}
Optimization	Weight decay	1×10^{-3}
Data augmentation	Weak augmentation noise	0.02
Data augmentation	Strong augmentation noise	0.10
FixMatch	Confidence threshold	0.95
FixMatch	Unsupervised loss weight	1.0
UDA	Temperature	0.7
UDA	Unsupervised loss weight	1.0
SPRED	Regularization weight	0.1
Self-training	Confidence threshold	0.9
Self-training	Maximum iterations	10

All methods were trained and evaluated using identical data partitions, label-availability settings, input features, and evaluation metrics. Neural-network-based methods shared the same backbone architecture and optimization settings to ensure a fair comparison. Method-specific hyperparameters were fixed according to the unified benchmark configuration.

Supplementary Table 23. Definitions of key datasets and classification schemes used in this study.

Term	Definition
Base dataset	The original dataset containing cells with complete lifetime labels used for model development and evaluation.
Extended dataset	Additional battery data incorporated to increase data diversity and simulate practical limited-label scenarios.
Labeled extended dataset	Extended dataset samples whose lifetime labels are available during model training.
Unlabeled extended dataset	Extended dataset samples whose lifetime labels are intentionally treated as unavailable during model training.
Batch dataset	A dataset constructed by grouping cells according to manufacturing batch. Cells within the same batch serve as the reference population for batch-level lifetime classification.
Material dataset	A dataset constructed by grouping cells according to chemistry or material system, regardless of manufacturing batch.
Batch-based classification	Lifetime classification performed relative to the lifetime distribution of cells within the same batch, primarily used for manufacturing quality assessment and consistency evaluation.
Material-based classification	Lifetime classification performed relative to the lifetime distribution of cells sharing the same material system, primarily used for cross-batch comparison and technology benchmarking.
Label availability	The proportion of samples with accessible lifetime labels during model training.
Lifetime category (L/N/H)	Relative lifetime labels indicating Low, Normal, and High lifetime performance within a specified reference population.

Supplementary Note 1. Relationship between battery manufacturing workflow and early-cycle quality assessment



As illustrated in Figure, the manufacturing process of a typical lithium-ion battery consists of three major stages: Electrode Fabrication, Cell Assembly, and Activation and Testing. Electrode fabrication includes slurry mixing, coating, calendaring, and die-cutting, whereas cell assembly involves electrode winding/stacking, housing and welding, and electrolyte injection. The present study focuses on the final stage, Activation and Testing, because it represents the last manufacturing-end quality control procedure before cells are released for downstream applications and therefore provides the most direct basis for industrial quality evaluation.

The Activation and Testing stage mainly comprises three sequential procedures: formation, aging, and capacity grading. During formation, newly assembled cells undergo initial charge-discharge activation, promoting electrochemical conversion of active materials and the formation of stable solid electrolyte interphase (SEI) layers. Following formation, cells enter the aging stage, during which they are stored under controlled conditions for several days to facilitate electrolyte wetting and further stabilization of the SEI layer. A key quality-control objective during aging is the identification of abnormal self-discharge behaviour, which may originate from micro-short circuits, metallic contaminants, separator defects, or other latent manufacturing imperfections. Subsequently, cells undergo capacity grading, where standardized charge–discharge tests are performed to determine practical capacity and internal resistance. Based on these measurements, cells are classified into different quality grades for subsequent utilization.

The datasets used in this work originate from individual cells that successfully passed industrial grading procedures and had not yet been assembled into battery packs. Importantly, battery degradation is a gradual and continuous process rather than a discrete event. Consequently, the electrochemical state of a cell during its first few operational cycles remains highly similar to the state established immediately after formation, aging, and capacity grading. From this perspective, early-cycle behaviour can be regarded as a direct continuation of the manufacturing-end cell condition.

Although the data analyzed in this study do not contain explicit upstream manufacturing parameters such

as coating thickness, calendaring density, or electrolyte filling volume, the early-cycle voltage–capacity characteristics inherently reflect the cumulative effects of the preceding manufacturing and quality-control processes. Variations introduced during electrode fabrication, cell assembly, and activation procedures can influence the electrochemical state of the cell and subsequently manifest as subtle differences in early-cycle behaviour. Therefore, early-cycle signals provide observable end-of-line signatures of manufacturing-induced quality variation, even when direct process information is unavailable.

Current industrial quality grading primarily relies on capacity, internal resistance, and self-discharge measurements. While these indicators are effective for detecting severe defects and assessing immediate cell performance, they provide limited information regarding long-term degradation trajectories. As a result, cells exhibiting similar grading metrics may still experience substantially different ageing behaviour during subsequent operation. The central objective of this study is therefore not to reconstruct manufacturing parameters or identify specific process root causes, but rather to extract latent quality information from routinely collected early-cycle data and establish its relationship with future lifetime outcomes.

Accordingly, the proposed framework should be viewed as an outcome-oriented manufacturing quality assessment approach. By linking end-of-line testing signatures to subsequent degradation behaviour, the framework provides a practical means of evaluating manufacturing-related quality variation under realistic industrial conditions where direct manufacturing process data are often inaccessible.

Supplementary Note 2. Analysis of Data Heterogeneity

The DELTA framework is designed to achieve robust generalization across both dataset domains and material domains. Accordingly, 4 major material systems (LFP, NCA, NCM, and NCANCM), drawn from the six datasets, were included in the evaluation. Even when nominal material labels are identical, the underlying chemical compositions differ across datasets. For example, the NCM chemistry in the XJTU dataset is $LiNi_{0.5}Co_{0.2}Mn_{0.3}O_2$, whereas the TJU dataset adopts $Li_{0.86}Ni_{0.86}Co_{0.11}Mn_{0.07}O_2$, and the Stanford dataset uses $LiNi_{0.5}Mn_{0.3}Co_{0.2}O_2$. Similarly, the NCA chemistry in the TJU dataset is $Li_{0.86}Ni_{0.86}Co_{0.11}Al_{0.03}O_2$, while the CAS dataset employs $LiNi_xCo_yAl_{1-x-y}O_2$. Batteries belonging to the same material system are therefore grouped together, and material differences are treated as a primary driver of ageing behaviour. In addition, even within the same material category, cells from different datasets exhibit variations in nominal capacity and cutoff voltage. To mitigate the influence of these factors, voltage signals were uniformly sampled over the entire voltage window, and normalization was applied during HI construction. All cells follow a CCCV charging protocol and CC discharging protocol, with the main operational differences arising from charge-rate settings. As the extracted HIs are derived exclusively from charging-phase data, only charge-rate groupings are considered. Taken together, material system, temperature and charge rate are identified as the three dominant factors governing battery ageing, with material effects being the most pronounced. Consequently, in the subsequent LME model, material is treated as a random effect, while temperature and charge rate are modelled as fixed effects, enabling quantification of material-dependent ageing behaviour under controlled temperature and charge-rate conditions.

Supplementary Note 3. Three-class scheme based on 1sigma boundaries of a Gaussian distribution

In extracting EOL labels, two key questions must be addressed: (1) how many EOL cycle-life categories should be defined; and (2) which specific partitioning strategy should be adopted for a review of related literature. In early-life screening applications, the primary objective is to identify cells with potentially large performance deviations, thereby ensuring consistency within battery packs.

From the perspective of category design, cycle-life classification generally includes binary classification and multi-class classification. Binary classification typically distinguishes between "good" and "bad" batteries, which in practical applications often corresponds to identifying "long-life batteries" and "short-life batteries." Multi-class classification provides a more refined categorization. In practice, batteries are commonly divided into three categories: short-life, medium-life, and long-life. The objective is to remove abnormally short-lived or excessively long-lived cells from a large population of otherwise normal batteries, thereby ensuring stable and controllable operation of battery packs.

Regarding specific label assignment methods, existing studies primarily adopt two strategies: proportion-based partitioning and threshold-based partitioning. Proportion-based partitioning assigns labels based on the quantiles of the cycle-life distribution across the entire dataset. For example, Lucas Murphy[8] defined batteries within the top 25% quantile of cycle life as long-life cells, considering them suitable for secondary use. Minzheng Hu[9] classified batteries into three categories, L (long-life), M (medium-life), and S (short-life), based on their quantile positions within the overall dataset.

Threshold-based partitioning, in contrast, classifies batteries by setting explicit cycle-life thresholds, often in combination with proportion-based methods. For instance, Xuelu Wang[10] first divided the dataset according to cycle-life quantiles and then defined corresponding thresholds for three-class classification. Sandro Stock[11] directly adopted 150 cycles as the threshold for binary classification and used 150 and 300 cycles as thresholds for three-class classification. Zicheng Fei[12] selected 550 cycles as the binary classification threshold and used 550 and 1200 cycles for three-class classification, while Yongzhi Zhang[13] also used 550 cycles as the binary classification threshold.

It should be noted that the selection of thresholds should depend on the characteristics of the batteries under investigation; therefore, threshold-based partitioning is often combined with proportion-based methods to improve rationality and interpretability.

As shown in **Fig. 2c**, the distribution of battery cycle life in practical scenarios is often approximately Gaussian, with the majority of cells exhibiting intermediate lifetimes and relatively few cells at the extremes. Accordingly, a three-class scheme based on the 1σ boundaries of a Gaussian distribution, defined by the quantiles of the dataset-specific cycle-life distribution, better reflects practical requirements.

Specifically, cells with cycle life falling within the interval $\mu \pm 1\sigma$ (where μ and σ denote the mean and

standard deviation, respectively) exhibit high concentration and limited performance variability and therefore constitute the core population for assembling highly consistent battery packs. By contrast, cells located below $\mu - 1\sigma$ or above $\mu + 1\sigma$, each accounting for approximately 16% of the population, are of particular interest for early screening. Cells in the lower tail $(0, \mu - 1\sigma)$ exhibit significantly reduced lifetime and may suffer from manufacturing-related defects, such as electrode material imperfections, non-uniform coating, or insufficient electrolyte filling, and are therefore likely to become "weak links" if incorporated into a pack. Cells in the upper tail $(\mu + 1\sigma, +\infty)$, although longer-lived, deviate substantially from the mainstream population and may introduce system-level issues, including capacity imbalance and increased difficulty in equalization. Consequently, both groups warrant careful review to ensure stable, consistent and reliable battery-pack performance in large-scale applications. To emulate incomplete label availability in real battery production settings, we further constructed a label-missing simulation mechanism based on feature dispersion. Specifically, the variance of the sample features was used as a proxy for sample complexity, with samples exhibiting higher dispersion assigned a higher probability of label missingness.

Supplementary Note 4. Economic analysis under end-of-line screening scenario

At the end of battery production, rapid and reliable screening of potentially risky cells is essential for ensuring the consistency, reliability, and safety of downstream battery pack integration. However, in practical manufacturing environments, end-of-life (EOL) labeled data are costly and time-consuming to obtain, and therefore typically exist only in limited quantities. In contrast, large volumes of early-cycle operational data are routinely generated during manufacturing but often remain unlabeled and underutilized.

To address this realistic industrial setting, as shown in Fig. S1, this study considers an end-of-line screening scenario characterized by scarce labeled data and abundant unlabeled data. The proposed DELTA framework uses the first $k$early cycles as input features and performs a three-class classification of battery EOL cycle life under a semi-supervised learning setting where only a fraction $r$of the samples are labeled. The three classes correspond to abnormally high, normal, and abnormally low lifetimes, enabling the identification of potentially risky batteries at the production stage. Batteries classified as normal can proceed directly to downstream processes, whereas batteries predicted as risky require additional inspection. Early identification of such batteries helps mitigate potential safety risks and improve pack-level consistency.

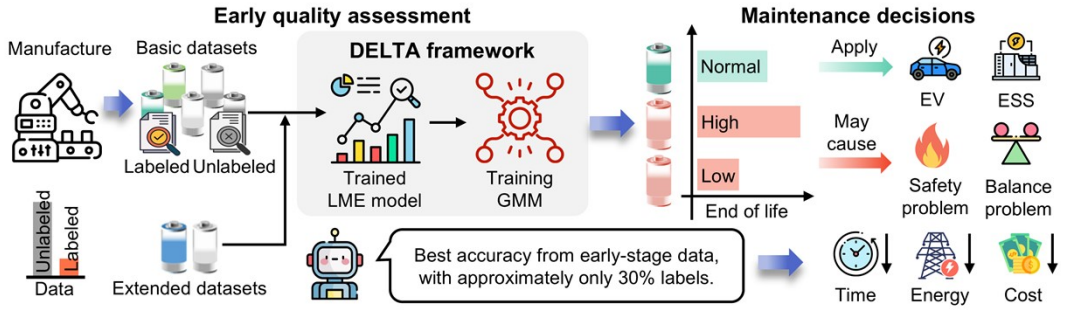


Fig. S1. Deployment of DELTA framework

To evaluate the economic feasibility of the proposed framework, we construct an expected economic cost model considering three main components:

1. Risk-related losses caused by undetected risky batteries
2. Additional re-inspection costs for batteries flagged as risky
3. Label acquisition costs required for model training

Economic assumptions

The economic analysis is conducted under a representative industrial production scale corresponding to 1.3 TWh[14] annual battery production. Assuming a nominal energy capacity of 100 Wh per cell, the annual production volume is approximately

$$N_{annual} = 1.3 \times \frac{10^{12}}{100} \approx 1.3 \times 10^{10}\ cells\ per\ year \quad (1)$$

For batteries that are incorrectly classified as normal (false negatives), a potential field failure may occur during downstream use. The probability that such a missed risky battery leads to a failure event is assumed to lie within

$$p_{fail} \in [10^{-5}, 2 \times 10^{-5}] \quad (2)$$

and the midpoint value is used in the calculation.

The economic loss associated with a single field failure event is assumed to lie within

$$c_{event} \in [20{,}000, 80{,}000]\ USD \quad (3)$$

Therefore, the expected loss per undetected risky battery is

$$c_{\text{risk}} = p_{\text{fail}} \times c_{\text{event}} \quad (4)$$

For batteries predicted as risky, an additional end-of-line re-inspection procedure is required. This includes equipment operation, labor, and additional testing costs. The unit re-inspection cost is assumed to lie within

$$c_{recheck} \in [0.08, 0.25]\ USD \quad (5)$$

Meanwhile, labeled samples required for model training must be obtained through full lifetime testing. The unit labeling cost is assumed to lie within

$$c_{label} \in [500, 1000]\ USD \quad (6)$$

In this study, the number of labeled samples collected per year is assumed to be

$$N_{label} = 10{,}000 \quad (7)$$

In addition, the deployment of the AI-based screening system introduces an annual operational cost including computing infrastructure, data management, and model maintenance. This annual system cost is assumed to lie within

$$C_{system} \in [1.5, 4.0]\ million\ USD \quad (8)$$

Midpoint values of these ranges are used in the economic calculations.

Expected economic cost model

Let TP, FP, and FN denote the numbers of correctly detected risky batteries, falsely flagged normal batteries, and missed risky batteries in the evaluation dataset, respectively. Let N denote the total number of evaluated samples.

The corresponding rates are

$$FN\ rate = \frac{FN}{N}, flag\ rate = \frac{TP + FP}{N} \tag{9}$$

The annual expected economic cost of the proposed screening strategy is defined as

$$S_{DELTA} = FN_{rate} \times N_{annual} \times c_{risk} + flag\ rate \times N_{annual} \times c_{recheck} + N_{label} \times c_{label} + C_{system} \tag{10}$$

This formulation captures the trade-off among risk mitigation, inspection workload, and model deployment cost.

Baseline strategies

Two baseline strategies are introduced for comparison.

Baseline 1: No screening.

In this strategy, all batteries are directly deployed without end-of-line screening. Although no inspection or labeling cost is incurred, all risky batteries remain undetected. The expected annual cost therefore arises entirely from potential risk-related losses:

$$S_{B1} = risk\ rate \times N_{annual} \times c_{risk} \tag{11}$$

Baseline 2: Random sampling screening

In this strategy, a proportion $s$of batteries is randomly selected for inspection. Because the sampling process does not target risky batteries, its risk coverage capability is limited. An effective risk coverage coefficient $\eta_{B2} = 0.5$ is introduced to represent the fraction of risky batteries that can be detected through random inspection. The residual risk rate is therefore

$$(1 - \eta_{B2} s) \times risk\ rate \tag{12}$$

The total annual cost of this strategy becomes

$$S_{B2} = (1 - \eta_{B2} s) \times risk\ rate \times N_{annual} \times c_{risk} + s \times N_{annual} \times c_{recheck} \tag{13}$$

In this study, a representative sampling ratio $s = 0.2$ is considered.

Economic evaluation results

Under the above assumptions, the proposed DELTA framework yields an average annual economic cost of approximately 2433.67 million USD at the considered production scale of 1.3 TWh per year. In comparison, the no-screening strategy results in an expected annual cost of approximately 6500.16 million USD, as all risky batteries remain undetected and may lead to downstream losses. The random sampling strategy with sampling ratio $s = 0.2$ yields an annual cost of approximately 6279.15 million USD, providing only limited reduction in risk-related losses due to its lack of targeted detection capability. Overall, the proposed framework reduces the expected annual economic cost by approximately 4066 million USD, corresponding to a 62.6% reduction relative to the no-screening strategy.

Discussion

The economic analysis further demonstrates the practical value of the proposed DELTA framework in large-scale battery manufacturing. At the production scale considered in this study (1.3 TWh annual output), the proposed method yields an expected annual economic cost of approximately 2433.67 million USD. In comparison, the no-screening baseline results in an expected cost of 6500.16 million USD per year, as all risky batteries remain undetected and may lead to downstream failures or costly corrective actions.

As a result, the proposed framework achieves an annual economic saving of approximately 4066.49 million USD, corresponding to a 62.57% reduction in overall economic cost relative to the no-screening strategy. This large reduction primarily arises from the improved ability of the proposed method to identify potentially risky batteries early, thereby preventing costly downstream failures while maintaining a manageable inspection workload.

By contrast, the random sampling screening strategy with a sampling ratio of $s = 0.2$results in an annual cost of approximately 6279.15 million USD, which is slightly lower than the no-screening strategy but still significantly higher than the proposed method. In fact, relative to the no-screening baseline, the random sampling strategy yields an additional cost of approximately 221.02 million USD, corresponding to a 3.52% economic loss, indicating that untargeted inspection may introduce unnecessary inspection expenses without effectively mitigating risk.

Overall, these results demonstrate that the DELTA framework provides a significantly more cost-effective end-of-line screening strategy than both no-screening and random sampling. By combining targeted risk identification with semi-supervised learning that leverages abundant unlabeled manufacturing data, the proposed method substantially reduces risk-related losses while maintaining economically acceptable inspection and labeling costs. This highlights its strong potential for deployment in gigafactory-scale battery production systems.

Supplementary Note 5. Feature extraction

Input signals To enable accurate EOL cycle-life classification using early-cycle data, a set of HIs is constructed based on the evolution of the capacity degradation curve and the Q-V curve. All selected features are extracted exclusively from early-cycle data and are screened through correlation analysis with EOL cycle life. The resulting features fall into three categories: early normalized capacity features, Q-V curve-based features, and capacity degradation curve features.

Early normalized capacity features Let Q_{rated} denote the nominal capacity of a cell, and Q_k the discharge capacity at the k-th cycle. To eliminate the influence of differences in nominal capacity across cells, the normalized capacity is defined as

$$\tilde{Q}_k = \frac{Q_k}{Q_{rated}} \tag{14}$$

The normalized capacities at the 1st, 3rd and 5th cycles are selected as HI_1, HI_2 and HI_3, respectively. These features characterize the initial health state immediately after formation and the early capacity retention behaviour of the cell and represent the most intuitive and physically interpretable inputs for lifetime assessment.

Q-V curve–based features Let $Q_k(V)$ denote the Q-V relationship at the k-th cycle. To capture changes in curve morphology during early cycling, a differential Q-V function between the 3rd and 5th cycles is defined as

$$Q(V) = Q_3(V) - Q_5(V) \tag{15}$$

This function reflects the distribution of capacity changes induced by ageing across the entire voltage window. Based on $\Delta Q(V)$, the following statistical descriptors are extracted as health indicators:

$$HI_4 = \frac{E[\Delta Q(V)]}{Q_{rated}}, HI_5 = Kurt(\Delta Q(V)), HI_6 = Skew(\Delta Q(V)) \tag{16}$$

Here, kurtosis quantifies the concentration of the distribution, while skewness captures its asymmetry. Together, these metrics reflect localized degradation heterogeneity in the Q-V curves over specific voltage regions.

Capacity degradation curve features To further characterize the trend of capacity decay with cycling, linear fitting is applied to early-cycle capacity data. For the normalized capacities from the 1st to the 3rd cycles, a linear model is fitted as

$$\tilde{Q}_k = a_1 k + b_1 \tag{17}$$

and the intercept term is defined as a health indicator:

$$HI_7 = b_1 \tag{18}$$

Similarly, a linear fit is performed for the normalized capacity data from the 3rd to the 5th cycles, and the corresponding intercept is defined as

$$HI_8 = b_2 \tag{19}$$

All health indicators are constructed solely from early-cycle data and are subjected to correlation analysis with EOL cycle life. Only features exhibiting a strong correlation with EOL are retained as model inputs. In total, 15 candidate features are initially extracted, of which the eight most informative are selected, thereby maintaining predictive performance while reducing feature redundancy.

Supplementary Note 6. Linear mixed-effects model

In practical battery operation, temperature, charge rate, and battery material are the most influential factors governing ageing behaviour, with material effects typically exhibiting the largest variability. To investigate and quantify the impact of material on battery ageing, the LME model treats temperature and charge rate as fixed effects and material as a random effect, thereby capturing and quantifying material-level random variation. The empirical best linear unbiased prediction (EBLUP) of the material random effect is subsequently extracted as the feature HI_m and used as input to the subsequent GMM-based semi-supervised classification framework.

Assume that for the i-th material cluster ($i = 1,2,...,g$), the observed cycle life of the j-th cell ($j = 1,2,...,m_i$) is denoted by $Cycle_{ij}$, with T_{ij} and C_{ij} representing the corresponding temperature and charge-rate groups, respectively. Here, g denotes the number of material clusters, and m_i the number of cells within the i-th cluster. The LME model is formulated as

$$Cycle_{ij} = \beta_0 + \beta_T T_{ij} + \beta_C C_{ij} + \alpha_i + e_{ij} \tag{20}$$

where β_0 is the fixed-effect intercept, β_T and β_C are the regression coefficients for temperature and charge rate, respectively, α_i denotes the random effect associated with the i-th material cluster, and e_{ij} is the residual error term. The random effects $\{\alpha_i\}$ and residuals $\{e_{ij}\}$ are assumed to be mutually independent with zero means and variances σ_α^2 and σ_e^2, respectively. No normality assumption is imposed on their distributions. The model treats different clusters as independent and assumes homogeneous within-cluster correlation. The estimated random effects $\hat{\alpha}_i$ are extracted as the material-related health indicator HI_m and combined with the eight dynamic health indicators to form the input feature vector for classification.

Supplementary Note 7. Semi-supervised Gaussian mixture model

Battery health states are modelled using a finite Gaussian mixture model. Let the feature vector of the j-th cell be

$$y_j = (y_{j1}, y_{j2}, \ldots, y_{jp})^{\top} \in R^p \tag{21}$$

which is assumed to be generated from one of g Gaussian components:

$$Y_j \sim N(\mu_i, \Sigma_i) \text{ with probability } \pi_i \tag{22}$$

where μ_i and Σ_i denote the mean vector and covariance matrix corresponding to the i-th health state, and π_i is the prior probability satisfying $\sum_{i=1}^{g} \pi_i = 1$.

Under the semi-supervised learning setting, the EOL labels of a subset of samples are observed, while the class memberships of the remaining samples are treated as latent variables during model estimation. For labeled samples, class information is directly incorporated into parameter estimation, whereas for unlabeled samples, class assignments are inferred via posterior probabilities.

Modelling the label-missing mechanism

In practical battery testing, missing health-state labels are often correlated with the degree of battery ageing. To account for this dependency, a label-missing indicator variable $m_j$is introduced, where $m_j = 1$ indicates that the class label of the j-th sample is missing, and $m_j = 0$ indicates that the label is observed. The probability of label missingness is assumed to depend solely on the observed feature vector y_j and is characterized through posterior classification uncertainty:

$$Pr(m_j = 1 | y_j) = q(y_j; \theta, \xi) \tag{23}$$

where $q(\cdot)$ is a logistic function whose input is the Shannon entropy computed from the posterior class probabilities τ_{ij}:

$$e_j = -\sum_{i=1}^{g} \tau_{ij} log \tau_{ij} \tag{24}$$

By introducing the label-missing mechanism, the complete partially classified log-likelihood can be expressed as

$$log L_{pc}^{full}(\Psi; x_{pc}) = log L_{pc}^{ig}(\Psi; x_{pc}) + log L_{pc}^{miss}(\Psi; x_{pc}) \tag{25}$$

where Ψ denotes the vector of all unknown parameters, x_{pc} represents the partially classified samples,

and $\log L_{pc}^{ig}$ corresponds to the likelihood obtained by ignoring the missingness mechanism. The likelihood contribution of the missingness process is given by

$$logL_{pc}^{miss}(\theta,\xi;x_{pc}) = \sum_{j=1}^{n}[(1-m_j)log\{1-q(y_j;\theta,\xi)\} + m_j log q(y_j;\theta,\xi)] \tag{26}$$

Here, m_j is the missing-label indicator, and $q(y_j;\theta,\xi)$ is the entropy-based logistic function. The detailed procedures for computation and parameter estimation are provided in the Methods section. By explicitly modelling the label-missing mechanism, the proposed framework is able to exploit missingness information to improve both the accuracy and generalization of early-stage ageing classification.

Supplementary Note 8. Evaluation Metrics

Accuracy measures the overall proportion of correctly classified samples and is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

Recall quantifies the model's ability to identify positive samples and is defined as

$$Recall = \frac{TP}{TP + FN} \tag{28}$$

Precision represents the proportion of samples predicted as positive that are truly positive. Together with Recall, it is used to construct F-score-based composite metrics. Based on Precision and Recall, this study adopts both the F1 score and the F2 score as summary performance measures. The F1 score assigns equal importance to Precision and Recall and is defined as

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{29}$$

whereas the F2 score places greater emphasis on recall of the positive class and is defined as

$$F2 = \frac{5 \cdot Precision \cdot Recall}{4 \cdot Precision + Recall} \tag{30}$$

To further evaluate classification consistency under class-imbalanced conditions, the Matthews correlation coefficient (MCC) is employed. MCC jointly considers TP, FP, TN and FN and is defined as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{31}$$

In addition, to assess overall model performance across different decision thresholds, the area under the precision–recall curve (PR-AUC) is used. PR-AUC is obtained by integrating the precision–recall curve and provides a more informative measure of a model's ability to identify positive samples under imbalanced class distributions. By jointly considering these evaluation metrics, model performance in early-stage battery health-state classification is systematically analysed from multiple complementary perspectives.

Reference

1. Chen, C. *et al.* The Operation Dependence of $C - N$ Fatigue for Lithium-Ion Batteries. *Advanced Energy Materials* **13**, 2300942 (2023).
2. Weng, A. *et al.* Predicting the impact of formation protocols on battery lifetime immediately after manufacturing. *Joule* **5**, 2971–2992 (2021).
3. Li, W. *et al.* One-shot battery degradation trajectory prediction with deep learning. *Journal of Power Sources* **506**, 230024 (2021).
4. Cui, X. *et al.* Data-driven analysis of battery formation reveals the role of electrode utilization in extending cycle life. *Joule* **8**, 3072–3087 (2024).
5. Zhu, J. *et al.* Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. *Nat Commun* **13**, 2261 (2022).
6. Wang, F., Zhai, Z., Zhao, Z., Di, Y. & Chen, X. Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. *Nat Commun* **15**, 4332 (2024).
7. Tan, R. *et al.* BatteryLife: A Comprehensive Dataset and Benchmark for Battery Life Prediction. Preprint at https://doi.org/10.48550/arXiv.2502.18807 (2025).
8. Murphy, L. & Crawford, C. Data-Driven Classification of Lithium-Ion Batteries for Second-Life Applications. *Journal of Energy Storage* https://doi.org/10.2139/ssrn.5119726 (2025) doi:10.2139/ssrn.5119726.
9. Hu, M., Tao, S., Wang, Y. & Sun, Y. A Data-Driven Approach for Lithium-ion Battery Lifetime Classification Based on Early Cycles. in *2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2)* 2208–2213 (IEEE, Hangzhou, China, 2023). doi:10.1109/EI259745.2023.10512808.
10. Wang, X., Meng, J. & Azib, T. A Comparative Study of Data-Driven Early-Stage End-of-Life Classification Approaches for Lithium-Ion Batteries. *Energies* **17**, 4485 (2024).
11. Stock, S. *et al.* Early Quality Classification and Prediction of Battery Cycle Life in Production Using Machine Learning. *Journal of Energy Storage* **50**, 104144 (2022).
12. Fei, Z., Zhang, Z., Yang, F. & Tsui, K.-L. Deep learning powered rapid lifetime classification of lithium-ion batteries. *eTransportation* **18**, 100286 (2023).
13. Zhang, Y., Zhao, M. & Xiong, R. Online data-driven battery life prediction and quick classification based on partial charging data within 10 min. *Journal of Power Sources* **594**, 234007 (2024).
14. Tao, S. *et al.* Collaborative and privacy-preserving retired battery sorting for profitable direct recycling via federated machine learning. *Nat Commun* **14**, 8032 (2023).