

AI for Lithium-ion battery research under data-scarce scenarios

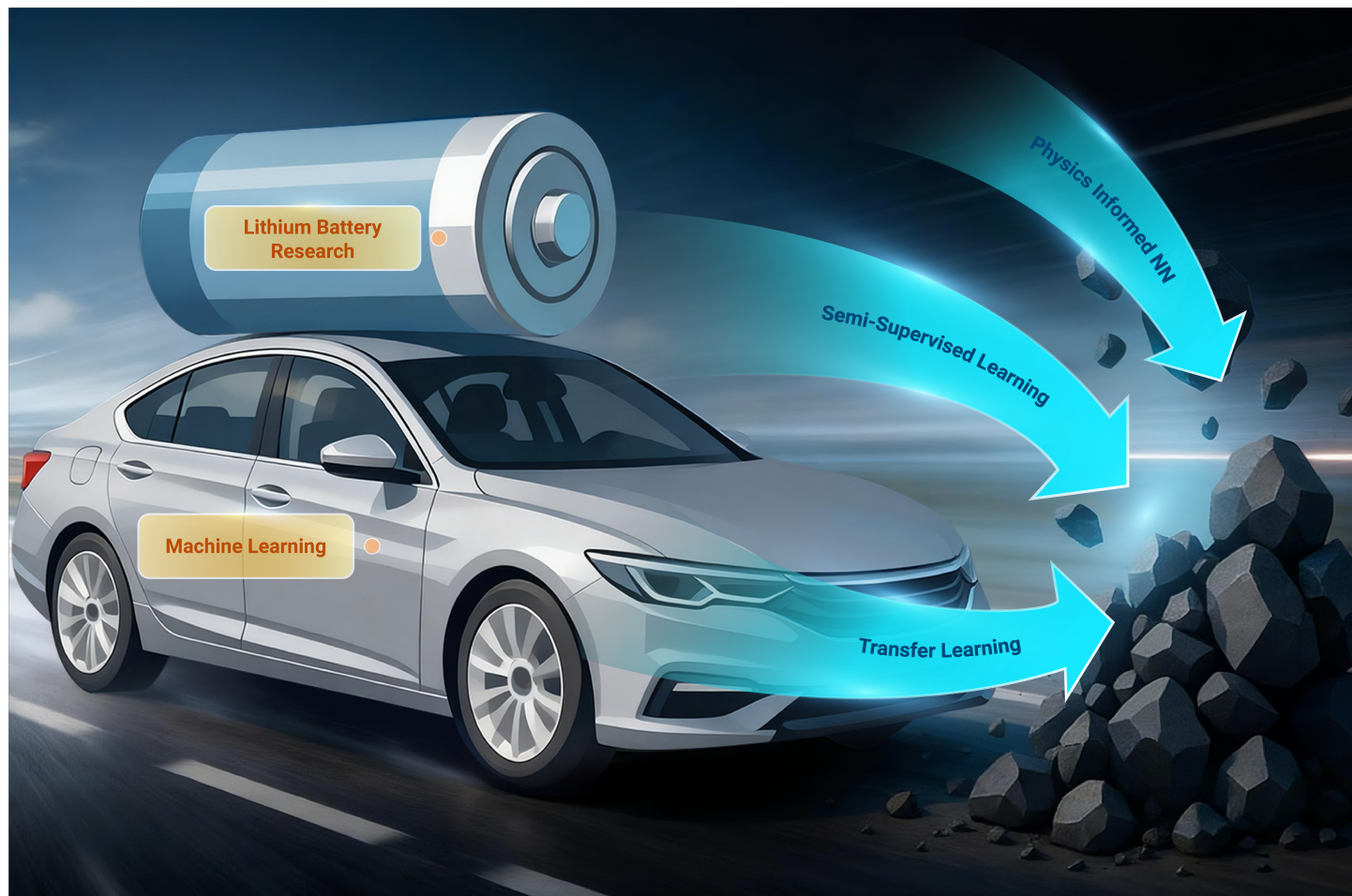
Junzhe Zhang,^{1,6} Xiaolang Zhai,^{2,6} Qiang Zhang,^{1,3,6} Peiyi Li,¹ Mian Li,¹ Kai Wu,^{4,*} Yang Liu,^{1,*} and Jiayu Wan^{1,5,*}

*Correspondence: wuk@catl.com (K.W.); liuyang130@sjtu.edu.cn (Y.L.); wanjy@sjtu.edu.cn (J.W.)

Received: July 21, 2025; Accepted: March 13, 2026; Published Online: March 13, 2026; <https://doi.org/10.59717/j.xinn-mater.2026.100209>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Root causes and consequences of data scarcity in LIB research are critically examined.
- Materials-level and device-level solutions for mitigating data scarcity are systematically presented.
- Future avenues for further reducing the impact of data scarcity across the field are proposed.

AI for Lithium-ion battery research under data-scarce scenarios

Junzhe Zhang,^{1,6} Xiaoang Zhai,^{2,6} Qiang Zhang,^{1,3,6} Peiyi Li,¹ Mian Li,¹ Kai Wu,^{4,*} Yang Liu,^{1,*} and Jiayu Wan^{1,5,*}

¹Future Battery Research Center and Smart Center, Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, China

²School of Computer Science and Technology, Donghua University, Shanghai 201620, China

³School of Energy and Power Engineering, North University of China, Taiyuan 030051, China

⁴Contemporary Amperex Technology Co., Limited, Ninde 352100 China

⁵Key Laboratory of Artificial Intelligent Creation Extreme Energy Materials, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China

⁶These authors contributed equally

*Correspondence: wuk@catl.com (K.W.); liuyang130@sjtu.edu.cn (Y.L.); wanji@sjtu.edu.cn (J.W.)

Received: July 21, 2025; Accepted: March 13, 2026; Published Online: March 13, 2026; <https://doi.org/10.59717/j.xinn-mater.2026.100209>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Zhang J., Zhai X., Zhang Q., et al. (2026). AI for Lithium-ion battery research under data-scarce scenarios. *The Innovation Materials* 4:100209.

The global challenge on energy crises has promoted the extensive research on energy storage, which positions the lithium-ion batteries (LIBs) as a core technology due to the exceptional performance. While conventional approaches for LIB research are limited by long development cycles, high resource demands, and heavy reliance on the expertise of researchers, data-driven methods powered by ML exhibit remarkable potentials on enabling the efficient and accurate predictions of key battery performance metrics, significantly reducing the time and economic costs on the R&D of batteries. Despite these advances, data scarcity remains a major obstacle to the widespread applications of these techniques due to the high costs and prolonged cycles associated with battery experiments. To investigate appropriate approaches for addressing these challenges, a comprehensive review on the data-driven methods for lithium battery research is provided, offering insights into mitigating the impact of data scarcity. To reach this goal, this review systematically examines the implementations of data-driven methods associated with the applications in the battery domain. Detailed studies have been conducted to investigate the root causes of data scarcity at various levels ranging from materials to devices and effective strategies to accommodate these issues. Finally, the review discusses future directions, emphasizing the need for collaborative data-sharing frameworks and adaptive ML models that balance domain expertise with computational innovation. By addressing data scarcity through these strategies, this work aims to accelerate the development of next-generation LIB while retaining the core insights of traditional methodologies.

INTRODUCTION

In the modern energy transition, LIB have emerged as the cornerstone of electrification and renewable energy integration, driven by their unparalleled energy density, extended cycle life, and lightweight design.¹⁻⁵ These attributes not only solidify their dominance in portable electronics and electric vehicles but also underscore their pivotal role in achieving global decarbonization goals.⁶⁻⁹ However, advancing LIB technologies demands continuous breakthroughs in both materials innovation—such as discovering high-capacity cathodes or stable solid electrolytes¹⁰ and device optimization, where accurate state-of-charge prediction and degradation monitoring are critical for safe and reliable battery management systems (BMS).¹¹

Traditional approaches for LIB development rely heavily on empirical models and iterative experimentation, which are inherently constrained by slow iteration cycles,¹² resource-intensive validation,^{9,13-15} and domain-specific expertise.⁸⁻¹¹ While physics-based models provide foundational insights, they often struggle to capture the multi-scale complexity of real-world battery systems, from atomistic interfacial phenomena to cell-level performance decay.¹⁶⁻¹⁸ In contrast, machine learning (ML) has revolutionized materials and device research by automating pattern recognition in high-dimensional datasets, uncovering nonlinear relationships imperceptible to conventional methods (Figure 1).¹⁹⁻²² Yet, the full potential of ML for battery research remains bottlenecked by data scarcity: the prohibitively high cost of generating high-fidelity experimental data, inconsistencies in data collection protocols, and limited access to proprietary industrial datasets.¹¹ Statistically sparse training data exacerbates overfitting, where models fixate on experimental noise other than generalizable electrochemical principles,²³⁻²⁵ulti-

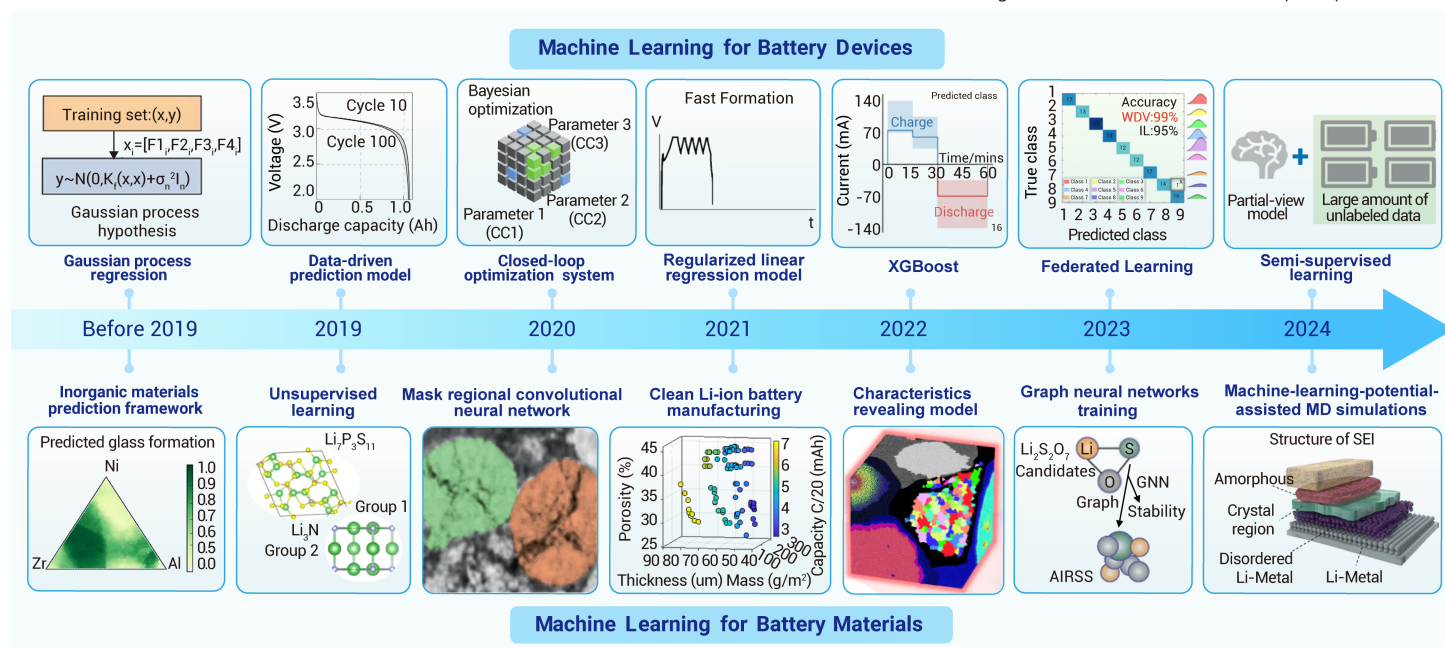


Figure 1. Recent advances of ML assisted methods for LIB research.²⁶⁻³⁹ Copyright 2019, Springer Nature. Copyright 2018, Elsevier. Copyright 2020, Springer Nature. Copyright 2021, Cell Press. Copyright 2022, Springer Nature. Copyright 2023, Springer Nature. Copyright 2024, Cell Press. Copyright 2016, Springer Nature. Copyright 2019, Springer Nature. Copyright 2020, Springer Nature. Copyright 2021, Elsevier. Copyright 2022, American Chemical Society. Copyright 2023, Springer Nature. Copyright 2024, Royal Society of Chemistry.

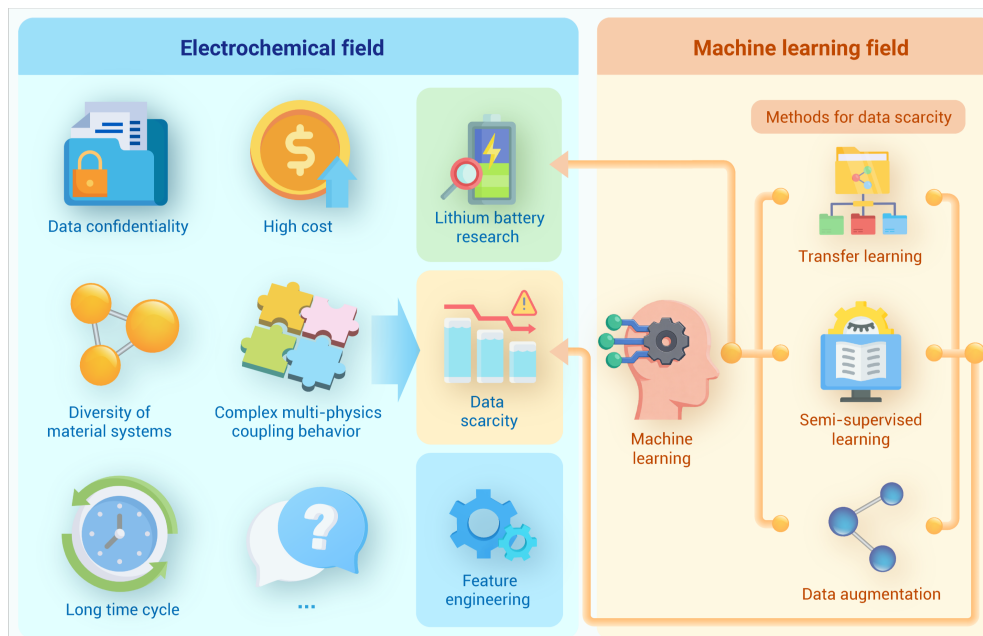


Figure 2. Machine learning methods to tackle data-scarcity issue.

gories, while detection focuses on identifying anomalies like manufacturing defects.^{51,52} The feasibility of these standard tasks is often directly challenged by the pervasive issue of data scarcity.

This challenge manifests in distinct ways, with specific causes compounding the problem for each task type. For regression, the primary cause of data scarcity is the prohibitive time and cost associated with long-term experiments, such as cycling test of batteries.⁵³ Consequently, the resulting datasets are often too sparse to allow models to accurately capture complex, nonlinear degradation pathways. In classification tasks, data scarcity frequently arises from the inherent rarity of critical events, such as thermal runaway or specific internal shorts. This leads to severe class imbalance, where models trained on such

mately undermining their reliability across diverse battery chemistries and operating conditions.

Recent efforts to mitigate data scarcity integrate external knowledge to compensate for limited datasets. Data augmentation techniques synthetically expand datasets by replicating the statistical distributions of experimental measurements.⁴⁰ Semi-supervised learning leverages abundant unlabeled data—often underutilized in battery studies—to reduce dependency on costly labeled examples,⁴¹ while transfer learning adapts pre-trained models from related domains to accelerate LIB-specific tasks.⁴² Despite these advances, existing reviews lack a systematic analysis of the root causes of data scarcity and its interplay with ML workflows in battery research (Figure 2).

This review bridges this gap by systematically analyzing how data scarcity arises at critical stages of LIB development, from materials discovery to device management. By mapping existing strategies to specific challenges, a coherent roadmap is provided for integrating computational innovation with experimental pragmatism, empowering researchers to overcome data scarcity and accelerate the development of next-generation energy storage systems with artificial intelligence (AI).

DATA-DRIVEN METHODS

Data-driven methods, which leverage ML to extract insights from historical data, hold immense promise for advancing LIB research.^{43–45} However, their practical application is critically hampered by the pervasive challenge of data scarcity. To deconstruct this bottleneck, the standard ML workflow is examined and summarized as a unified roadmap in Figure 3, which serves as the structural backbone of this study. This workflow consists of six key stages: task definition, data collection, data preprocessing, feature engineering, modeling, and deployment. Unlike in data-rich domains, each stage in battery research is profoundly affected by the limited availability of high-quality data.^{46,47} In the following subsections, the review adopts this roadmap as a framework to systematically analyze how data scarcity impacts the entire pipeline. The analysis will focus on identifying the root causes of data scarcity at its source, examining its adverse effects on model development at each stage, and reviewing the specialized strategies designed to overcome these limitations.

Task definition

Before applying ML to facilitate LIB research, it is important to clearly define the research objectives into ML tasks. The common tasks in Lithium battery research can generally be categorized into regression, classification, and detection. Regression tasks, such as battery lifetime prediction, aim to forecast continuously distributed values.^{49,50} Classification is usually used for tasks such as fault diagnosis by projecting input data into predefined cate-

data become biased towards predicting the common healthy state and fail to identify rare but crucial faults.⁵² Detection tasks face a dual challenge where data becomes scarce due to both the natural infrequency of manufacturing defects and the labor-intensive process of expert annotation required to label them. This combination of limited data samples and high defect variability makes it difficult for models to generalize across different defect types or imaging conditions.⁵⁴

To address these issues, a plausible approach is to redefine the task to better exploit the available data. For example, binary classification with scarce positive examples can be reframed as an anomaly or one-class detection task, whereby a model is trained on abundant negative samples and tasked with identifying deviations.⁴¹ Similarly, a regression task with sparse target coverage can be transformed through discretizing the continuous prediction target into interval bins, converting the problem into a more manageable coarse-grained classification task.⁵⁵ In object detection, few-shot and weakly supervised approaches can be employed to leverage auxiliary data or meta learning strategies for rare classes.⁵⁶ Such task redefinitions can enable the application of specialized algorithms and inductive biases that improve robustness and generalization under data-scarce scenarios, thereby enhancing overall modeling efficacy.

Data collection

In battery research, data collection spans two complementary levels: material level and device level. At the material level, characterization techniques such as X-ray diffraction, electron microscopy, and spectroscopy provide essential insights into structural, morphological, and chemical properties.⁵⁷ At the device level, long-term cycling experiments yield time-resolved voltage, current, and temperature data, which are critical for understanding performance degradation and safety.⁵⁸ The processes for data acquisition are the primary sources of the data-scarcity challenge.⁵⁹ Conventional material characterization often requires expensive, specialized equipment and intricate sample preparation, severely limiting throughput and the amount of data required for ML tasks. Similarly, device-level cycling tests can take months to complete for a single batch, with inherent cell-to-cell variability necessitating costly and time-consuming replications.⁶⁰ Moreover, acquiring high-fidelity electrochemical data necessitates advanced battery cyclers capable of precise current and voltage control, coupled with rigorously thermostatically controlled environments.⁶¹ These stringent experimental requirements impose significant economic burdens on the data acquisition process itself. These fundamental constraints of high cost, long duration, and required expertise mean that comprehensive, large-scale datasets, especially for emerging chemistries, are inherently rare.⁶²

Quantitatively, at the materials level, large public databases may list around

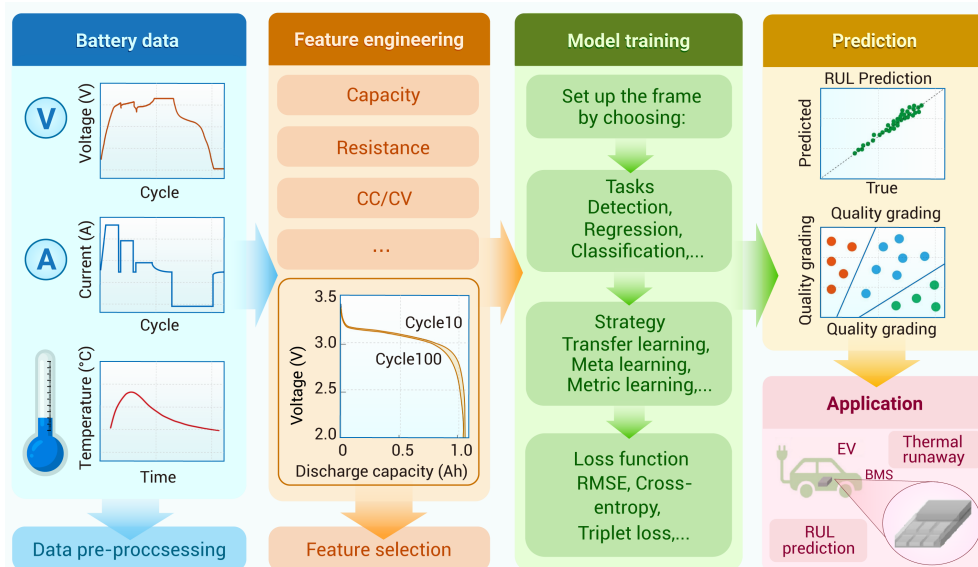


Figure 3. Workflow of data-driven methods for LIB research.²⁶⁻⁴⁸ Copyright 2024, Wiley-VCH. Copyright 2019, Springer Nature. Copyright 2024, Elsevier.

generate extra examples for rare classes (e.g., specific fault types) by interpolating between existing minority-class samples, thereby creating more balanced datasets for training.⁶⁶⁻⁶⁹

Pre-processing

As the data collected from practical scenarios are frequently plagued by non-idealities, pre-processing is essential to ensure data quality and consistency, whose importance is even more magnified under conditions of data scarcity. Data acquisition processes in real-world applications are susceptible to issues such as equipment malfunctions, environmental or signal interference, and data transmission or storage faults, thereby leading to the

loss of data (resulting in missing values) and the generation of anomalies, directly jeopardizing the performance of prediction models.⁷⁰ While problematic data points can often be discarded with minimal impact in data-rich scenarios, every data sample is valuable in data-scarce context, and discarding data exacerbates the scarcity issue.⁷¹ However, retaining corrupted data can introduce systematic biases that mislead the learning process, in which scenario models will overfit to noisy information in contrast to capturing true electrochemical principles.

Therefore, pre-processing facilitates both data cleaning and intelligently information recovery in battery research. For this purpose, common practice like smoothing and normalization are applied to time-series data to filter out random fluctuations and ensure fair feature comparisons, highlighting the underlying degradation trends.^{72,73} For image data, denoising is critical for accurate defect identification. Beyond simple data cleaning, data imputation techniques take on a more significant role. These methods reconstruct missing values by exploiting useful patterns in existing data. For instance, in time-series data, interpolation can fill gaps in cycling curves based on adjacent measurements, effectively "repairing" the dataset without the cost of re-running experiments. The recovery of missing information can be regarded as a form of low-cost data augmentation, as it increases the quantity of usable, continuous data available for training.

Additionally, a critical trade-off that must be handled during pre-processing is that overly aggressive filtering or smoothing can inadvertently remove subtle but critical features indicative of early-stage degradation or impending failure. For instance, a small voltage anomaly can signal the emergence of an internal short, which can easily be mis-detected as noise and smoothed away. Therefore, a potential solution is to apply domain knowledge as a guidance of pre-processing methods, thereby selecting appropriate algorithms and parameters to preserve genuine signals while suppressing noise. Ultimately, in the context of data scarcity, meticulous pre-processing is not merely a preliminary chore but a strategic necessity for maximizing the value of every collected data point and building robust, generalizable models.

To overcome these data-collection bottlenecks, a potential solution is to shift from exhaustive, conventional methods to intelligent, data-efficient strategies. For instance, active learning directly tackles the issue of expensive labeling and experimentation through iteratively identifying the most informative data samples that can diversify the training dataset instead of passively data collection, enabling the training of high-performance models with significantly fewer data samples and thus reducing experimental costs and time.^{63,64} Importantly, the feasibility of active learning has also been substantiated in industrially relevant scenarios. The framework proposed by Zhu et al.⁶⁴ was validated using laboratory-scale lithium-ion battery degradation tests that closely resemble industrial aging processes. These tests explicitly account for the high electricity consumption from repeated charge-discharge cycling and the operation of temperature chambers, which dominate the cost structure of large-scale industrial testing. The results demonstrate that active learning can substantially reduce the usage of the battery aging data by up to 40% for exhaustive run-to-failure experiments while maintaining equivalent accuracy, thereby confirming its applicability in practical industrial collaborations rather than being limited to purely theoretical or synthetic settings. A complementary strategy is data augmentation, which expands limited datasets using synthetic data. Generative models like generative adversarial networks and variational autoencoders can learn the underlying pattern of real battery data—such as cycling curves or impedance spectra—and generate synthetic data samples that are statistically realistic.⁶⁵ These synthetic data can be used to enrich training sets, improving model robustness. For classification tasks plagued by imbalanced data, over-sampling techniques like synthetic minority over-sampling technique can

loss of data (resulting in missing values) and the generation of anomalies, directly jeopardizing the performance of prediction models.⁷⁰ While problematic data points can often be discarded with minimal impact in data-rich scenarios, every data sample is valuable in data-scarce context, and discarding data exacerbates the scarcity issue.⁷¹ However, retaining corrupted data can introduce systematic biases that mislead the learning process, in which scenario models will overfit to noisy information in contrast to capturing true electrochemical principles.

Therefore, pre-processing facilitates both data cleaning and intelligently information recovery in battery research. For this purpose, common practice like smoothing and normalization are applied to time-series data to filter out random fluctuations and ensure fair feature comparisons, highlighting the underlying degradation trends.^{72,73} For image data, denoising is critical for accurate defect identification. Beyond simple data cleaning, data imputation techniques take on a more significant role. These methods reconstruct missing values by exploiting useful patterns in existing data. For instance, in time-series data, interpolation can fill gaps in cycling curves based on adjacent measurements, effectively "repairing" the dataset without the cost of re-running experiments. The recovery of missing information can be regarded as a form of low-cost data augmentation, as it increases the quantity of usable, continuous data available for training.

Additionally, a critical trade-off that must be handled during pre-processing is that overly aggressive filtering or smoothing can inadvertently remove subtle but critical features indicative of early-stage degradation or impending failure. For instance, a small voltage anomaly can signal the emergence of an internal short, which can easily be mis-detected as noise and smoothed away. Therefore, a potential solution is to apply domain knowledge as a guidance of pre-processing methods, thereby selecting appropriate algorithms and parameters to preserve genuine signals while suppressing noise. Ultimately, in the context of data scarcity, meticulous pre-processing is not merely a preliminary chore but a strategic necessity for maximizing the value of every collected data point and building robust, generalizable models.

Feature engineering

In battery research, raw data often span a high-dimensional feature space.⁷⁴ For instance, a single impedance spectrum can contain hundreds of data points, and the time-series of cycling data can encompass thousands of measurements over the lifespan of a battery. When training data is limited, the high dimensionality can drastically increase the "curse of dimensionality" risk, such that a model may end up learning random noise other than the useful underlying signal, leading to severe overfitting and poor generalization. Therefore, feature engineering is a critical step acting as a powerful regularization strategy to mitigate this risk by reducing dimensionality. Another important effect of feature engineering is to enable efficient learning through directly embedding domain-specific knowledge into data, thereby uncovering highly complex underlying correlations.⁷⁵

For aforementioned reasons, feature engineering generally follows two complementary steps: feature extraction and feature selection. Feature extraction methods transform the original high-dimensional data into a smaller set of informative and explicit features. A classic example is principal component analysis (PCA), which identifies orthogonal components that capture the maximum variance in the data. In the context of battery aging, PCA can distill complex degradation phenomena—reflected across numerous voltage or impedance data points—into a few latent variables representing dominant aging modes such as loss of lithium inventory or loss of active material.⁷⁶ This not only reduces model complexity but also enhances its ability to focus on the most useful patterns, improving the robustness against the noise prevalent in limited datasets. Furthermore, kernel-based extensions of classic approaches like Kernel PCA can be employed to capture more complex nonlinear relationships.

In parallel to feature extraction, feature selection aims to identify and retain the most predictive subset of features while discarding redundant or irrelevant ones, which is crucial for building interpretable and robust models from limited data. This is particularly important in battery research, where features derived from domain knowledge often hold more predictive power than raw data points. For example, features like the duration of specific voltage plateaus (e.g., the dQ/dV peak area) can serve as highly effective proxies for capacity fade or state of health (SOH) in contrast of the entire discharge curve.⁷⁷ Methodologically, feature selection can be implemented in several ways. Filter methods use statistical metrics to measure the importances of features, which is independent to the specific model, and remove the less critical ones. More sophisticated wrapper and embedded methods, such as recursive feature elimination or tree-based models (e.g., random forests and XGBoost), select features based on their contribution to the performance of a specific model.⁷⁸ These model-based approaches can uncover non-obvious feature interactions, while their reliability on small datasets must be validated carefully, often through cross-validation, to ensure the generalization ability of selected features.

In summary, both feature extraction and selection are indispensable for building reliable models from limited battery data. They transform a high-risk, high-dimensional problem into a more manageable and robust version by ensuring the model learning from physically meaningful and information-dense representations, thereby avoiding getting lost in the noise of a sparse, high-dimensional space.

Modeling

The challenges associated with the selection and training of ML models are multifaceted in data-scarce environments, which involves the appropriate fitting to the context of data, the control of model complexity, and the avoidance of overfitting risks, especially when the availability of data is limited. In addition to model selection, another critical issue is the applications of key strategies—from regularization to advanced training paradigms—to enable robust learning from limited battery data.

A wide range of ML models, from simple linear models to complex deep neural networks (DNNs), have been adopted for battery research due to their distinct advantages in data-scarce scenarios.⁷⁹ Within these models, linear models are often the preferred starting point due to the low complexity, strong interpretability, thereby offering the transparency to analyze the electrochemical principles simultaneously with data-based observations and reducing the risk of overfitting.^{80–83} Ensembled decision tree models (i.e., random forest and gradient boosting machines such as XGBoost, LightGBM, and CatBoost), further enhance the accuracy and robustness by aggregating predictions from multiple decision trees, which automatically accommodate nonlinear relationships and missing data accommodation.^{84,85} In data-scarce scenarios, their performance can be further improved through techniques like pruning, while their ability to provide feature importance scores also makes them highly valuable for interpretable battery health diagnosis.⁸⁶ Despite these advantages, their reliance on handcrafted features limits their ability to capture intricate degradation dynamics present in raw, high-dimensional data.

On the other end of the spectrum, DNNs offer unparalleled capabilities in automatic representation learning through accommodating diverse structures of input data using flexible and customized architectures.^{87,88} For

instance, the family of convolutional neural networks (CNNs) are adept at processing grid-like data, such as microscopic images of electrode materials, to identify structural defects.^{89,90} For time-series data, recurrent neural networks (RNNs) and their advanced variants, long short-term memory (LSTM) models, are designed to capture temporal dependencies, though they often struggle with long-range effects.^{91,92} More recently, transformers, with innovative self-attention mechanism, have shown great promise in modeling long-term battery degradation.⁹³ However, the high expressivity of these models is a double-edged sword: without sufficient data, their large parameter spaces make them highly prone to memorizing noise, thereby resulting in overfitting. Therefore, successfully deploying DNNs in data-scarce scenarios highly relies on the sophisticated training strategies.

To bridge the aforementioned gap and bolster model generalization, various techniques are employed, where regularization is a primary defense against overfitting. It works by involving a penalty to the loss function, thereby constraining model complexity through imposing a form of prior knowledge. Standard methods like Lasso and Ridge regularization achieve this by encouraging smaller or sparser model weights through incorporating Laplacian and Gaussian prior to their distributions.^{94,95} For neural networks, a specific technique widely used to avoid overfitting is dropout, which randomly deactivates a fraction of neurons during training, preventing complex co-adaptations and forcing the network to learn more robust features.⁹⁶

More advanced regularization approaches directly embed domain knowledge to compensate for data sparsity. For example, causal regularization, such as causal structure learning, aims to learn the underlying causal graph between variables, helping the model distinguish true causal relationships from spurious correlations and thus improving its robustness to distributional shifts between laboratory and real-world conditions.⁹⁷ Similarly, physics-informed neural networks (PINNs) incorporate physical laws (e.g., electrochemical equations from a Doyle-Fuller-Newman model) as a soft constraint in the loss function. This guides the model to produce physically plausible predictions even in regions with no training data, making it a powerful tool for fusing data-driven methods with first-principles understanding.⁹⁸ And Beyond PINNs, physics-informed Gaussian process regression (GPR) offers another concrete route to knowledge-driven model construction in LIB research. For instance, a recent physics-informed multiview collaborative semisupervised framework for early lifespan prediction first extracts mechanistic aging indicators such as loss of lithium inventory via electrode-level open-circuit potential analysis.⁹⁹ Then, the framework augments them with phase-transition descriptors obtained from Lorentzian-decomposed incremental capacity curves and operating-condition features (temperature, resistance growth and charging time). These views are coupled to dedicated physics-informed kernels (e.g., Arrhenius-type temperature kernels and Matérn kernels for resistance and charge-time evolution), so that the GPR prior is explicitly biased toward electrochemically plausible degradation pathways while still being trained on a small labeled set together with abundant unlabeled early-cycle data. In this way, the model regularization is no longer purely statistical but directly reflects known aging mechanisms, leading to improved extrapolation, tighter uncertainty quantification and physically interpretable predictions of battery lifespan.

Beyond penalizing complexity, advancing training strategies is crucial for maximizing the utility of data. As a powerful tool in battery research, transfer learning adapts external data or pre-trained models to the target domain with limited data, leveraging existing knowledge of battery degradation patterns to dramatically reduce the data demanded for new tasks.¹⁰⁰ Semi-supervised learning offers another avenue by utilizing large amounts of unlabeled data—which is often more convenient to obtain—to help guide the learning process on a small labeled dataset, for instance, by learning a better feature representation before supervised fine-tuning.¹⁰¹ Finally, reliable model evaluation and hyperparameter tuning in low-data settings are critically dependent on robust validation schemes. k-fold cross-validation is standard for obtaining a stable estimation of generalization performance, while variants like leave-one-out cross-validation are particularly useful for extremely small datasets.¹⁰² These methods, often guided by efficient hyperparameter searching algorithms like Bayesian optimization (BO), ensure that model performance is genuine and not an artifact of a lucky train-test split.¹⁰³ By synergizing an appropriate model selection with advanced regularization and

training techniques, it is possible to build reliable, high-performance models even when faced with the fundamental challenge of data scarcity.

Deployment

Deploying ML models in real-world LIB systems, such as an on-board BMS, is often impacted by the challenge of "lab-to-field gap", which is characterized by the limited computational resource and complex operation environment. Due to the limited computational resources of embedded devices, relatively large and accurate laboratory models can hardly generate timely on-device inference in an energy-efficient manner. Furthermore, models trained on clean, controlled lab data often fail to generalize to the noisy and dynamic conditions of real-world operation, which is known as the domain shift issue. These challenges can be magnified if the real-world labeled data is insufficient. To address these issues, two complementary strategies, transfer learning and knowledge distillation, offer a powerful toolkit to overcome these deployment hurdles.

Transfer learning is employed to address the challenge of data-efficient domain adaptation. Instead of training a model from scratch in application scenarios that requires extensive real-world data, it leverages a robust model pre-trained under laboratory conditions and fine-tunes it using a relatively small amount of labeled data from the application environment.¹⁰⁴ The benefit of this approach is two-folded, such that it offers a cost-efficient solution in deployment scenarios where obtaining labeled operational data is impractical or costly, while allowing the model to rapidly adapt to new operational dynamics, ensuring continued accuracy without the need for complete retraining.

Knowledge distillation primarily mitigates the computational footprint for resource-constrained hardware. It works by compressing a large, high-performing "teacher" model into a compact and efficient "student" model that can run in real-time on a BMS. It is worth noting that its effectiveness in data-scarce scenarios is enabled by the transfer of "dark knowledge". Other than training the student on sparse, hard labels (e.g., "good" or "bad"), it is trained on the teacher's soft outputs, such as class probabilities. This provides a much richer, more nuanced supervisory signal that encodes the teacher's "reasoning" and inter-class similarities.¹⁰⁵ This rich signal acts as a powerful regularizer, enabling the student model to learn more effectively and achieve higher performance than training from scratch on the same small, hard-labeled dataset.

In summary, these two strategies form a synergistic pair for successful deployment of ML models in practical battery research scenarios, where transfer learning adapts the model to the rapid dynamic of application environment with minimal new data, and knowledge distillation enables efficient inference on resource-constrained devices. The joint utilization of them will further provide a robust pathway for developing compact, context-aware, and reliable models that can be effectively deployed in real-world LIB systems.

AI AND DATA-SCARCITY IN BATTERY MATERIALS

In LIB research, ML has emerged as a critical tool for improving material performance and accelerating the discovery of new materials.¹⁰⁶ However, its application in material development is fundamentally constrained by data scarcity, a challenge rooted in the high cost, long duration, and inherent complexity of both experimental and computation processes. This section delves into how ML addresses data scarcity across three critical stages of material development: materials discovery, material characterization, and battery manufacturing.

AI assisted material discovery under data-scarcity

Material simulation methods like density functional theory (DFT), molecular dynamics (MD), and the finite element method (FEM) provide foundational insights into material properties.^{107,108} However, their high computational cost is a primary source of data scarcity for ML applications. Generating large, diverse datasets required to train robust ML models by running these simulations across vast chemical or structural spaces is often computationally expensive.¹⁰⁹ For instance, DFT is computationally demanding, which limits its use in large scale simulations, while traditional MD relies on pre-defined force fields that may lack accuracy for novel systems. This scarcity of high-fidelity simulation data severely restricts the ability of ML models to learn

complex structure-property relationships and generalize to unseen materials.

To overcome this bottleneck, a powerful strategy has emerged: integrating ML directly with physics-based simulations. This approach mitigates data scarcity in two primary ways: by using ML to accelerate simulations, thereby generating more data, or by using ML to intelligently guide a limited number of expensive simulations. One prominent example is the use of ML interatomic potentials (ML-potentials) to accelerate MD simulations. As shown in Figure 4A, Xu et al. developed a moment tensor potential (MTP) to study polyanion rotation in solid electrolytes.¹¹⁰ The MTP, trained on a small set of accurate DFT calculations, learns the potential energy surface and allows for ultralong-time MLMD simulations. This approach extends the simulation timescale from picoseconds to the microsecond level, enabling the generation of extensive trajectory data to investigate rare events like Li^+ diffusion—data that is inaccessible with traditional *ab initio* MD due to time constraints, thus can effectively overcome data scarcity.

Another strategy is ML-guided discovery, which uses ML to efficiently screen candidate materials before committing to costly simulations. In Figure 4B, Sendek A D et al. aims to accelerate the discovery of solid LIB conductors for all-solid-state batteries through ML.¹¹¹ Trained on a sparse dataset of only 40 experimental data points, their logistic regression model used a set of structural descriptors to predict superionic conductivity. This ML model then guided the selection of candidates for DFT-MD validation, revealing 8 new fast conductors from ML-guided selections versus only 3 from random screening.

Furthermore, ML can be engineered to learn effectively even when simulation data is exceptionally scarce. Chouchane M Xu et al. aimed to enhance the rate capability of dry-processed thick nickel manganese cobalt oxide (NMC) 811 electrodes through a coupled FEM and ML approach as shown in Figure 4C.¹¹² To address data scarcity in 3D electrode modeling, researchers made full use of data by treating individual particles as data points rather than entire simulations. This strategy enabled effective model training with only 10 FEM-generated electrode configurations, achieving a 7.3% average relative error while reducing computation time by two orders of magnitude.

These combined methods effectively use the foundational physical and chemical principles to supplement and enhance the quality of the dataset. By integrating ML with computational methods like DFT, MD, and FEM, high-quality data can be generated more efficiently or used more effectively. Further extending this integration of physics and ML, related approaches like PINNs take a more direct route by incorporating physical laws (e.g., differential equations) as soft constraints directly into the loss function. This embeds physical principles as strong prior knowledge, preventing data-driven models from making unphysical predictions and improving their generalization from limited experimental data. Additionally, this integration provides physical insights that allow researchers to understand and validate predictions, reducing reliance on "black-box" models.¹¹⁴

AI assisted materials characterization under data-scarcity

In the characterization of LIB materials, ML offers powerful tools for automated image analysis, but its effectiveness is often hampered by data scarcity. This scarcity manifests in two primary forms: the limited availability of raw image data and, more critically, the lack of extensive, high-quality annotated data.

The first challenge, a scarcity of raw images, arises because image-based characterization relies on advanced and expensive instruments like electron microscopes, synchrotron sources, or computed tomography (CT) scanners. The high operational costs and slow acquisition speeds inherently limit the volume of data that can be collected. Furthermore, due to the diversity and complexity of images in material characterization, standardized image datasets are relatively scarce. Additionally, the experimental data obtained during material characterization are often high-dimensional and complex, making manual analysis extremely challenging and leading to a lack of sufficient labeling and structured information. The second and more pervasive challenge is the scarcity of annotated data. Raw images are often high-dimensional and complex, requiring expert-driven, labor-intensive manual analysis to label key features like phase fractions, particle sizes, or defects. This annotation process is a significant bottleneck, resulting in datasets with

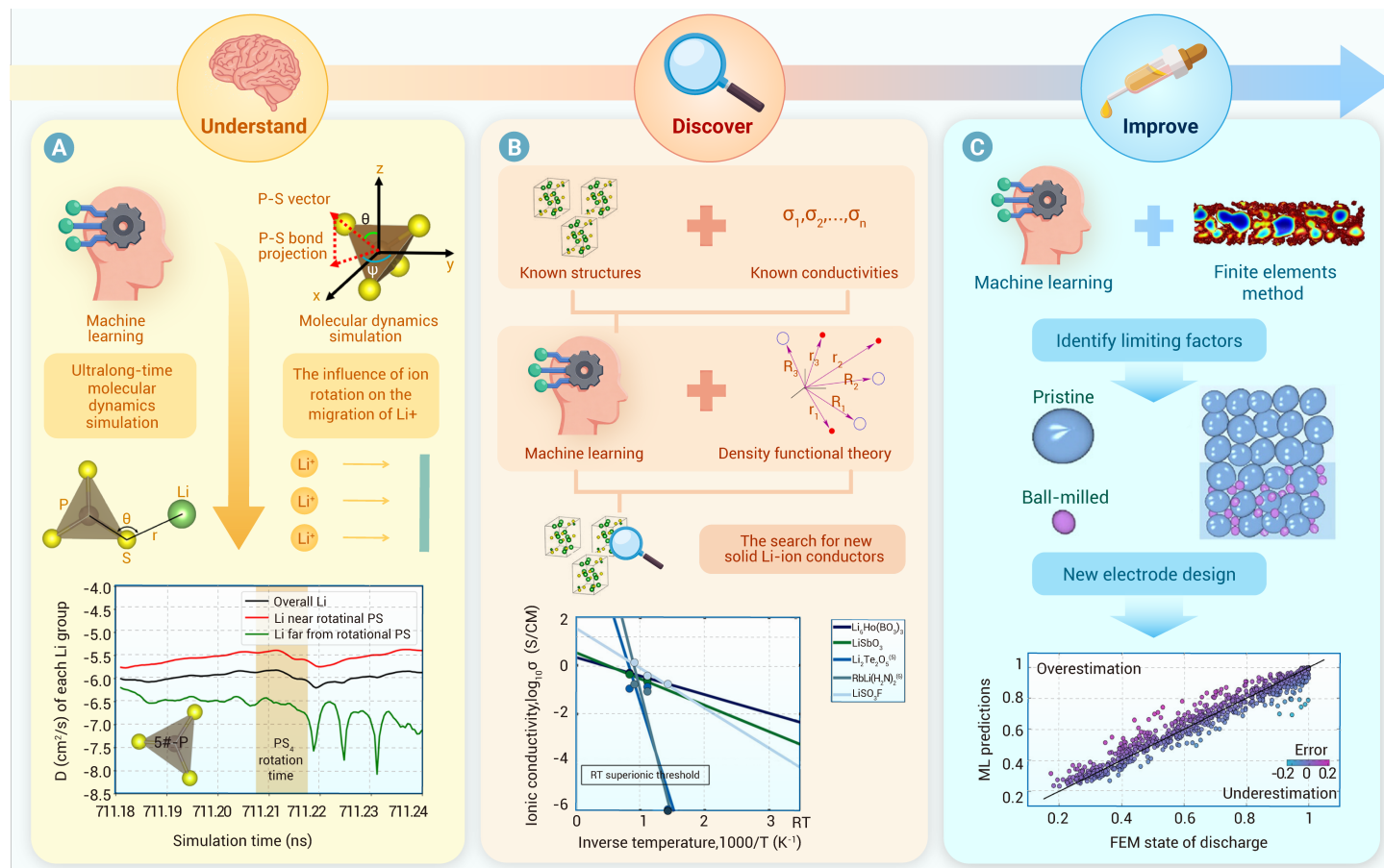


Figure 4. ML assisted MD, DFT and FEM (A) MD simulations using ML interatomic potentials extend timescales from picoseconds to microseconds while preserving DFT accuracy.¹¹⁰ Copyright 2023, Springer Nature. (B) ML predicts lithium-ion conductivity, identifies potential materials, and validates results through DFT-MD simulations.^{111,113} Copyright 2018, American Chemical Society. (C) ML assisted FEM creates a performance model for thick electrodes, identifies limitations, and validates structured bilayer designs via simulation and experiments.¹¹² Copyright 2024, American Chemical Society.

very few labeled examples available for training supervised ML models.

To address the bottleneck of manual annotation, ML-powered image segmentation has become an indispensable tool. As shown in Figure 5A, Jiang et al. investigated the particle-carbon/binder domain detachment in battery cathodes.³⁶ They employed a Mask R-CNN model to automate the segmentation of over 650 particles from advanced imaging data, a task that would be prohibitively time-consuming if done manually. This automated approach not only dramatically increases the throughput of analysis but also extracts quantitative microstructural parameters (e.g., detachment degree, volume) from the raw images. By transforming unlabeled pixels into meaningful, structured data, ML segmentation effectively alleviates the annotation scarcity and enables robust statistical analysis.

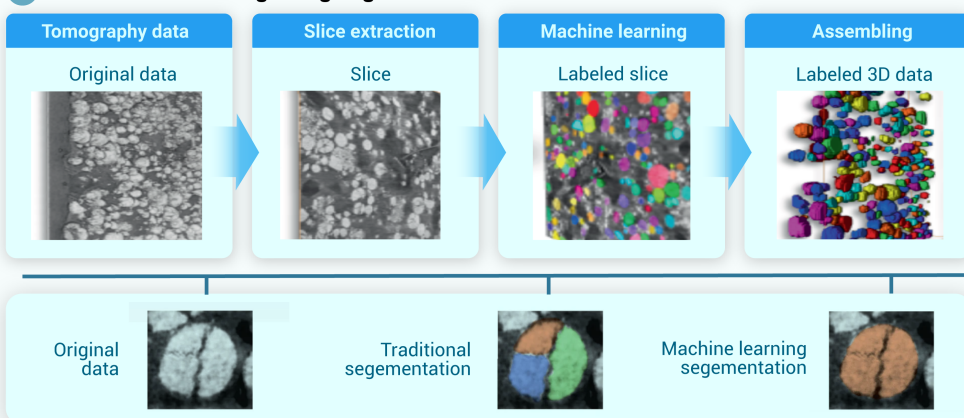
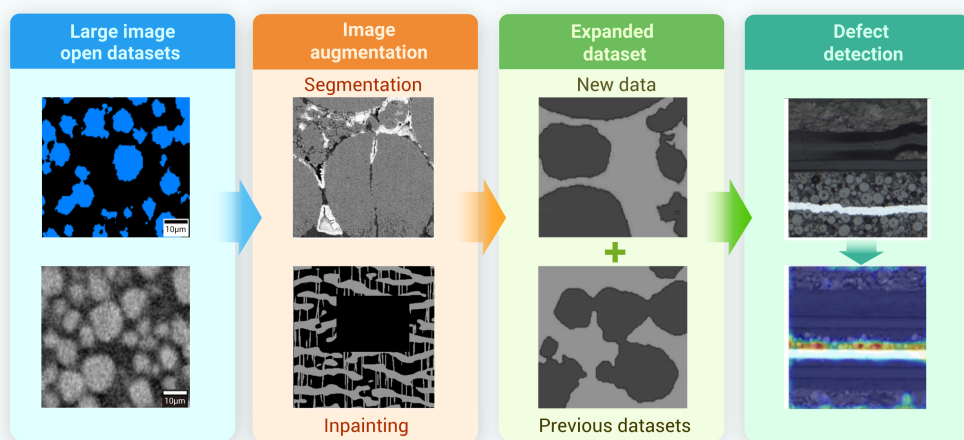
However, even with automated tools, the initial model training still requires some annotated data, which might be extremely limited.¹¹⁵ To overcome this fundamental data and label scarcity, data augmentation is a widely adopted solution. Figure 5B shows a workflow that the data augmentation is used in defects detection. Data augmentation artificially expands the training set by applying a series of transformations—such as rotation, flipping, scaling, and color variations—to the original annotated images.¹¹⁶ In material characterization, this is particularly valuable for microscopy or X-ray images, as it can simulate variations in experimental conditions like sample orientation, lighting, or imaging noise. By generating diverse, synthetic training samples, data augmentation helps the model learn more robust and generalizable features, mitigating overfitting and improving its performance even when the initial annotated dataset is small.

AI assisted manufacturing under data-scarcity

The increasing demand for efficient conversion from laboratory research towards mass production creates an opportunity for the introduction of

ML.^{119–121} However, applying ML to manufacturing process, such as the electrode preparation shown in Figure 6A, faces challenges from both the high cost of advanced sensors and inherent process variability. First, obtaining precise real-time process data, such as temperature, humidity, and pressure during electrode manufacturing, requires advanced sensors and monitoring systems. Such technological infrastructure is typically available on a limited number of high-end production lines, leading to a restricted volume of high-quality data. Second, variations in material batches, equipment performance fluctuations, and process instabilities introduce inconsistencies in production data. This process variability means that data from one batch may not be representative of another, fragmenting the available data into small, inconsistent sets and making it difficult for ML models to learn generalizable process-property relationships.

To overcome these challenges, several data-efficient ML strategies have been developed. A key approach is to use data clustering to consolidate fragmented datasets. By grouping production data with similar characteristics, clustering helps identify underlying patterns that are consistent across different batches, despite superficial variations. For instance, as shown in the workflow in Figure 6B, Tan et al. proposed a data-driven approach that begins with data generation and clustering.¹²² Clustering is used to group intermediate product properties from different experiments, revealing correlations that might be obscured by batch-to-batch noise. This allows for the consolidation of data from disparate sources into a more robust dataset, thereby directly addressing the issue of insufficient and inconsistent data from individual experiments. Subsequently, ML models can be trained on this enriched dataset to predict final electrode properties. The workflow concludes with a critical post-modeling stage, which employs sensitivity analysis and 3D output representation to interpret the model's findings. This final step is vital in data-scarce contexts, as it allows researchers to understand the influence

A Detachment revealing using segmentation**B Electrode defect detection**

of various process parameters on final properties, thereby guiding future, more targeted and efficient experimentation.

Another powerful strategy is to enhance data efficiency through intelligent feature selection, ensuring that the model learns from the most informative signals within a limited dataset. As illustrated in Figure 6C, Liu et al. aimed to predict LIB electrode mass load and quantify feature importance in production processes using ML.¹²³ They employed a gaussian process regression (GPR) framework with an automatic relevance determination (ARD) kernel to analyze four key process features. The ARD mechanism automatically identified the most predictive features and assigned them higher importance weights, while down-weighting or eliminating less relevant ones.¹²⁴ By eliminating low-weight features, training time reduced by 9.4% with minimal accuracy loss, demonstrating effective data utilization. This approach maximizes the value of scarce data by focusing the model on the most critical process parameters, leading to robust and accurate predictions even with limited information.

AI AND DATA-SCARCITY IN BATTERY DEVICES

Lithium battery device research centers on critical operational parameters such as remaining useful life (RUL), SOH, state of charge (SOC), thermal management, and fast-charging protocol optimization. While ML can extract hidden patterns from heterogeneous operational data to address the limitations of conventional analysis, its application at the device level faces significant data scarcity challenges. The primary sources of this scarcity are the prohibitive costs and long durations of lifetime testing, which limit the availability of high-fidelity datasets capturing full degradation cycles and rare failure modes. Furthermore, proprietary data silos and inconsistent measurement standards further impede cross-scenario generalization.

ML for lifetime management

Predicting battery lifetime, a cornerstone of device management, is fundamentally challenged by data scarcity. In laboratory settings, this scarcity

Figure 5. ML image segmentation and data augmentations in material characterization (A) Mask R-CNNs automatically identify and segment NMC particles in 3D volumes to enable quantitative analysis of detachment, volume, and relative intensity characteristics.³⁶ Copyright 2020, Springer Nature. (B) New data generated by data augmentation combines with existing datasets to create larger training sets for model development.^{37,117-118} Copyright 2022, American Chemical Society. Copyright 2023, Elsevier. Copyright 2020, Springer Nature.

stems from lengthy and costly aging tests, while in real-world applications, it is exacerbated by the rarity of failure events, making it difficult to collect sufficient fault data.¹²⁶ To address these challenges, several key ML strategies have been developed, which can be categorized by how they leverage or compensate for limited data.

In many scenarios, unlabeled operational data is abundant and cheap to collect, while labeled data is scarce. Semi-supervised learning leverages this large amount of unlabeled data alongside a small amount of labeled data for training. As shown in Figure 7A, Guo et al. address data scarcity in lifetime prediction using a semi-supervised method called partial Bayesian co-training (PBCT).³² PBCT employs a dual-model framework: a "partial-view" model, trained on easily accessible features from a few labeled samples, generates pseudo-labels for the vast unlabeled data. A more comprehensive "complete-view" model then integrates all features and co-trains with the pseudo-labeled data. Quantitatively, PBCT matches the accuracy

of traditional methods while using 26.9% fewer labeled samples. This efficiency reduced the experimental cost for data acquisition by 2,836.5 cycles compared to the conventional Lasso method, resulting in an economic saving of \$4,685.1. The experimental and economic costs of other representative data-driven techniques have also been analyzed following the computational logic provided by Ref.³², as summarized in Table 1. In specific, the experimental cost is computed as the total number of tested cycles contained in the training data (both labeled and unlabeled) required to attain a common accuracy target (RMSE = 0.1 log₁₀ (cycle), error reported on normalized data), which is converted to economic cost based on the rental expense of battery testing equipment. It can be observed from Table 1 that the ML methods specifically designed for data-scarce scenarios can achieve remarkable reductions on experimental and economic cost compared with conventional techniques.

When only a small set of labeled data is available, strategies that maximize its value are critical. Transfer learning mitigates data scarcity by transferring knowledge from a data-rich source task to a data-scarce target task. As illustrated in Figure 7B, Lin et al. proposed an approach where prior knowledge from a source battery group (calendar aging) was transferred to predict the lifetime of a target group (full life data).⁴² They developed two methods: Bayesian model fusion (BMF) and weighted orthogonal matching pursuit (W-OMP). Experimental results across LIB datasets demonstrate significant improvements, with W-OMP achieving a 41% reduction in root mean square error (RMSE) compared to baseline models. Beyond that, the Shapley additive explanation (SHAP) analysis is employed to ensure the explainability of prediction models by quantifying the contribution of each input feature to the prediction outcome. For instance, Lin et al. utilized SHAP to discover important features, such as the discharge capacity of cycle 2, as a most indicator that describes the battery aging state between datasets in a straightforward manner, which provides useful insights to guide the next generation battery designs.

Apart from the improvement of prediction accuracy in data-scarce scenar-

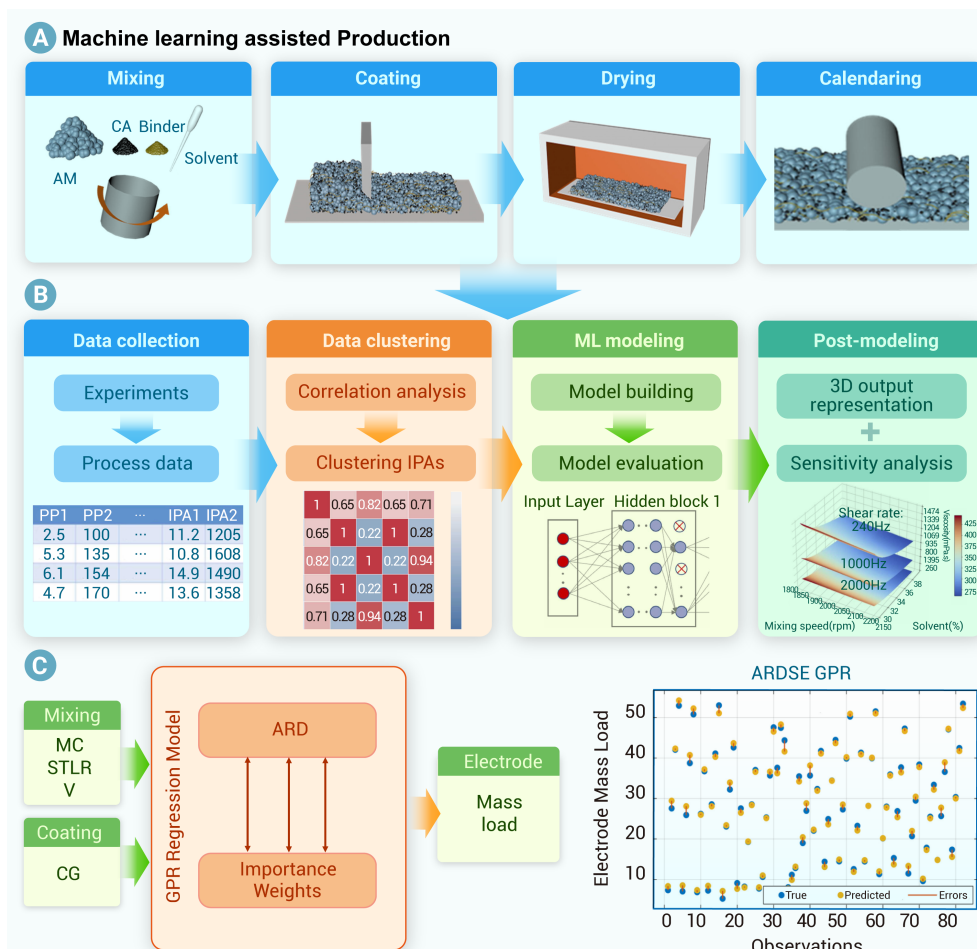


Figure 6. ML involved production workflow (A) Four general steps of electrode production.¹²⁵ Copyright 2022, Elsevier. (B) Experimental or simulation data are clustered to group intermediate product attributes, then modeled, evaluated and post-modeling analyzed using ML.¹²² Copyright 2023, Elsevier. (C) A ML framework based on GPR with an ARD kernel analyzes feature importance and predicts electrode mass loading in LIB.¹²³ Copyright 2021, Elsevier.

reduces required training samples by 75% compared to conventional methods while maintaining < 1% estimation errors across different battery chemistries.

While ensemble methods address the challenge of model generalization from small datasets, real-world data is often not just small but also imperfect. It frequently suffers from missing values or measurement noise, which, if not handled properly, can lead to the loss of already scarce and valuable samples. To tackle this, data imputation methods "repair" incomplete data. As shown in Figure 7C, Zhai et al. introduced a ML framework leveraging the multiple imputation chained equations (MICE) method.¹²⁹ MICE cyclically predicts missing values for each feature based on the others, creating multiple plausible complete datasets. On a challenging test set where 80% of the samples were incomplete with a 60% missing rate, this approach achieved an RMSE of 155.1 cycles, a 23.2% reduction versus using only the complete data.

Beyond missing data, addressing measure-

ment noise and uncertainty is critical. To handle dimension-variant measurement errors, Yu et al. introduced a framework that combines the expectation-maximization (EM) and total-least-squares (TLS) methods, referred to as the EM-TLS framework.¹³⁰ By iteratively estimating noise distributions, it achieves a 29.9% lower RMSE than standard TLS under high noise. Complementarily, Lu et al. tackled overfitting under noisy conditions by integrating TLS with L0-norm regularization to identify critical features, reducing modeling error by 11.95% on public datasets.¹³¹ Both methods address data limitations, EM-TLS mitigates measurement uncertainty through probabilistic noise modeling, while feature selection in TLS optimizes sparse models to prevent overfitting with limited samples. Compared to data generation methods, data imputation and denoising techniques can more efficiently utilize the true distribution of existing data, while preserving original physical features, avoiding the introduction of synthetic biases, and typically incurring lower computational costs.

ML for state monitoring

Accurate real-time monitoring of battery states, such as SOH and SOC, is critical for safe and reliable operation. Data-driven algorithms excel at learning the complex nonlinear relationships inherent in battery operation, but like lifetime prediction, state monitoring is challenged by data scarcity.^{132,133}

This scarcity arises because obtaining reliable SOH/SOC labels often requires destructive or time-consuming offline tests, making it impractical to generate large, diverse datasets covering the full operational spectrum. Consequently, models must learn from limited and often non-diverse labeled data. To address this, advanced strategies focus on embedding prior knowledge or quantifying model uncertainty. A primary strategy is to use Bayesian neural networks (BNNs) to explicitly model the uncertainty that arises from limited data. Unlike standard neural networks that produce single-point predictions, BNNs model neural network weights as probability distributions instead of fixed values, enabling uncertainty quantification through variational inference and Monte Carlo sampling.¹³⁴ This approach inherently

ios, the semi-supervised learning and transfer learning methods have also demonstrated robustness and generalization ability through the validations on datasets with discrepancies. To further emphasize this, Table 2 provides a comparative summary on the characteristics of datasets used for semi-supervised learning³² and transfer learning⁴², as well as the reported performance across these datasets in Ref. 32, making the findings more explicit and transparent. From Table 2, it can be observed that can consistently generate accurate predictions despite the variations across conditions for data acquisition. The semi-supervised learning method proposed in Ref.³² have achieved RMSEs between 0.070 and 0.100 for the prediction on the logarithm of lifetime across various data acquisition conditions, while substantially outperforming conventional techniques (Lasso and ElasticNet) given limited labeled training data. Transfer learning is inherently robust to multi-source heterogeneous data scenarios due to the build-in domain-adaptive characteristics. As the battery lifetime prediction model trained on Dataset1 (training dataset) usually performs less reliable on Dataset3 (secondary testing dataset),²⁶ two transfer learning methods have been proposed in Ref. 42, substantially reducing the prediction RMSE from 295.22 cycles to 201.29 cycles. Additionally, BatLiNet is proposed in Ref.¹²⁷, which further demonstrates the cross-type robustness of transfer learning by adapting the lifetime prediction model achieved using LFP battery to NMC/NCA/LCO batteries.

Except for semi-supervised learning and transfer learning, ensemble learning improves generalization and robustness by aggregating predictions from multiple models, which is crucial when limited data could cause a single model to overfit. For instance, Sui et al. proposed an optimized ensemble learning framework (BaggELM) for SOH estimation.¹²⁸ The research combines voltage sequence optimization from partial charging curves with bagging extreme learning machines, using Euclidean distance analysis to preserve aging features. The framework achieves remarkable accuracy with limited data, showing RMSE $\leq 0.4\%$ and mean absolute percentage error (MAPE) $\leq 0.25\%$ even with 1:2 training-validation splits. The approach

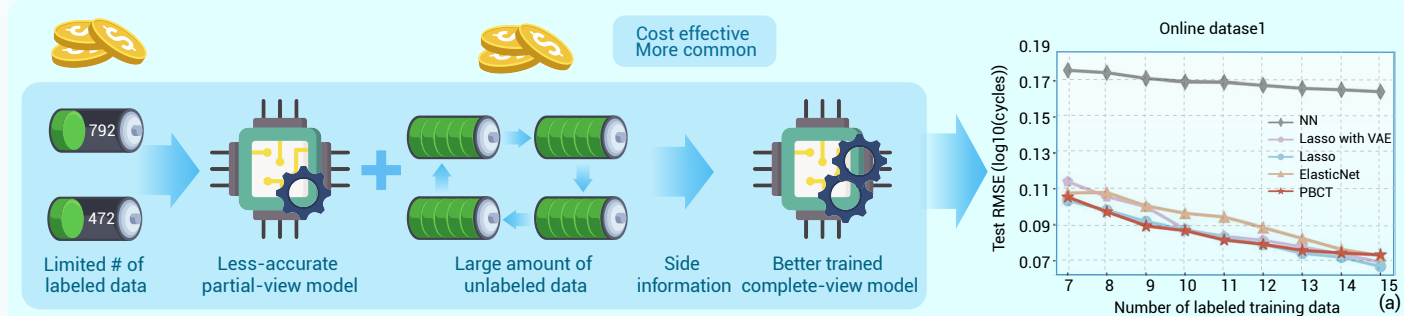
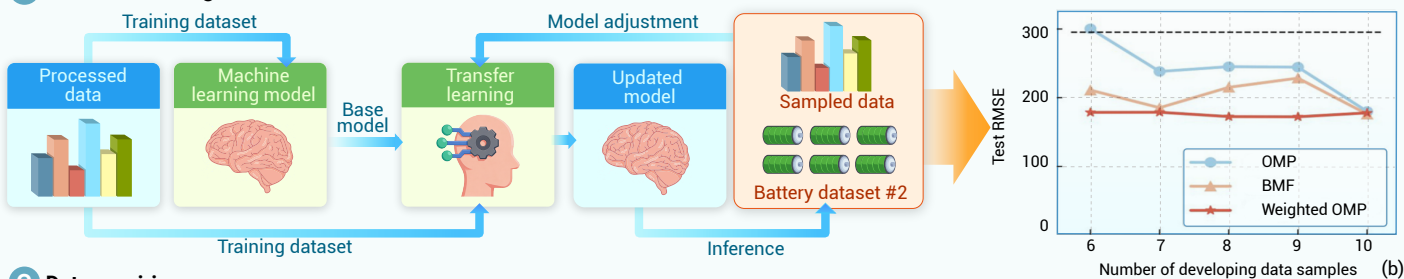
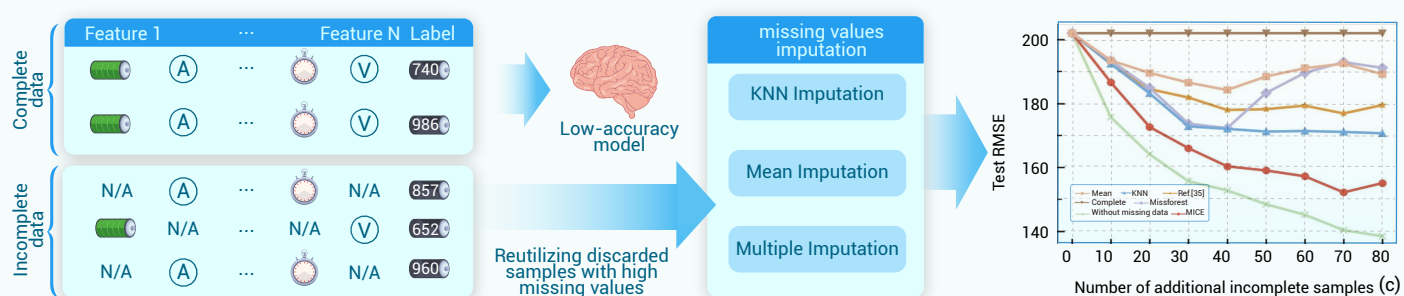
A Semi-supervised Learning**B Transfer Learning****C Data repairing**

Figure 7. ML guided efficient data utilization in life management (A) Pseudo-labels generated from unlabeled data are combined with a small amount of labeled data for joint training to enhance model interpretability and reducing costs.³² Copyright 2024, Elsevier. (B) Transfer learning methods balance dataset information to enabling high-accuracy and interpretable battery life prediction under state transitions.⁴² Copyright 2024, Elsevier. (C) The MICE significantly improves the accuracy and interpretability of battery lifespan prediction by enhancing training data through repairing incomplete data samples.¹²⁹ Copyright 2025, Elsevier.

Table 1. The economic analysis of some methodologies mentioned in the manuscript.

Algorithm	Cycles Required to Reach RMSE = 0.1 $\log_{10}(\text{cycle})$ or equivalent	Saved Cycles (Saved Cycle)	Saved Cost (Saved Money, \$)
Traditional Fully Supervised Testing Method	15000	0	0
Partial Bayesian Co-training (PBCT) ³²	7700.5	7300	12060
Weighted Orthogonal Matching Pursuit (W-OMP) ⁴²	7800	7200	11894
Bayesian Model Fusion (BMF) ⁴²	8300	6700	11068
Bagging Extreme Learning Machine (BaggELM) ¹³⁰	8800	6200	10242
Lasso Regression ³²	10537	4463	7373
Orthogonal Matching Pursuit (OMP) ⁴²	10900	4100	6773
ElasticNet ³²	11200	3800	6278
Support Vector Machine (SVM) ³²	12500	2500	4130

reduces overfitting in data-scarce scenarios by providing a natural form of regularization. As shown in Figure 8A, Zhang et al. tackled SOH estimation by developing a Bayesian deep neural network (BDNN).¹³⁵ The model integrates Bayesian convolutional and LSTM layers to quantify prediction uncertainty, while degradation pattern (DP) and gramian angular fields (GAF) enrich input representations. Trained on only 1%-50% of three datasets (MIT-Stanford,

CALCE, Oxford), the BDNN achieved mean absolute error $\leq 0.47\%$ and RMSE ≤ 0.006 with 1% data, outperforming benchmarks. Quantitatively, BDNN reduced errors by 60%-80% in multi-step predictions compared to conventional models, demonstrating robustness to varied operating conditions. This framework enables reliable, data-efficient SOH estimation with quantified uncertainty, critical for real-world applications with sparse training data.

Table 2. The characteristics of datasets used for semi-supervised learning and transfer learning,^{32,42} as well as the reported performance for battery lifetime prediction in Ref.³².

Sub-dataset	Number of Cells	Test conditions	Test RMSE (log ₁₀ (cycle))
Dataset 1 (Training dataset)	41	Short storage (no extra calendar aging); charging protocol identical to the reference set; same C-rate range and SOC switching schedule; nominal 25°C.	0.088
Dataset 2 (Primary testing dataset)	43	Moderate storage time (different from D1); slightly modified charging protocol (minor change in current steps/SOC windows); nominal 25°C	0.081
Dataset 3 (Secondary testing dataset)	40	Long storage / calendar-aged; charging protocol differs from D1/D2 (narrower C-rate range and altered SOC switching setpoints); noticeably lower initial capacity; nominal 25°C.	0.073

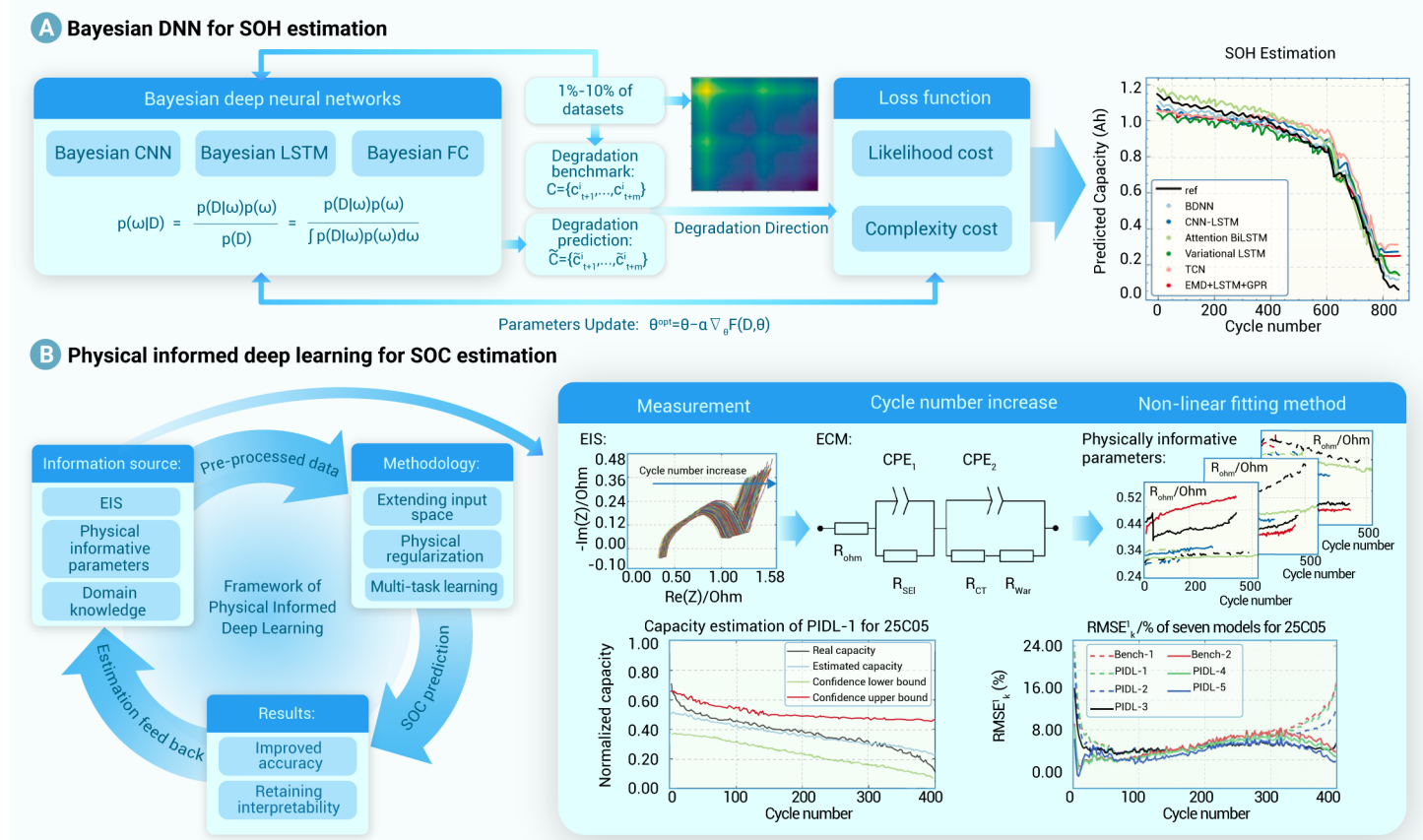


Figure 8. SOX estimation based on ML methodologies (A) A BDNN method is proposed for estimating the SOC of LIB. It extracts DP and GAF features and leverages a Bayesian network to handle small-sample data effectively.¹³⁵ Copyright 2023, Elsevier. (B) The PIDL can enhance the accuracy and interpretability of battery SOC estimation by integrating physical principles, measurement data, and domain knowledge.¹³⁷ Copyright 2023, Elsevier.

Another powerful approach is to embed physical principles directly into the model to compensate for data scarcity. Physics-informed deep learning (PIDL) integrates domain knowledge from physical models with data-driven methods. Specifically, PINNs incorporate physical constraints, such as energy conservation, charge balance, and chemical kinetics equations, into the loss function.¹³⁶ As shown in Figure 8B, Lin et al. proposed a PIDL framework for SOC estimation.¹³⁷ This framework addresses data scarcity by fusing two information sources: physical parameters extracted from EIS via equivalent circuit modeling with raw EIS data through two fusion strategies, while employing multi-task learning and physical regularization to leverage domain knowledge and mitigate limited data. The deep ensemble approach further quantifies predictive uncertainty. Tested on EIS datasets from 8 commercial LIB, the framework achieved significant improvements: PIDL-5 reduced RMSE to 7.22% and 3.73% for two test batteries, with confidence intervals shortened by 33%-38% compared to baseline models. By grounding the model in physical laws, this method ensures that predictions remain physically plausible even in data-sparse regions, demonstrating effective resolution of data scarcity through physics-guided feature enrichment.

ML for operational optimization

Operational optimization, which aims to enhance battery safety, efficiency, and lifespan, is fundamentally a high-dimensional search problem challenged by extreme data scarcity.¹³⁸ The goal is to find optimal strategies for thermal management or fast-charging by navigating a vast parameter space, where exhaustive experimental exploration is prohibitively expensive. To overcome this, ML strategies have been developed to enable rapid adaptation to new scenarios and intelligent exploration of the parameter space.

When developing optimization strategies for new battery types or operating conditions, data is exceptionally scarce. Meta-learning, or "learning to learn", addresses this by training a model on a variety of related tasks to learn a generalized initialization or adaptation strategy.^{139,140} This enables the model to achieve few-shot adaptation to new scenarios through learned initialization of model parameters in high-dimensional optimization spaces. As shown in Figure 9A, Ding et al. proposed a meta-learning-based multimodal neural network (Meta-TRFNN) for multistep thermal runaway (TR) forecasting.¹⁴¹ The method integrates high-dimensional thermal images with low-dimensional temperature/voltage data through a neural architecture combining

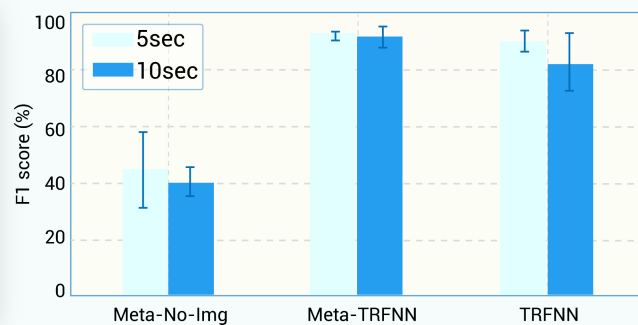
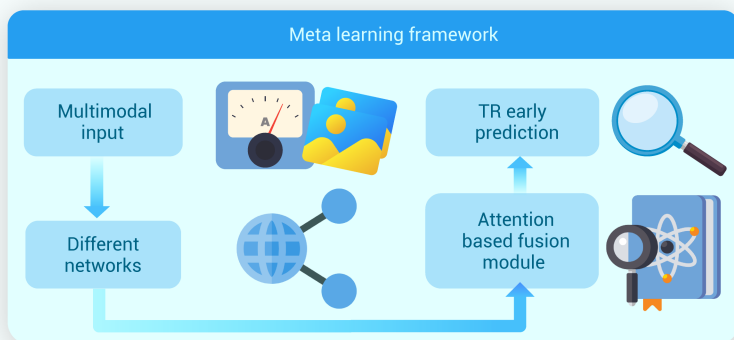
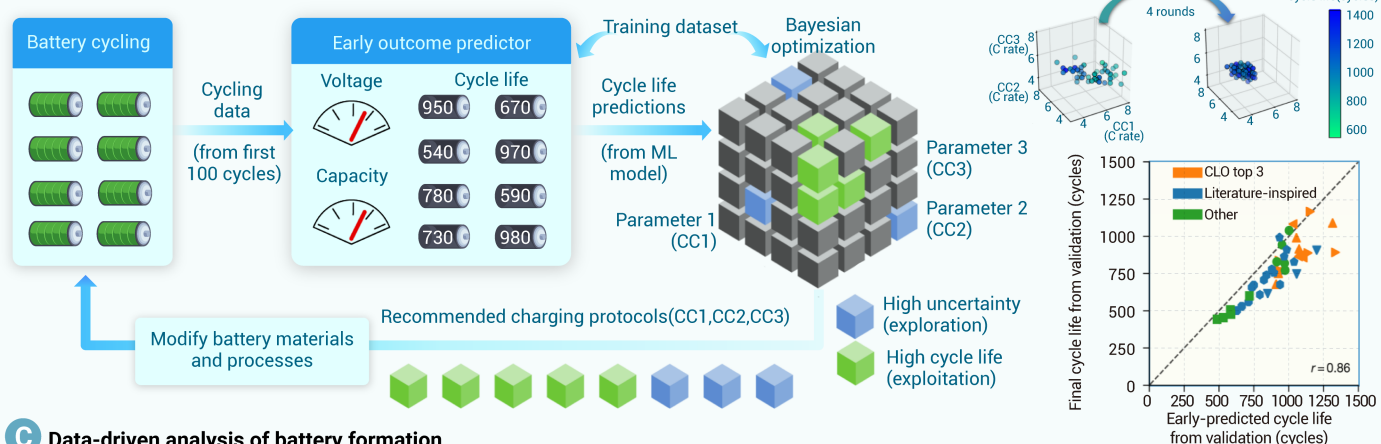
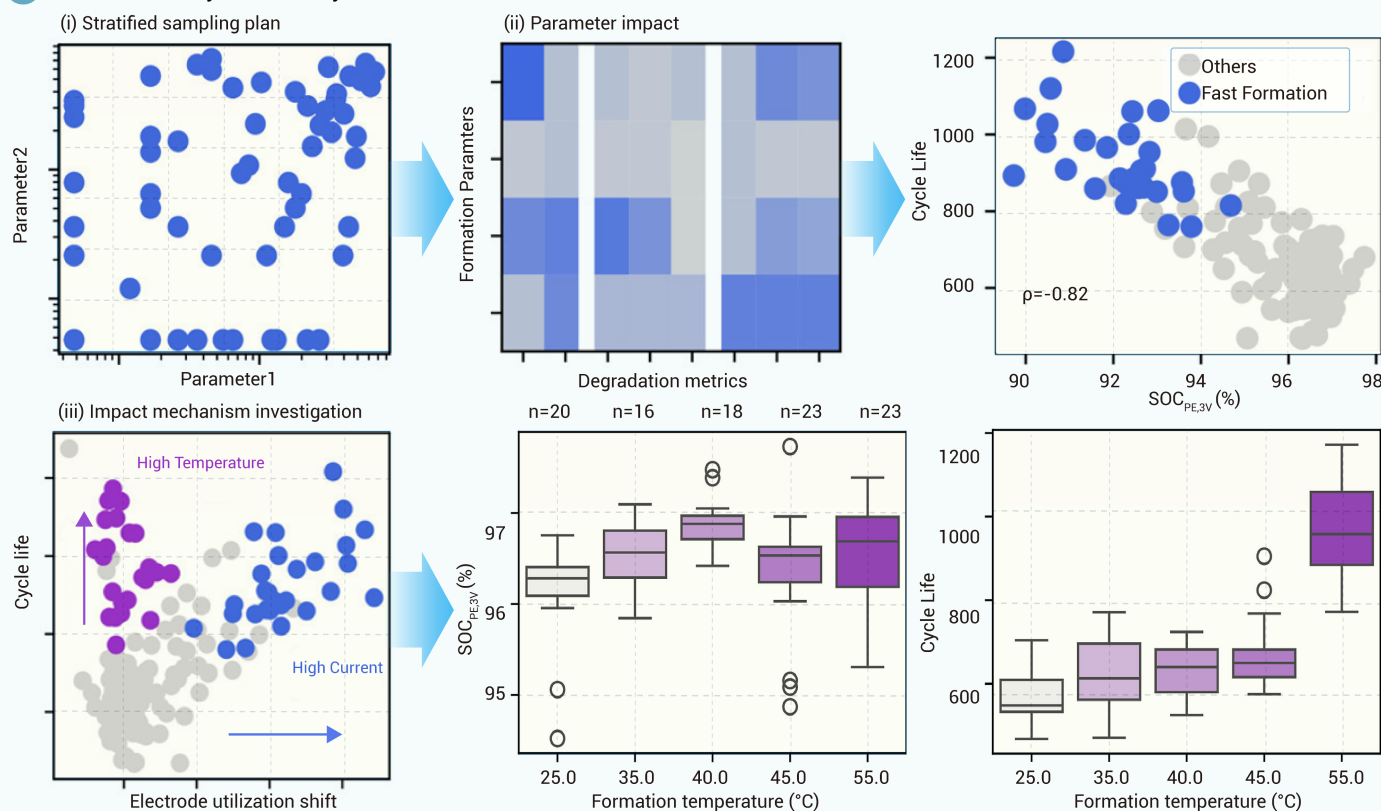
A Meta thermal runaway forecasting neural network**B Optimization of fast-charging protocols****C Data-driven analysis of battery formation**

Figure 9. ML assisted predictive and adaptive battery management (A) A meta-learning based neural network using attention to combine multimodal data for multistep ahead battery thermal runaway forecasting.¹⁴¹ Copyright 2020, IEEE Xplore. (B) Using early lifetime prediction and BO can reduce searching time cost and focus on best charging protocols while close loop system adjusts test strategy to maximize data utilization.²⁸ Copyright 2020, Springer Nature. (C) Using a random forest model and SHAP analysis, complex data is efficiently processed to reveal the impact of various parameters on cycle life during battery formation. Combined with differential voltage analysis and half-cell and hybrid pulse power characterization tests, the mechanisms by which rapid formation affect battery cycle life are explored.¹⁴² Copyright 2024, Elsevier.

gated recurrent units, CNNs, and attention-based fusion to capture comprehensive thermal profiles. To overcome data scarcity, the meta-learning framework leverages 10% of training data samples to enhance few-shot learning capabilities. Experiments on simulated and real-world datasets demonstrated significant performance: Meta-TRFNN achieved 94.8–99.35% accuracy and 91.46%–98.96% F1 scores on simulated data, outperforming baseline models by 13%–18% F1. The integration of thermal images improved the F1 score by over 50% compared to non-image models. With this combined approach, it is possible to enhance feature extraction and data fusion efficiency while maximizing the value of limited data.

To efficiently search for optimal protocols within a given parameter space, BO has emerged as a powerful tool. BO builds a surrogate model of the objective function (e.g., cycle life) and uses an acquisition function to intelligently select the next set of parameters to test, balancing exploration (sampling in regions of high uncertainty) and exploitation (sampling in regions of predicted high performance). As shown in Figure 9B, Attia et al. proposed a closed-loop optimization (CLO) framework that combines an early-prediction model with BO to discover optimal fast-charging protocols.²⁸ The early-prediction model estimated cycle life from the first 100 cycles, and BO used these predictions to navigate the 224-protocol parameter space. BO outperformed random search, achieving comparable results in 500 battery-hours versus 7,700 hours, demonstrating efficiency gains. The optimized protocols achieved an average validated cycle life of 895 cycles, surpassing literature-inspired designs.

In a similar vein, intelligent experimental design can maximize information gain from a limited number of experiments. As shown in Figure 9C, Cui et al. developed a data-driven workflow to optimize battery formation protocols.¹⁴² To efficiently explore the 6-dimensional parameter space, they used stratified Latin hypercube sampling, a method that ensures broad and uniform coverage with minimal sample points, thereby directly addressing data scarcity. This reduced the required experiments by 67% compared to conventional grid searches. The resulting model achieved an R^2 of 0.85 in cycle life prediction, with SHAP analysis revealing the key influencing factors. This workflow demonstrates how targeted sampling and explainable ML can extract mechanistic insights from limited datasets, effectively resolving the trade-off between parameter exploration depth and experimental feasibility.

SUMMARY AND OUTLOOK

This review has systematically analyzed the pervasive challenge of data scarcity in ML-driven battery research across the entire workflow from materials discovery to device deployment. While the high cost and complexity of battery research make data scarcity an inherent issue, a paradigm shift towards data-efficient AI offers a powerful path forward. Our analysis reveals three core, complementary strategies that researchers are adopting to overcome these limitations through data preparation, model construction, and data augmentation. The data preparation stage contains meticulous pre-processing to repair imperfect data, sophisticated feature engineering to embed domain knowledge, and active learning to intelligently guide new experiments toward the most informative outcomes. These data-centric approaches contribute to enhance data quality and information density, thereby maximizing the utilization of every available data point. The model construction stage establishes a knowledge-driven paradigm that integrates fundamental scientific principles into the ML framework, which regularizes the models with physical laws or causal relationships to reduce the reliance on large volumes of data, while ensuring to learn physically plausible and interpretable patterns. The data augmentation stage focuses on incorporating information from synthetic data, external data, and unlabeled data, powered by generative models, transfer learning, and semi-supervised learning, to expand and diversify the available data pool. These strategies provide valuable and comprehensive paths to alleviate the data-scarcity issue in battery research.

Looking forward, the synergy between AI and battery science is poised to unlock even more transformative capabilities. Therefore, several key directions are foresaw that will further address the data scarcity challenge in battery research. (A) **Foundation models for battery science:** the development of large-scale “foundation models” pre-trained on massive, aggregated datasets from diverse battery chemistries and conditions will hold immense

promise, which can provide a strong basis for researchers to fine-tune for their specific, data-scarce tasks, such as property prediction of novel materials or lifetime prediction for new types of battery cells, to dramatically reduce the need for extensive de novo data collection. While such battery foundation pre-trained models are highly promising, their development also faces several challenges specific to the pre-training paradigm. Aggregating data from many sources and chemistries creates strong distribution imbalance. In this case, over-represented systems can dominate the learned representation, while under-represented but important chemistries are poorly captured. Scaling these models to multi-modal, high-resolution data also raises model-specific issues such as optimization instability, catastrophic forgetting when new data are added, and high training and inference cost, which in turn makes robust transfer across different battery chemistries (e.g., Li-ion, solid-state, Na-ion) particularly difficult. (B) **Large language model (LLM)-enhanced knowledge synthesis:** as LLMs are becoming powerful tools for scientific knowledge management and synthesis, their ability to understand semantic context allows them to process and organize useful information from vast corpora of unstructured text, such as research papers and patents.^{143,144} This can help researchers rapidly identify limitations in existing research, synthesize material property data scattered across literature, and generate data-driven hypotheses, thereby guiding more targeted and efficient experimental data collection. For instance, BatteryGPT is proposed and proven a ability to extract fragmented data into experimental recommendations.¹⁴⁴ And retrieval-augmented generation enhanced LLMs can automatically provide comprehensive insights through literature reviews. (C) **Federated learning frameworks for data sharing:** data scarcity is frequently compounded by the challenges of cross-entity data sharing due to confidentiality and proprietary restrictions. Federated learning offers a promising solution by enabling fragmented information aggregation without physically exposing raw data to external platforms. This inherent privacy-preserving capability allows models to benefit from diverse datasets while mitigating data leakage risks.

Overall, the discussions above suggest that the primary bottleneck for AI-enabled LIB research in data-scarce settings lies less in algorithm availability than in establishing reliable evidence—namely, representative data coverage, rigorous validation under distribution shifts, and trustworthy uncertainty quantification. Continued progress is therefore expected to depend on stronger community practices (standardized reporting and benchmarks) and tighter coupling between modeling and experimentation (e.g., closed-loop design), thereby providing a solid foundation for the promising path forward outlined below. While achieving full intelligentization in battery research is a work in progress, recent advancements have already mapped out a promising path forward. By fostering collaboration across disciplines and embracing emerging data-efficient and knowledge-integrated paradigms, the research community will be well-equipped to overcome the challenge of data scarcity and usher in a new era of accelerated discovery and development for next-generation energy storage techniques.

REFERENCES

1. Tarascon J.-M. and Armand M. (2001). Issues and challenges facing rechargeable lithium batteries. *Nature* **414**:359–367. DOI:10.1038/35104644
2. Wang H., Yu Z., Kong X., et al. (2021). Dual-solvent Li-ion solvation enables high-performance Li-metal batteries. *Adv. Mater.* **33**:2008619. DOI:10.1002/adma.202008619
3. Wang Y., Chen C., Xie H., et al. (2017). 3D-printed all-fiber Li-ion battery toward wearable energy storage. *Adv. Funct. Mater.* **27**:1703140. DOI:10.1002/adfm.201703140
4. Xu K. (2004). Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chem. Rev.* **104**:4303–4418. DOI:10.1021/cr030203g
5. Zhang Y., Li Y., Guo Z., et al. (2024). Health monitoring by optical fiber sensing technology for rechargeable batteries. *eScience* **4**:100174. DOI:10.1016/j.esci.2023.100174
6. Bresser D., Hosoi K., Howell D., et al. (2018). Perspectives of automotive battery R&D in china, germany, japan, and the USA. *J. Power Sources* **382**:176–178. DOI:10.1016/j.jpowsour.2018.02.039
7. Dunn B., Kamath H. and Tarascon J.-M. (2011). Electrical energy storage for the grid: a battery of choices. *Science* **334**:928–935. DOI:10.1126/science.1212741
8. Nykvist B. and Nilsson M. (2015). Rapidly falling costs of battery packs for electric vehicles. *Nat. Clim. Chang.* **5**:329–332. DOI:10.1038/nclimate2564
9. Schmuck R., Wagner R., Hörpel G., et al. (2018). Performance and cost of materials

- for lithium-based rechargeable automotive batteries. *Nat. Energy* **3**:267–278. DOI:10.1038/s41560-018-0107-2
10. Vikström H., Davidsson S. and Höök M. (2013). Lithium availability and future production outlooks. *Appl. Energy* **110**:252–266. DOI:10.1016/j.apenergy.2013.04.005
11. Liu Y., Chen S., Li P., et al. (2024). Status, challenges, and promises of data-driven battery lifetime prediction under cyber-physical system context. *IET Cyber-Phys. Syst.* **9**:207–217. DOI:10.1049/cps.2.12086
12. Xue Z., Zhang Y., Cheng C., et al. (2020). Remaining useful life prediction of lithium-ion batteries with adaptive unscented Kalman filter and optimized support vector regression. *Neurocomputing* **376**:95–102. DOI:10.1016/j.neucom.2019.09.074
13. Lipu M.S.H., Hannan M.A., Hussain A., et al. (2018). A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: challenges and recommendations. *J. Clean. Prod.* **205**:115–133. DOI:10.1016/j.jclepro.2018.09.065
14. Zheng X. and Fang H. (2015). An integrated unscented Kalman filter and relevance vector regression approach for lithium-ion battery remaining useful life and short-term capacity prediction. *Reliab. Eng. Syst. Saf.* **144**:74–82. DOI:10.1016/j.res.2015.07.013
15. Azis N.A., Joelianto E. and Widyotriatmo A. (2019). State of charge (SoC) and state of health (SoH) estimation of lithium-ion battery using dual extended Kalman filter based on polynomial battery model. *2019 6th Int. Conf. Instrum., Control Autom.* **2019**:88–93. DOI:10.1109/ica.2019.8916734
16. Guha A., Patra A. and Vaisakh K.V. (2017). Remaining useful life estimation of lithium-ion batteries based on the internal resistance growth model. *2017 Indian Control Conf.* **2017**:33–38. DOI:10.1109/indiancc.2017.7846448
17. Li D.Z., Wang W. and Ismail F. (2014). A mutated particle filter technique for system state estimation and battery life prediction. *IEEE Trans. Instrum. Meas.* **63**:2034–2043. DOI:10.1109/tim.2014.2303534
18. Ahwiadi M. and Wang W. (2019). An enhanced mutated particle filter technique for system state estimation and battery life prediction. *IEEE Trans. Instrum. Meas.* **68**:923–935. DOI:10.1109/tim.2018.2853900
19. Weng C., Cui Y., Sun J., et al. (2013). On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression. *J. Power Sources* **235**:36–44. DOI:10.1016/j.jpowsour.2013.02.012
20. Weng C., Feng X., Sun J., et al. (2016). State-of-health monitoring of lithium-ion battery modules and packs via incremental capacity peak tracking. *Appl. Energy* **180**:360–368. DOI:10.1016/j.apenergy.2016.07.126
21. Zhang Y., Tang Q., Zhang Y., et al. (2020). Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nat. Commun.* **11**:15135. DOI:10.1038/s41467-020-15235-7
22. Ding R., Wang R., Ding Y., et al. (2020). Designing AI-aided analysis and prediction models for nonprecious metal electrocatalyst-based proton-exchange membrane fuel cells. *Angew. Chem.* **132**:19337–19345. DOI:10.1002/ange.202006928
23. Ying X. (2019). An overview of overfitting and its solutions. *J. Phys.: Conf. Ser.* **1168**:022022. DOI:10.1088/1742-6596/1168/2/022022
24. Dietterich T. (1995). Overfitting and undercomputing in machine learning. *ACM Comput. Surv.* **27**:326–327. DOI:10.1145/212094.212114
25. Hawkins D.M. (2003). The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**:1–12. DOI:10.1021/ci0342472
26. Severson K.A., Attia P.M., Jin N., et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**:383–391. DOI:10.1038/s41560-019-0356-8
27. Yang D., Zhang X., Pan R., et al. (2018). A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. *J. Power Sources* **384**:387–395. DOI:10.1016/j.jpowsour.2018.03.015
28. Attia P.M., Grover A., Jin N., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **578**:397–402. DOI:10.1038/s41586-020-1994-5
29. Weng A., Mohtat P., Attia P.M., et al. (2021). Predicting the impact of formation protocols on battery lifetime immediately after manufacturing. *Joule* **5**:2971–2992. DOI:10.1016/j.joule.2021.09.015
30. Jones P.K., Stimming U. and Lee A.A. (2022). Impedance-based forecasting of lithium-ion battery performance amid uneven usage. *Nat. Commun.* **13**:6324. DOI:10.1038/s41467-022-32422-w
31. Tao S., Liu H., Sun C., et al. (2023). Collaborative and privacy-preserving retired battery sorting for profitable direct recycling via federated machine learning. *Nat. Commun.* **14**:8883. DOI:10.1038/s41467-023-43883-y
32. Guo N., Chen S., Tao J., et al. (2024). Semi-supervised learning for explainable few-shot battery lifetime prediction. *Joule* **8**:1820–1836. DOI:10.1016/j.joule.2024.02.020
33. Ward L., Agrawal A., Choudhary A., et al. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**:16028. DOI:10.1038/npjcompumats.2016.28
34. Zhang Y., He X., Chen Z., et al. (2019). Unsupervised discovery of solid-state lithium ion conductors. *Nat. Commun.* **10**:5267. DOI:10.1038/s41467-019-13214-1
35. Jiang Z., Li J., Yang Y., et al. (2020). Machine-learning-revealed statistics of the particle-carbon/binder detachment in lithium-ion battery cathodes. *Nat. Commun.* **11**:2737. DOI:10.1038/s41467-020-16233-5
36. Niri M.F., Liu K., Apachitei G., et al. (2021). Machine learning for optimised and clean Li-ion battery manufacturing: Revealing the dependency between electrode and cell characteristics. *J. Clean. Prod.* **324**:129272. DOI:10.1016/j.jclepro.2021.129272
37. Finegan D.P., Squires I., Dahari A., et al. (2022). Machine-learning-driven advanced characterization of battery electrodes. *ACS Energy Lett.* **7**:4368–4378. DOI:10.1021/acscenergylett.2c01996
38. Merchant A., Batzner S., Schoenholz S.S., et al. (2023). Scaling deep learning for materials discovery. *Nature* **624**:80–85. DOI:10.1038/s41586-023-06735-9
39. Ren F., Wu Y., Zuo W., et al. (2024). Visualizing the SEI formation between lithium metal and solid-state electrolyte. *Energy Environ. Sci.* **17**:2743–2752. DOI:10.1039/d3ee03536k
40. Naaz F., Herle A., Channegowda J., et al. (2021). A generative adversarial network-based synthetic data augmentation technique for battery condition evaluation. *Int. J. Energy Res.* **45**:19120–19135. DOI:10.1002/er.7013
41. Laskowski F.A.L., McHaffie D.B. and See K.A. (2023). Identification of potential solid-state Li-ion conductors with semi-supervised learning. *Energy Environ. Sci.* **16**:1264–1276. DOI:10.1039/d2ee03499a
42. Lin T., Chen S., Harris S.J., et al. (2024). Investigating explainable transfer learning for battery lifetime prediction under state transitions. *eScience* **4**:100280. DOI:10.1016/j.esci.2024.100280
43. Finegan D.P., Zhu J., Feng X., et al. (2021). The application of data-driven methods and physics-based learning for improving battery safety. *Joule* **5**:316–329. DOI:10.1016/j.joule.2020.11.018
44. Richardson R.R., Birkel C.R., Osborne M.A., et al. (2019). Gaussian process regression for in situ capacity estimation of lithium-ion batteries. *IEEE Trans. Ind. Inform.* **15**:127–138. DOI:10.1109/tii.2018.2794997
45. Lin C.P., Cabrera J., Yu D.Y.W., et al. (2020). SoH estimation and SoC recalibration of lithium-ion battery with incremental capacity analysis & cubic smoothing spline. *J. Electrochem. Soc.* **167**:090537. DOI:10.1149/1945-7111/ab8f56
46. Zhang J. and Lee J. (2011). A review on prognostics and health monitoring of Li-ion battery. *J. Power Sources* **196**:6007–6014. DOI:10.1016/j.jpowsour.2011.03.101
47. Verma P., Maire P. and Novák P. (2010). A review of the features and analyses of the solid electrolyte interphase in Li-ion batteries. *Electrochim. Acta* **55**:6332–6341. DOI:10.1016/j.electacta.2010.05.072
48. Yang M., Sun X., Liu R., et al. (2024). Predict the lifetime of lithium-ion batteries using early cycles: A review. *Appl. Energy* **376**:124171. DOI:10.1016/j.apenergy.2024.124171
49. Lin M., You Y., Meng J., et al. (2023). Lithium-ion battery degradation trajectory early prediction with synthetic dataset and deep learning. *J. Energy Chem.* **85**:534–546. DOI:10.1016/j.jechem.2023.06.036
50. Iftikhar M., Shoaib M., Altaf A., et al. (2024). A deep learning approach to optimize remaining useful life prediction for Li-ion batteries. *Sci. Rep.* **14**:77427. DOI:10.1038/s41598-024-77427-1
51. Ren X. and Malik J. (2003). Learning a classification model for segmentation. *Proc. IEEE Int. Conf. Comput. Vis.* **2003**:10–17. DOI:10.1109/iccv.2003.1238308
52. Bolloju S., Vangapally N., Elias Y., et al. (2025). Electrolyte additives for Li-ion batteries: classification by elements. *Prog. Mater. Sci.* **147**:101349. DOI:10.1016/j.pmatsci.2024.101349
53. Li X., Yuan C., Li X., et al. (2020). State of health estimation for Li-ion battery using incremental capacity analysis and Gaussian process regression. *Energy* **190**:116467. DOI:10.1016/j.energy.2019.116467
54. Nahar L., Awrangzeb M. and Islam M.S. (2026). AI-enabled defect detection in industrial products: A comprehensive survey, key insights and future research challenges. *Adv. Eng. Inform.* **69**:104067. DOI:10.1016/j.aei.2025.104067
55. Shen S., Sadoughi M., Li M., et al. (2020). Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries. *Appl. Energy* **260**:114296. DOI:10.1016/j.apenergy.2019.114296
56. Zhang J., Wang Y., Jiang B., et al. (2023). Realistic fault detection of Li-ion battery via dynamical deep learning. *Nat. Commun.* **14**:1226. DOI:10.1038/s41467-023-41226-5
57. Ghadbeigi L., Harada J.K., Lettiere B.R., et al. (2015). Performance and resource considerations of Li-ion battery electrode materials. *Energy Environ. Sci.* **8**:1640–1650. DOI:10.1039/c5ee00685f
58. Paulson N.H., Kubal J., Ward L., et al. (2022). Feature engineering for machine learning enabled early prediction of battery lifetime. *J. Power Sources* **527**:231127. DOI:10.1016/j.jpowsour.2022.231127
59. Urban A., Seo D.-H. and Ceder G. (2016). Computational understanding of Li-ion batteries. *npj Comput. Mater.* **2**:16002. DOI:10.1038/npjcompumats.2016.2
60. Ma G., Xu S., Jiang B., et al. (2022). Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. *Energy Environ. Sci.* **15**:4083–4094. DOI:10.1039/d2ee01676a
61. Zhu J., Wang Y., Huang Y., et al. (2022). Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. *Nat. Commun.* **13**:3837. DOI:10.1038/s41467-022-29837-w
62. Das Goswami B.R., Mastrogiorio M., Ragone M., et al. (2024). A combined multiphysics modeling and deep learning framework to predict thermal runaway in cylindrical Li-ion batteries. *J. Power Sources* **595**:234065. DOI:10.1016/j.jpowsour.2024.234065

63. Settles B. (2012). Active learning. Springer. DOI:10.1007/978-3-031-01560-1
64. Zhu R., Chen Y., Peng W., et al. (2022). Bayesian deep-learning for RUL prediction: an active learning perspective. *Reliab. Eng. Syst. Saf.* **228**:108758. DOI:10.1016/j.res.2022.108758
65. Kim S., Choi Y.Y. and Choi J.-I. (2022). Impedance-based capacity estimation for lithium-ion batteries using generative adversarial network. *Appl. Energy* **308**:118317. DOI:10.1016/j.apenergy.2021.118317
66. Fernandez A., Garcia S., Herrera F., et al. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**:863–905. DOI:10.1613/jair.1.11192
67. Chawla N.V., Bowyer K.W., Hall L.O., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**:321–357. DOI:10.1613/jair.953
68. Camacho L., Douzas G. and Bacao F. (2022). Geometric SMOTE for regression. *Expert Syst. Appl.* **193**:116387. DOI:10.1016/j.eswa.2021.116387
69. Han H., Wang W.Y. and Mao B.H. (2005). borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Lect. Notes Comput. Sci.* **3644**:878–887. DOI:10.1007/11538059_91
70. Xu S., Lu B., Baldea M., et al. (2015). Data cleaning in the process industries. *Rev. Chem. Eng.* **31**:453–490. DOI:10.1515/revce-2015-0022
71. Wang Y., Yao Q., Kwok J.T., et al. (2020). Generalizing from a few examples. *ACM Comput. Surv.* **53**:63. DOI:10.1145/3386252
72. Velleman P.F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *J. Am. Stat. Assoc.* **75**:609–615. DOI:10.1080/01621459.1980.10477521
73. Patro S.G.K. and Sahu K.K. (2015). Normalization: A preprocessing stage. *IARJSET* **2**:20–22. DOI:10.17148/iarjset.2015.2305
74. Xiong R., Sun Y., Wang C., et al. (2023). A data-driven method for extracting aging features to accurately predict the battery health. *Energy Storage Mater.* **57**:460–470. DOI:10.1016/j.ensm.2023.02.034
75. Tao S., Sun C., Fu S., et al. (2023). Battery cross-operation-condition lifetime prediction via interpretable feature engineering assisted adaptive machine learning. *ACS Energy Lett.* **8**:3269–3279. DOI:10.1021/acsenenergylett.3c01012
76. Ma M., Li X., Gao W., et al. (2022). Multi-fault diagnosis for series-connected lithium-ion battery pack with reconstruction-based contribution based on parallel PCA-KPCA. *Appl. Energy* **324**:119678. DOI:10.1016/j.apenergy.2022.119678
77. Sarasketa-Zabala E., Martinez-Laserna E., Berecibar M., et al. (2016). Realistic lifetime prediction approach for Li-ion batteries. *Appl. Energy* **162**:839–852. DOI:10.1016/j.apenergy.2015.10.115
78. Liu Y. and Li X. (2018). Predictive modeling for advanced virtual metrology: A tree-based approach. *Proc. IEEE Int. Conf. Emerg. Technol. Factory Autom.* **2018**:845–852. DOI:10.1109/etfa.2018.8502480
79. Mosavi A., Salimi M., Faizollahzadeh Ardabili S., et al. (2019). State of the art of machine learning models in energy systems, a systematic review. *Energies* **12**:1301. DOI:10.3390/en12071301
80. Oh H.-S. (2013). Introduction to linear regression analysis, 5th edition by montgomery, douglas c. , Peck, Elizabeth A., and Vining, G. Geoffrey. *Biometrics* **69**:1087. DOI:10.1111/biom.12129
81. Weisberg S. (2005). Applied linear regression. Wiley. DOI:10.1002/0471704091.
82. Su X., Yan X. and Tsai C. (2012). Linear regression. *WIREs Comput. Stat.* **4**:275–294. DOI:10.1002/wics.1198
83. Vilsen S.B. and Stroe D.-I. (2021). Battery state-of-health modelling by multiple linear regression. *J. Clean. Prod.* **290**:125700. DOI:10.1016/j.jclepro.2020.125700
84. Sun J., Fan C. and Yan H. (2024). SoH estimation of lithium-ion batteries based on multi-feature deep fusion and XGBoost. *Energy* **306**:132429. DOI:10.1016/j.energy.2024.132429
85. Prokhorenkova L., Gusev G., Vorobev A., et al. (2018). CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **31**:6638–6648. DOI:10.13140/RG.2.2.30029.96485
86. Niu S., Liu Y., Wang J., et al. (2020). A decade survey of transfer learning (2010–2020). *IEEE Trans. Artif. Intell.* **1**:151–166. DOI:10.1109/tai.2021.3054609
87. Samek W., Montavon G., Lapuschkin S., et al. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **109**:247–278. DOI:10.1109/jproc.2021.3060483
88. Miikkulainen R., Liang J., Meyerson E., et al. (2024). Evolving deep neural networks. Artificial intelligence in the age of neural networks and brain computing. Elsevier. DOI:10.1016/b978-0-323-96104-2.00002-6
89. Li Z., Liu F., Yang W., et al. (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **33**:6999–7019. DOI:10.1109/tnnls.2021.3084827
90. Gu J., Wang Z., Kuen J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognit.* **77**:354–377. DOI:10.1016/j.patcog.2017.10.013
91. Sherstinsky A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* **404**:132306. DOI:10.1016/j.physd.2019.132306
92. Yu Y., Si X., Hu C., et al. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**:1235–1270. DOI:10.1162/neco_a_01199
93. Varis D. and Bojar O. (2021). Sequence length is a domain: Length-based overfitting in transformer models. *Proc. Conf. Empir. Methods Nat. Lang. Process.* **2021**:3042–3051. DOI:10.18653/v1/2021.emnlp-main.650
94. Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**:267–288. DOI:10.1111/j.2517-6161.1996.tb02080.x
95. Owen A.B. (2007). A robust hybrid of lasso and ridge regression. *Contemp. Math.* **443**:59–71. DOI:10.1090/conm/443/08555
96. Wu X. (2023). A dropout optimization algorithm to prevent overfitting in machine learning. *Mach. Learn. Theory Pract.* **4**:18–26. DOI:10.38007/ML.2023.040103
97. Kyono T., Zhang Y. and van der Schaar M. (2020). CASTLE: Regularization via auxiliary causal graph discovery. *Adv. Neural Inf. Process. Syst.* **33**:1501–1512. DOI:10.5555/3495724.3495939
98. Cuomo S., Di Cola V.S., Giampaolo F., et al. (2022). Scientific machine learning through physics-informed neural networks: Where we are and what's next. *J. Sci. Comput.* **92**:88. DOI:10.1007/s10915-022-01939-z
99. Xiao J., Wang Y., Xiang H., et al. (2025). A physics-informed multiview collaborative semisupervised framework for battery lifespan early prediction. *IEEE Trans. Ind. Inform.* **21**:9933–9944. DOI:10.1109/tii.2025.3609201
100. Weiss K., Khoshgoftaar T.M. and Wang D. (2016). A survey of transfer learning. *J. Big Data* **3**:9. DOI:10.1186/s40537-016-0043-6
101. Zhu X. and Goldberg A. (2009). Introduction to semi-supervised learning. Morgan & Claypool Publishers. DOI:10.7551/mitpress/6173.003.0003.
102. Wong T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit.* **48**:2839–2846. DOI:10.1016/j.patcog.2015.03.009
103. Vakharia V., Shah M., Nair P., et al. (2023). Estimation of lithium-ion battery discharge capacity by integrating optimized explainable-AI and stacked LSTM model. *Batteries* **9**:125. DOI:10.3390/batteries9020125
104. Zhu X., Chen N. and Shen C. (2020). A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions. *Mech. Syst. Signal Process.* **139**:106602. DOI:10.1016/j.ymssp.2019.106602
105. Gou J., Yu B., Maybank S.J., et al. (2021). Knowledge distillation: A survey. *Int. J. Comput. Vis.* **129**:1789–1819. DOI:10.1007/s11263-021-01453-z
106. Kilic A., Oral B., Eroglu D., et al. (2023). Machine learning for beyond Li-ion batteries: powering the research. *J. Energy Storage* **73**:109057. DOI:10.1016/j.est.2023.109057
107. Luo H., Gou Q., Zheng Y., et al. (2025). Machine learning-assisted high-donor-number electrolyte additive screening toward construction of dendrite-free aqueous zinc-ion batteries. *ACS Nano* **19**:2427–2443. DOI:10.1021/acsnano.4c13312
108. Luo H., Deng J., Gou Q., et al. (2023). Accelerated discovery of novel high-performance zinc-ion battery cathode materials by combining high-throughput screening and experiments. *Chin. Chem. Lett.* **34**:107885. DOI:10.1016/j.cclet.2022.107885
109. Lv C., Zhou X., Zhong L., et al. (2022). Machine learning: An advanced platform for materials development and state prediction in lithium-ion batteries. *Adv. Mater.* **34**:2101474. DOI:10.1002/adma.202101474
110. Xu Z., Duan H., Dou Z., et al. (2023). Machine learning molecular dynamics simulation identifying weakly negative effect of polyanion rotation on Li-ion migration. *npj Comput. Mater.* **9**:143. DOI:10.1038/s41524-023-01049-w
111. Sendek A.D., Cubuk E.D., Antoniak E.R., et al. (2019). Machine learning-assisted discovery of solid Li-ion conducting materials. *Chem. Mater.* **31**:342–352. DOI:10.1021/acs.chemmater.8b03272
112. Chouchane M., Yao W., Cronk A., et al. (2024). Improved rate capability for dry thick electrodes through finite elements method and machine learning coupling. *ACS Energy Lett.* **9**:1480–1486. DOI:10.1021/acsenenergylett.4c00203
113. Saad Y., Chelikowsky J.R. and Shontz S.M. (2010). Numerical methods for electronic structure calculations of materials. *SIAM Rev.* **52**:3–54. DOI:10.1137/060651653
114. Nascimento R.G., Corbetta M., Kulkarni C.S., et al. (2021). Hybrid physics-informed neural networks for lithium-ion battery modeling and prognosis. *J. Power Sources* **513**:230526. DOI:10.1016/j.jpowsour.2021.230526
115. Shorten C. and Khoshgoftaar T.M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* **6**:60. DOI:10.1186/s40537-019-0197-0
116. Chlap P., Min H., Vandenberg N., et al. (2021). A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **65**:545–563. DOI:10.1111/1754-9485.13261
117. Bailey J.J., Wade A., Boyce A.M., et al. (2023). Quantitative assessment of machine-learning segmentation of battery electrode materials for active material quantification. *J. Power Sources* **557**:232503. DOI:10.1016/j.jpowsour.2022.232503
118. Badmos O., Kopp A., Bernthaler T., et al. (2019). Image-based defect detection in lithium-ion battery electrode using convolutional neural networks. *J. Intell. Manuf.* **31**:885–897. DOI:10.1007/s10845-019-01484-x
119. Väyrynen A. and Salminen J. (2012). Lithium ion battery production. *J. Chem. Thermodyn.* **46**:80–85. DOI:10.1016/j.jct.2011.09.005
120. McLay A. (1996). Computational fluid dynamics: The basics with applications. *Aeronaut. J.* **100**:365. DOI:10.1017/s0001924000067129
121. Saw L.H., Ye Y., Yew M.C., et al. (2017). Computational fluid dynamics simulation on open cell aluminium foams for Li-ion battery cooling system. *Appl. Energy* **204**:1489–1499. DOI:10.1016/j.apenergy.2017.04.022
122. Tan C., Ardanese R., Huemiller E., et al. (2023). Data-driven battery electrode production process modeling enabled by machine learning. *J. Mater. Process.*

- Technol.* **316**:117967. DOI:10.1016/j.jmatprotec.2023.117967
123. Liu K., Wei Z., Yang Z., et al. (2021). Mass load prediction for lithium-ion battery electrode clean production: a machine learning approach. *J. Clean. Prod.* **289**:125159. DOI:10.1016/j.jclepro.2020.125159
124. Deringer V.L., Bartók A.P., Bernstein N., et al. (2021). Gaussian process regression for materials and molecules. *Chem. Rev.* **121**:10073–10141. DOI:10.1021/acs.chemrev.1c00022
125. Zhu P., Slater P.R. and Kendrick E. (2022). Insights into architecture, design and manufacture of electrodes for lithium-ion batteries. *Mater. Des.* **223**:111208. DOI:10.1016/j.matdes.2022.111208
126. Nair P., Vakharia V., Borade H., et al. (2023). Predicting Li-ion battery remaining useful life: An XDFM-driven approach with explainable AI. *Energies* **16**:5725. DOI:10.3390/en16155725
127. Zhang H., Li Y., Zheng S., et al. (2025). Battery lifetime prediction across diverse ageing conditions with inter-cell deep learning. *Nat. Mach. Intell.* **7**:270–277. DOI:10.1038/s42256-024-00972-x
128. Sui X., He S. and Teodorescu R. (2024). Small-sample-learning-based lithium-ion batteries health assessment: An optimized ensemble framework. *IEEE Trans. Ind. Appl.* **60**:4366–4380. DOI:10.1109/tia.2024.3351619
129. Zhai X., Liu G., Lu T., et al. (2025). Transforming waste to value: enhancing battery lifetime prediction using incomplete data samples. *J. Energy Chem.* **106**:642–649. DOI:10.1016/j.jechem.2025.03.011
130. Yu C., Lu T., Liu G., et al. (2025). Dimensional-noise-aware battery lifetime prediction via an EM-TLS framework. *Prog. Nat. Sci. Mater. Int.* **35**:146–155. DOI:10.1016/j.pnsc.2024.11.009
131. Lu T., Zhai X., Chen S., et al. (2024). Robust battery lifetime prediction with noisy measurements via total-least-squares regression. *Integr.* **96**:102136. DOI:10.1016/j.vlsi.2023.102136
132. Yu J., Yang J., Wu Y., et al. (2020). Online state-of-health prediction of lithium-ion batteries with limited labeled data. *Int. J. Energy Res.* **44**:11345–11360. DOI:10.1002/er.5750
133. Babaeiyazdi I., Rezaei-Zare A. and Shokrzadeh S. (2021). State of charge prediction of EV Li-ion batteries using EIS: a machine learning approach. *Energy* **223**:120116. DOI:10.1016/j.energy.2021.120116
134. Jospin L.V., Laga H., Boussaid F., et al. (2022). Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Comput. Intell. Mag.* **17**:29–48. DOI:10.1109/mci.2022.3155327
135. Zhang S., Liu Z. and Su H. (2023). State of health estimation for lithium-ion batteries on few-shot learning. *Energy* **268**:126726. DOI:10.1016/j.energy.2023.126726
136. Wang F., Zhai Z., Zhao Z., et al. (2024). Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. *Nat. Commun.* **15**:48779. DOI:10.1038/s41467-024-48779-z
137. Lin Y.-H., Ruan S.-J., Chen Y.-X., et al. (2023). Physics-informed deep learning for lithium-ion battery diagnostics using electrochemical impedance spectroscopy. *Renew. Sustain. Energy Rev.* **188**:113807. DOI:10.1016/j.rser.2023.113807
138. Nair P., Vakharia V., Shah M., et al. (2024). AI-driven digital twin model for reliable lithium-ion battery discharge capacity predictions. *Int. J. Intell. Syst.* **2024**:8185044. DOI:10.1155/2024/8185044
139. Vilalta R. and Drissi Y. (2002). A perspective view and survey of meta-learning. *Artif. Intell. Rev.* **18**:77–95. DOI:10.1023/a:1019956318069
140. Hospedales T.M., Antoniou A., Micaelli P., et al. (2022). Meta-learning in neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**:5149–5169. DOI:10.1109/tpami.2021.3079209
141. Ding S., Dong C., Zhao T., et al. (2021). A meta-learning based multimodal neural network for multistep ahead battery thermal runaway forecasting. *IEEE Trans. Ind. Inform.* **17**:4503–4511. DOI:10.1109/tii.2020.3015555
142. Cui X., Kang S.D., Wang S., et al. (2024). Data-driven analysis of battery formation reveals the role of electrode utilization in extending cycle life. *Joule* **8**:3072–3087. DOI:10.1016/j.joule.2024.07.024
143. Zhao S., Chen S., Zhou J., et al. (2024). Potential to transform words to watts with large language models in battery research. *Cell Rep. Phys. Sci.* **5**:101844. DOI:10.1016/j.xcrp.2024.101844
144. Liu R., Zou Z., Chen S., et al. (2025). Harnessing AI for understanding scientific literature: Innovations and applications of chat-agent system in battery recycling research. *Mater. Today Energy* **49**:101818. DOI:10.1016/j.mtener.2025.101818

FUNDING AND ACKNOWLEDGMENTS

The authors gratefully acknowledge the support from the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-01012), and National Natural Science Foundation of China (52272215). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Kai Wu, Yang Liu and Jiayu Wan supervised and revised the manuscript. Junzhe Zhang, Xiaolang Zhai and Qiang Zhang gathered the data and wrote the manuscript. Peiyi Li edited the manuscript. Mian Li revised the manuscript. All authors contributed to the discussion and composition of the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AND CODE AVAILABILITY

No new data were created or analyzed in this study.