nature communications

Article

https://doi.org/10.1038/s41467-025-58192-9

Data-driven energy management for electric vehicles using offline reinforcement learning

Received: 3 June 2024

Accepted: 12 March 2025

Published online: 22 March 2025

Check for updates

Yong Wang [1,2], Jingda Wu [1,2], Hongwen He[1,2], Zhongbao Wei [1] & Fengchun Sun[1]

Energy management technologies have significant potential to optimize electric vehicle performance and support global energy sustainability. However, despite extensive research, their real-world application remains limited due to reliance on simulations, which often fail to bridge the gap between theory and practice. This study introduces a real-world data-driven energy management framework based on offline reinforcement learning. By leveraging electric vehicle operation data, the proposed approach eliminates the need for manually designed rules or reliance on high-fidelity simulations. It integrates seamlessly into existing frameworks, enhancing performance after deployment. The method is tested on fuel cell electric vehicles, optimizing energy consumption and reducing system degradation. Real-world data from an electric vehicle monitoring system in China validate its effectiveness. The results demonstrate that the proposed method consistently achieves superior performance under diverse conditions. Notably, with increasing data availability, performance improves significantly, from 88% to 98.6% of the theoretical optimum after two updates. Training on over 60 million kilometers of data enables the learning agent to generalize across previously unseen and corner-case scenarios. These findings highlight the potential of data-driven methods to enhance energy efficiency and vehicle longevity through large-scale vehicle data utilization.

The automotive industry is undergoing a significant transformation, primarily due to the global focus on sustainability and environmental conservation. Electric vehicles (EVs) are leading this shift, playing a key role in mitigating environmental challenges and advancing sustainable transportation solutions[1,2]. Concurrently, the emergence of hybrid energy systems (HES) within the EV powertrain represents an emerging trend, offering superior solutions over single energy systems[3]. By integrating multiple energy sources such as batteries, fuel cells, and internal combustion engines, HES improves overall efficiency, sustainability, and reliability, while also providing adaptability to a wide range of driving conditions[4]. Propelled by rapid technological advancements and supportive policies, EVs equipped with HES, including Hybrid EVs (HEV), Plug-in hybrid EVs (PHEV), and Fuel cell EVs (FCEV), are gaining traction worldwide[5]. For example, the EV manufacturer BYD achieved sales of 3.02 million EVs in 2023, and PHEVs represent 47.9% of total sales. By 2024, the share of PHEVs increased significantly to 58.2% of total sales, highlighting their rising popularity in the market.

Given the increasing complexity and capabilities of HESs, effective energy management is crucial for optimizing their overall performance. An energy management strategy (EMS) serves as a vital component in EVs, specifically designed to regulate energy flow allocation among various sources within HES to achieve predefined operational objectives[6]. An EMS performs several critical functions essential for the optimal operation of EVs: (1) EMSs optimize system efficiency by intelligently allocating energy flow based on driving conditions,

[1]School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China. [2]National Key Laboratory of Advanced Vehicle Integration and Control, Beijing Institute of Technology, Beijing, China. e-mail: hwhebit@bit.edu.cn

reducing overall energy consumption and extending the driving range[7]; (2) EMSs optimize power delivery based on driver demand, enhancing acceleration and responsiveness while ensuring smooth power delivery for improved driving experience; and (3) By considering the unique characteristics of different power sources, EMSs extend the lifespan of HESs, thereby enhancing system reliability and safety[8]. Early EMSs primarily relied on various rule-based approaches to achieve energy-saving objectives. These strategies involve the creation of predefined rules and parameters tailored to specific driving conditions and vehicle characteristics, relying heavily on expert knowledge with iterative refinement via testing feedback[9]. However, while effective, this approach can be labor-intensive and time-consuming, requiring significant manual expertize/effort for rule formulation and extensive experimentation for parameter calibration. Moreover, the static nature of rule-based EMSs limits their adaptability to dynamic driving scenarios, thereby reducing their effectiveness in maximizing energy savings and overall performance.

To address the complex challenge of energy management, various EMS approaches have been developed over the past decades. Among these, optimal control theory-based methods, such as dynamic programming (DP)[10] and model predictive control (MPC)[11], are widely adopted. These strategies, often categorized as prediction-based EMSs, excel in achieving near-optimal energy allocation by leveraging systematic modeling and predictive capabilities. However, while DP and MPC provide sophisticated solutions utilizing future driving data, accurately forecasting future driving conditions based on historical speed patterns and dynamic traffic information remains a considerable challenge[12]. Additionally, prediction models and optimization algorithms often introduce considerable computational complexity, leading to suboptimal real-time performance[3]. Consequently, the application of prediction-based EMS methods in real-world vehicle scenarios remains limited.

In response to these challenges, advanced machine learning (ML) techniques have emerged as powerful tools for energy management in HESs, with reinforcement learning (RL) (especially deep reinforcement learning (DRL)) marking a key milestone in this field. The optimal formulation of RL is based on a Markov Decision Process (MDP), comprising an environment and an agent. In this framework, the RL agent interacts with the environment to learn strategies that maximize cumulative rewards over time[13]. Unlike prediction-based EMS methods, RL determines optimal actions through a trial-and-error process, eliminating the need for prior knowledge of system mathematical modeling or future driving conditions[14]. DRL algorithms, such as Deep Deterministic Policy Gradient (DDPG)[15], Soft Actor-Critic (SAC)[16], and Proximal Policy Optimization (PPO)[17], have demonstrated impressive results in addressing EMS problems. These approaches, collectively termed simulation-based EMSs, leverage high-fidelity EV simulators to safely train DRL agents in developing near-optimal strategies. A key advantage of DRL lies in its self-learning ability, which allows it to autonomously derive effective and adaptive strategies[18]. However, applying simulation-based DRL methods to real-world vehicle tasks is constrained by sample inefficiency and safety concerns. DRL adopts an online learning paradigm, where agents typically require extensive interactions with the environment to learn effective policies[19]. In real-world scenarios, direct interaction with vehicles poses safety risks, as the agent may execute suboptimal or unsafe actions during the learning process. Additionally, while existing studies often assume that simulation models accurately replicate real-world conditions, constructing high-fidelity models that comprehensively capture vehicle dynamics, powertrains, traffic scenarios, and driver behavior remains a considerable challenge[20]. This limitation can result in the “sim-to-real" problem, where EMS strategies developed in simulators fail to transfer effectively to real vehicles, further complicating the development and deployment of EMS solutions.

Recently, advancements in ML and the growing availability of large datasets have made data-driven methods essential for addressing major challenges in the EV industry[21,22]. With support from data collection platforms and open-access laboratory data, these data-driven approaches have revolutionized various aspects of battery management systems (BMSs)[23]. Some applications include the automatic discovery of complex battery aging mechanisms[24], prediction of battery safety envelopes[25], evaluation of safety conditions[26], estimation of battery state of health[27], and even enhancing battery lifetime prediction models using unlabeled data[28]. Notably, innovations in feature extraction and supervised ML techniques tailored for time-series data have significantly enhanced prediction accuracy[29]. This progress has sparked interest in exploring data-driven methods to sequential decision-making tasks, including improving energy management systems. A common approach for implementing a data-driven EMS involves using supervised learning, where the ML model captures complex and non-linear relationships between input features and corresponding control outputs. In refs. 30–32, deep neural networks are trained offline using substantial amounts of training data obtained from the global optimization strategy of DP, yielding a near-optimal EMS that closely approximates the DP. It is essential to distinguish this from prediction tasks in BMS applications, as EMS entails sequential decision-making. Although supervised learning can mimic the EMS policy through imitation learning, its heavy reliance on expert data may result in limited generalization to new and diverse scenarios[33,34]. In refs. 35,36, the application of data-driven DRL for energy management was explored, with the agent learning the EMS from data generated by online DRL. While this approach showed promise, a challenge emerged when the dataset quality was poor, hindering successful learning. Therefore, it is crucial to investigate alternative methods for learning EMS from non-expert or suboptimal data, a scenario commonly encountered in automotive applications.

In this study, we propose a data-driven EMS paradigm that learns solely from pre-generated data from existing EMS methods, eliminating the need for explicit rule design or online interaction. Our approach integrates DRL with supervised learning, advancing beyond traditional rule-based, prediction-based, and simulation-based RL approaches to enhance EMS capabilities, as illustrated in Fig. 1(a). Compared to existing data-driven methods, the main feature of our approach is its ability to learn a near-optimal EMS from large-scale, suboptimal data. Even with poor suboptimal data, it can still achieve successful learning, making it highly practical. Notably, this method can be seamlessly integrated with any EMS algorithm and continuously improve as new data is collected. We term this approach the offline reinforcement learning (ORL) agent, which incorporates a blended policy regularization (BPR) that facilitates effective exploration and ensures constraint satisfaction throughout the learning process. The ORL agent is trained using data from an augmented-reality EV platform, which combines real-world operational data from an EV monitoring and management system in China (Fig. 1(b)) with a high-fidelity simulated FCEV powertrain model. We focus on the FCEV as the subject, collecting data for learning power allocation strategies. Note that the proposed method is also applicable to EMS for other types of vehicles with HESs, including HEVs and PHEVs. Overall, the ORL agent introduced here exhibits four key features (Fig. 1(c)):

1. Purely data-driven EMS: By autonomously learning and optimizing from collected offline datasets, our approach facilitates the development of advanced EMS without necessitating expert knowledge or the creation of comprehensive high-fidelity simulators incorporating both vehicle models and traffic scenarios. This data-driven process significantly streamlines EMS development workflows.
2. Learning from non-optimal data: Our research demonstrates that the ORL agent can learn near-optimal EMS from non-optimal data, and even derive superior policies from poor suboptimal EMS data. This approach is less reliant on data quality, allowing for effective

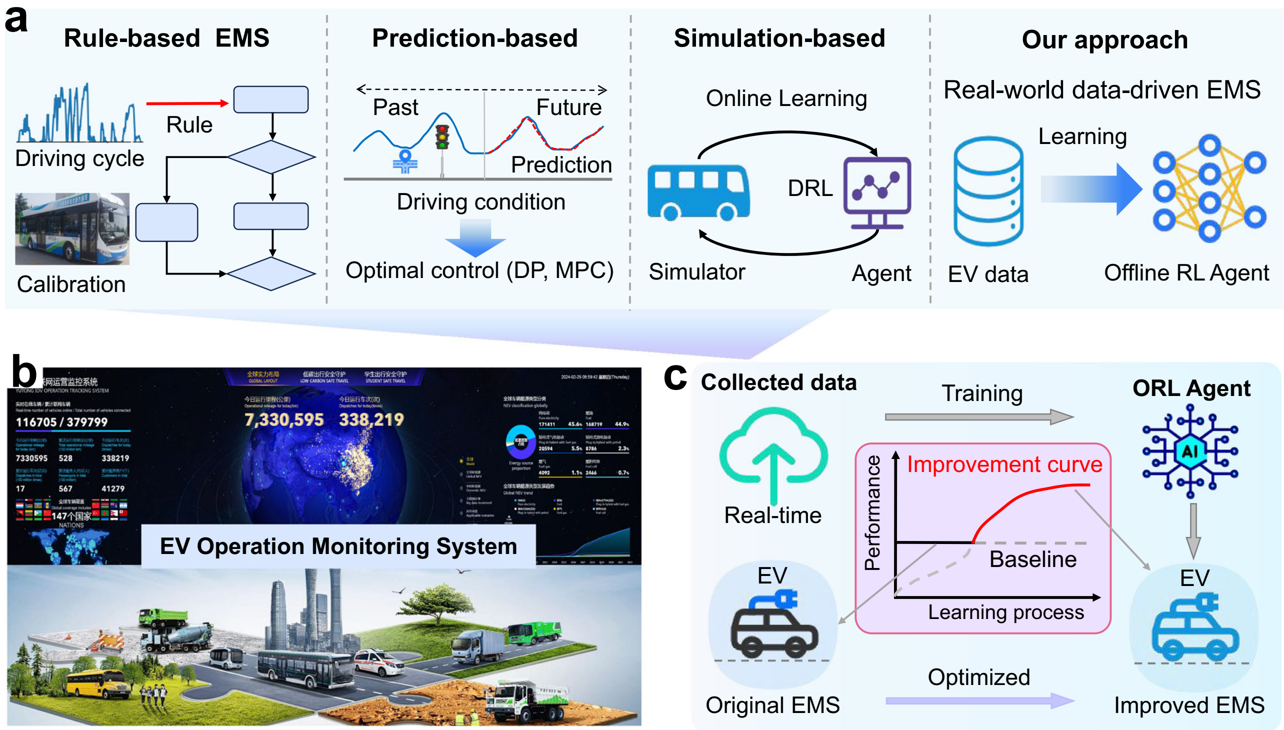

Fig. 1 | Overview of the proposed data-driven EMS framework. a Comparison of the four EMS paradigms: Traditional rule-based EMS relies on expert knowledge and calibration based on fixed driving cycles. Prediction-based methods, such as DP and MPC, rely on future driving data to operate. Simulation-based EMS requires high-precision models and entails transitioning from simulation analysis to real-world deployment, resulting in a gap between simulated and real-world performance (sim-to-real gap). In contrast, the proposed real-world data-driven EMS learns directly from actual data. **b** China has established a three-tier EV monitoring and management system involving the state, local governments, and public enterprises. The National Monitoring and Management Platform collects real-time operational data from over 20 million EVs[49]. **c** The ORL agent works with an existing EMS and continuously collects EV data to improve the EMS.

learning. Its practicality enables the use of raw data generated by actual vehicles, a common scenario in automotive applications.

3. Enhancement with increased data: The ORL agent supports continuous learning by collecting new data online and updating its knowledge offline. The performance improves as more training data is incorporated, highlighting its ability to continuously adapt and enhance EMS performance. By training across diverse datasets, it can potentially adapt to new driving conditions and produce favorable results, even in corner-case situations.
4. Compatibility with existing EMS: Our approach seamlessly integrates with existing rule-based or simulation-based EMS methods, leveraging data from onboard controllers to augment EMS performance. This ensures that baseline performance is maintained while facilitating further improvements via ORL, making it a valuable extension to conventional EMS methodologies.

Results

The overview of data-driven EMS

In Fig. 2(a), the framework overview of ORL for EMS is illustrated. We present the three phases of applying the proposed ORL algorithm to the EMS problem: data collection, offline learning, and evaluation. ORL is an subset of RL methods based on data-driven approaches. Unlike traditional simulation-based RL EMS scheduling methods, the data-driven learning process of ORL does not require online interaction with an EV simulation environment. Instead, it learns solely from pre-generated data obtained from existing EMS methods. To efficiently collect large-scale EV datasets and evaluate the performance of the trained ORL agent during the evaluation phase, an augmented-reality EV platform is developed. This platform integrates real-world operational data from an EV monitoring and management system with a high-fidelity simulated FCEV powertrain model. As shown in Fig. 2(b), the platform synchronizes the movements of a real FCEV on a physical test track with its virtual counterpart, enabling the real vehicle to interact with virtual vehicles in a realistic traffic environment. Real driving conditions serve as input to the powertrain system, which iteratively interacts with existing EMS methods to generate high-quality data. Specifically, raw data such as vehicle speed, battery voltage, current, FC system power, and motor power are collected through the EV monitoring and management system. These raw inputs are processed by the simulated FCEV powertrain system to generate refined outputs, such as acceleration, hydrogen and electricity consumption, FC degradation, battery state of charge (SOC), battery degradation, and accurately derived values for FC power, power variations, and motor power.

In the data collection phase, the augmented-reality platform facilitates the efficient creation of comprehensive, large-scale, and high-precision EMS datasets. Metrics like hydrogen consumption and degradation, which are difficult to measure directly and often lack accuracy in real-world tests, are precisely derived through the high-fidelity simulation model. Additionally, the EV monitoring and management system effectively captures real driving conditions and driver behavior, which are otherwise challenging to replicate in purely simulated environments. By combining both real-world data and simulated model, the platform ensures the generation of comprehensive and reliable EMS datasets. In this study, EMS datasets of varying sizes are constructed, with the largest dataset encompassing over 60 million kilometers of driving data. The collected data are further processed into a standardized format, suitable for EMS applications. During the data encoding process, raw data from the augmented-reality platform are converted into a transition dataset, $\mathcal{D}=(s,a,s',r)_i$, organized in a time-series structure (i). Here, s represents the state, a the action, s' the next state, and r the reward. Detailed descriptions of the states, actions, and reward functions are provided

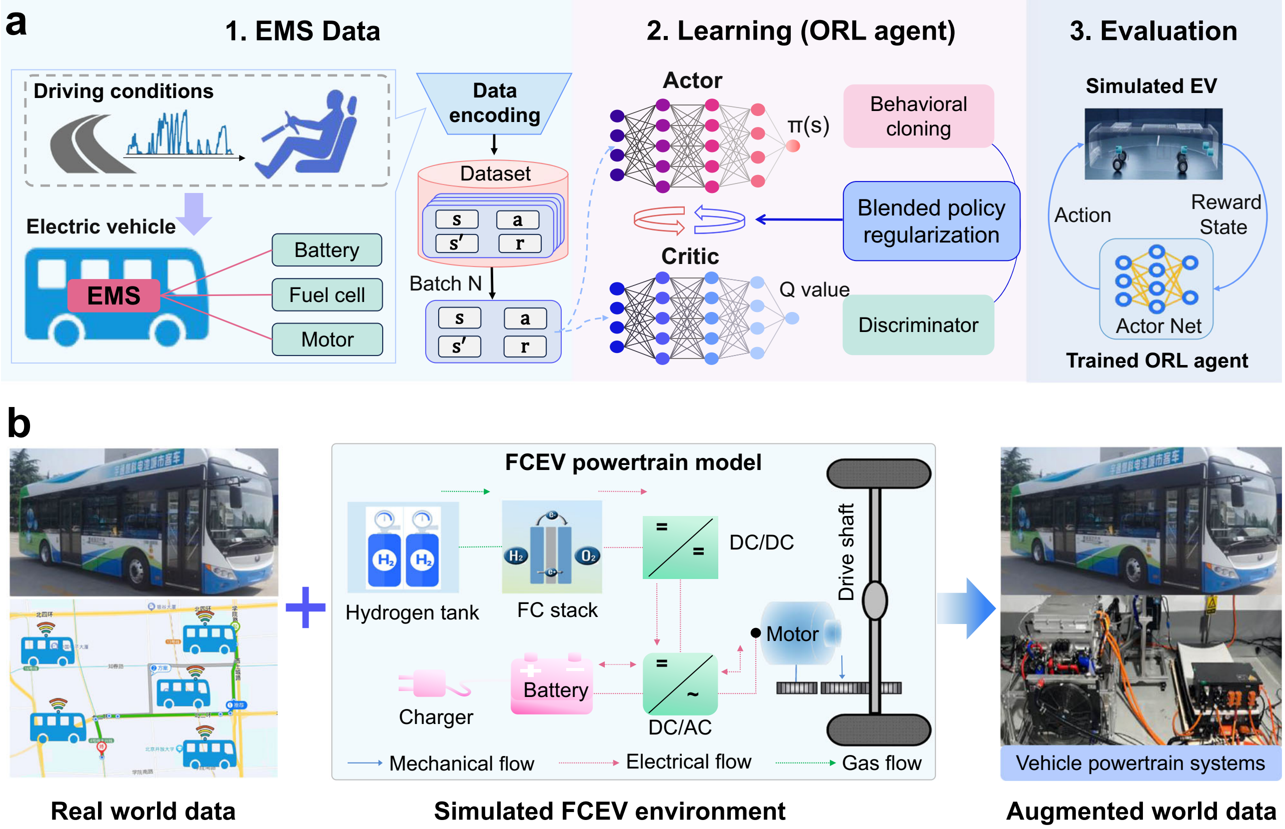

Fig. 2 | Overall diagram of the proposed data-driven EMS methodology. **a** The framework overview of ORL for EMS, including data collection, offline learning, and evaluation. **b** The augmented-reality testing platform enhances real-world operational data from an EV monitoring and management system with a high-fidelity simulated FCEV powertrain system, creating an efficient data collection and algorithm-testing environment for EVs. FC: fuel cell, DC/DC: DC-to-DC converter, DC/AC: DC-to-AC converter, DC: direct current, AC: alternating current.

in the Methods section. The encoded state-action-reward sequences are stored in an experience replay buffer, serving as the basis for subsequent policy learning.

In the offline learning phase, we propose Actor-Critic with BPR (AC-BPR), a novel ORL method addressing limitations of traditional approaches. AC-BPR incorporates BPR, which combines Behavior Cloning (BC)[37] and Discriminator-based Regularization (DR)[38]. BC ensures policy alignment with expert behavior by minimizing divergence, while DR employs an adversarial module to encourage exploration in high Q-value regions. This strategic balance between conservatism and exploration allows AC-BPR to effectively improve performance, even when learning from suboptimal or lower-quality datasets. In each training step, mini-batches of transitions (s, r, a, s') stored in a data buffer are sampled to update the Actor-Critic networks. The Actor network selects actions based on the current policy ($\pi(s)$), which is guided by both expert behavior (via BC) and exploration of high Q-value(through DR). The Critic network evaluates the actions by estimating the Q-values for state-action pairs, providing feedback to the Actor for policy refinement. AC-BPR can be seamlessly integrated with any Actor-Critic algorithm and is practically implemented using the Twin Delayed Deep Deterministic Policy Gradient (TD3)[39] framework, featuring a highly efficient design (as detailed in the Methods section). Both empirical results and theoretical analysis demonstrate that AC-BPR effectively mitigates distribution shifts in ORL by introducing blended regularization.

Upon completion of the training phase, the neural network parameters representing the EMS of the ORL agent are saved for future use. Subsequently, the trained ORL agent is evaluated to gauge its effectiveness and performance. Utilizing the FCEV environment established during the data collection phase, we conduct experiments with three standard driving cycles (WTVC: World Transient Vehicle Cycle; CHTC: China Heavy-duty Commercial Vehicle Test Cycle; FTP: Federal Test Procedure 75) to evaluate energy costs. Additionally, various real-world driving scenarios are incorporated to comprehensively assess the trained EMS. Following the evaluation, adjustments to the agent's hyperparameters or training process may be made to improve its performance. This iterative process of training, evaluation, and refinement continues until the EMS attains the desired level performance. Once satisfactory performance is reached, the trained agent becomes eligible for deployment in real-world scenarios, where it can be utilized to efficiently optimize energy management systems.

Data for learning and analysis

We select the PPO as the expert EMS, as it demonstrates the best performance among online DRL algorithms for our EMS problem. Details regarding the performance of different EMS algorithms are presented in Table S1. Using PPO, we generate datasets, denoted as $\mathcal{D}^E$, comprising 300e3 time steps. Additionally, we employ a random agent that samples actions randomly, generating datasets, denoted as $\mathcal{D}^R$, which represent poor performance. To create settings with varying levels of data quality in the suboptimal offline dataset, we combine transitions from the expert datasets $\mathcal{D}^E$ and the random datasets $\mathcal{D}^R$ in different ratios. Specifically, we consider four different dataset compositions, denoted as D1, D2, D3, and D4, defined as follows: D1 (Data-1): Consists solely of transitions from the expert dataset $\mathcal{D}^E$, representing the expert policy. D2 (Data-2): Contains two-thirds of transitions from the expert dataset $\mathcal{D}^E$ and one-third from the random dataset $\mathcal{D}^R$, representing suboptimal data. D3 (Data-3): Comprises one-third of transitions from the expert dataset $\mathcal{D}^E$ and two-thirds from the random dataset $\mathcal{D}^R$, representing another form of suboptimal data. D4 (Data-4): Composed solely of

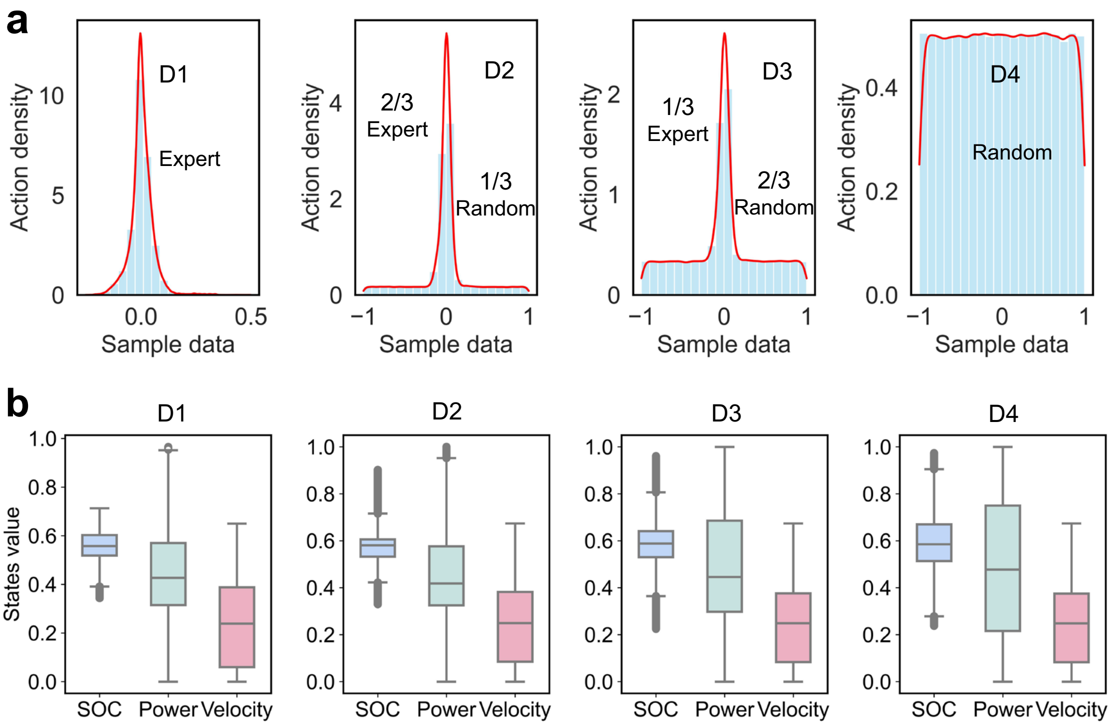

Fig. 3 | Comparison of different datasets. a Distribution of encoded actions for the four datasets, with each action normalized to the range [−1,1]. D1 represents data generated by the PPO expert policy; D2 and D3 denote suboptimal data generated by a combination of expert and random policies; and D4 comprises entirely random data. **b** The state distribution of four datasets, including battery SOC expressed as a percentage, fuel cell system output power scaled to the range [0, 1], and velocity also normalized to the range [0, 1].

transitions from the random dataset $\mathcal{D}^R$, representing the random policy.

Figure 3 (a) depicts the action distributions for the four datasets, revealing significant differences among the four EMS policies. The action range for D1 falls within (−0.2, 0.5), indicating relatively stable variations in FC power. In contrast, the introduction of random policy data broadens the action ranges for the other datasets, all spanning (−1, 1). Notably, D4 exhibits a uniformly distributed action range across (−1, 1), indicating that this policy is noisy and represents a poor EMS. Figure 3(b) illustrates the state distributions for the four datasets, where all states have undergone post-processing and scaling to the (0, 1) interval. Comparing the box plots of the four datasets reveals that the SOC of D1 remains within a reasonable range (0.38–0.7), adhering to EMS constraints for battery SOC. However, the SOC of the other datasets falls into unreasonable ranges, such as (0.2, 1) for D3. Additionally, with the increase in $\mathcal{D}^R$ data, the FC power distribution ranges in D3 and D4 become wider. Since the conditions of the four datasets are derived from fixed segments of standard driving cycles, the velocity distribution remains the same across all datasets.

Creating challenging datasets is practical as generating suboptimal or random data is more cost-effective than collecting expert-level data from real vehicles. Consequently, an effective data-driven EMS method must be able to effectively handle and learn from these suboptimal offline datasets.

Learning superior EMS from non-optimal data

We first examine the performance of the ORL agent with different datasets. To ensure a fair comparison, the algorithm employs uniform experimental settings and network parameters across all four datasets. Figure 4(a) illustrates the average reward during the training process for each dataset. This average is computed as the mean reward over every 1000 training steps and validated across 10 iterations using three standard driving cycles: WTVC, CHTC, and FTP. The training process involves utilizing a buffer comprising 300e3 samples, with the ORL agent randomly selecting 256 data points for each training iteration, totaling one million training epochs. For D1, which comprises exclusively expert data, convergence is observed after approximately 210e3 episodes. However, the ORL agent exhibits slower convergence speed during iterative learning on the D2, D3, D4 datasets, converging at around 330e3, 600e3, and 360e3 steps, respectively. This suggests that the data distribution considerably influences the learning speed. Nevertheless, the ORL agent ultimately succeeds in learning an effective EMS.

Figure 4 (b) presents the reward performance of trained ORL agents across three driving cycles. Notably, the absolute reward value achieved via ORL decreases in D1, from 323.4 to 297.3 in the CHTC cycle, representing an improvement of 8.8%. Surprisingly, even when trained on suboptimal datasets D2 and D3, ORL outperforms the expert strategy, achieving reward increases of 1.8 and 3.4%, respectively. Similarly, under WTVC and FTP conditions, the ORL agent showcases superior performance, learning more effective strategies from the suboptimal datasets D2 and D3. An exception occurs with D3 under the WTVC condition, likely due to the high-speed nature of this cycle, leading to larger reward values for SOC. Despite this, the final energy consumption results remain within reasonable limits. Particularly noteworthy is the exceptional performance of the ORL agent on the random dataset D4, where it closely approaches expert-level results across all three validation conditions, achieving rewards of 402, 325, and 395. Compared to the original average reward of 2637 for the D4 dataset, ORL has reduced the reward by 85.8%. The enhanced performance of the ORL agent on suboptimal or non-expert datasets can be attributed to the proposed AC-BPR algorithm. By employing blended regularization techniques, ORL effectively balances conservative imitation with exploratory learning. This enables the ORL agent to maintain robust performance across diverse data qualities, achieving superior results on D1, D2, and D3, while also effectively exploring and optimizing the random dataset D4.

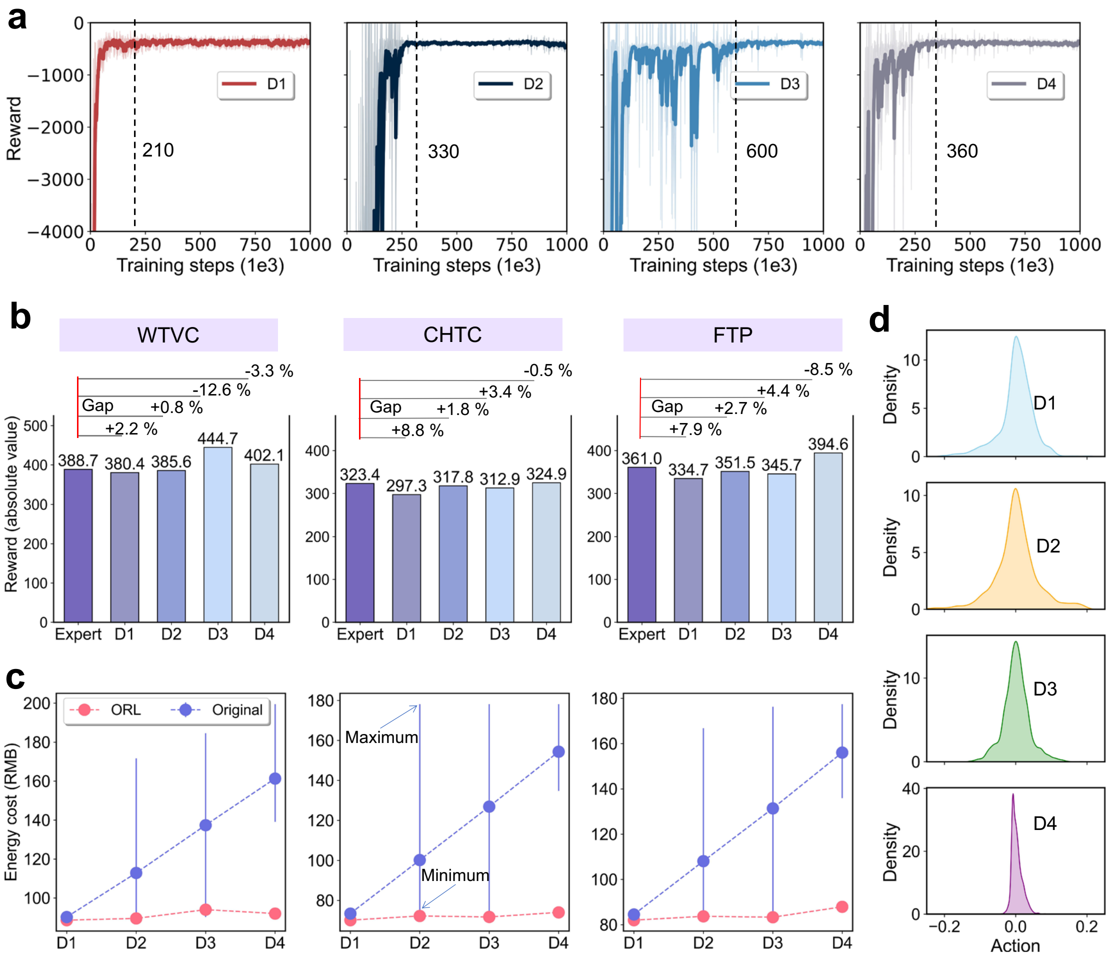

Fig. 4 | Learning performance of the ORL agent. a Learning curves of the ORL agent for the four different datasets. **b** The comparison of absolute rewards (original rewards are negative) under three validation conditions. Expert refers to the original D1 dataset generated by the PPO policy, while D1, D2, D3, and D4 correspond to the best rewards achieved by ORL after learning on each respective dataset. Notably, the ORL agent's performance on various datasets closely approximates or exceeds the expert policy. **c** The comparison of energy costs between the original EMS and the optimized EMS using ORL demonstrates a significant reduction in energy costs via the data-driven learning process. **d** The action distributions (FC power slopes) of the optimized EMS using ORL.

Figure 4 (c) provides a detailed comparison of the energy costs between the original EMS datasets and the optimized EMS using ORL. The blue dots represent the mean energy costs for the four original EMS datasets, with error bars indicating range between the maximum and minimum costs. In contrast, the red dots depict the energy costs incurred by ORL on the corresponding datasets. Despite the inclusion of random data, which degrades cost performance in the original datasets, ORL consistently achieves lower costs across all data sets. For instance, under the WTVC condition, the energy cost escalates from the initial 90 RMB in dataset D1 to 163 RMB in dataset D4. However, ORL consistently helps maintain costs within the narrow range of 90–95 RMB. In particular, the minimum cost values in the original D4 dataset significantly exceed those achieved by the ORL agent, which reduces costs by over 40% across all three conditions. These results underscore the ORL agent's ability not only to leverage expert EMS for superior outcomes but also to consistently deliver excellent performance from increasingly suboptimal datasets. Remarkably, the ORL agent even attains expert-level EMS performance when trained solely on noisy datasets.

To elucidate the rationale behind the performance improvements, Fig. 4(d) illustrates the action distributions of the optimized EMS using ORL. As different EMS policies can be reflected by the actions taken, in the context of the FCEV considered here, this pertains to the FC power slope under the same driving cycle. Comparing Figs. 3(a) and 4(d), significant changes are observed in D2, D3, and D4 with respect to Fig. 3(a). In D2, D3, and D4, the action distributions closely resemble those of expert data in D1, concentrating within the range of [-0.3, 0.3], as opposed to the wider range of [-1, 1] seen in Fig. 3(a). This change is particularly pronounced in D4, where the lack of expert data results in slight differences in the action distributions compared to D1, D2, and D3. However, all ORL policies consistently learn FC power variations with smaller ranges, ensuring smoother FC power output while effectively meeting the power demand requirements.

In conclusion, experimentation across three validation conditions and four datasets, our ORL agent demonstrates robust performance across diverse dataset conditions, including expert, suboptimal, and random datasets. By incorporating the AC-BPR algorithm, which balances BC with discriminator-based regularization, the ORL agent can effectively learn from non-optimal datasets.

Performance by comparative evaluation

To demonstrate the superior performance of ORL, we contrast it with simulation-based and imitation learning EMS approaches. Since imitation learning and ORL are closely related, both involve learning EMS from data. We first compare the performance of ORL with that of BC. Notably, BC typically employs a supervised learning paradigm, relying solely on expert data, while the ORL agent incorporates RL with exploration mechanisms. This distinctive learning mechanism results in significant performance differences between the two methods.

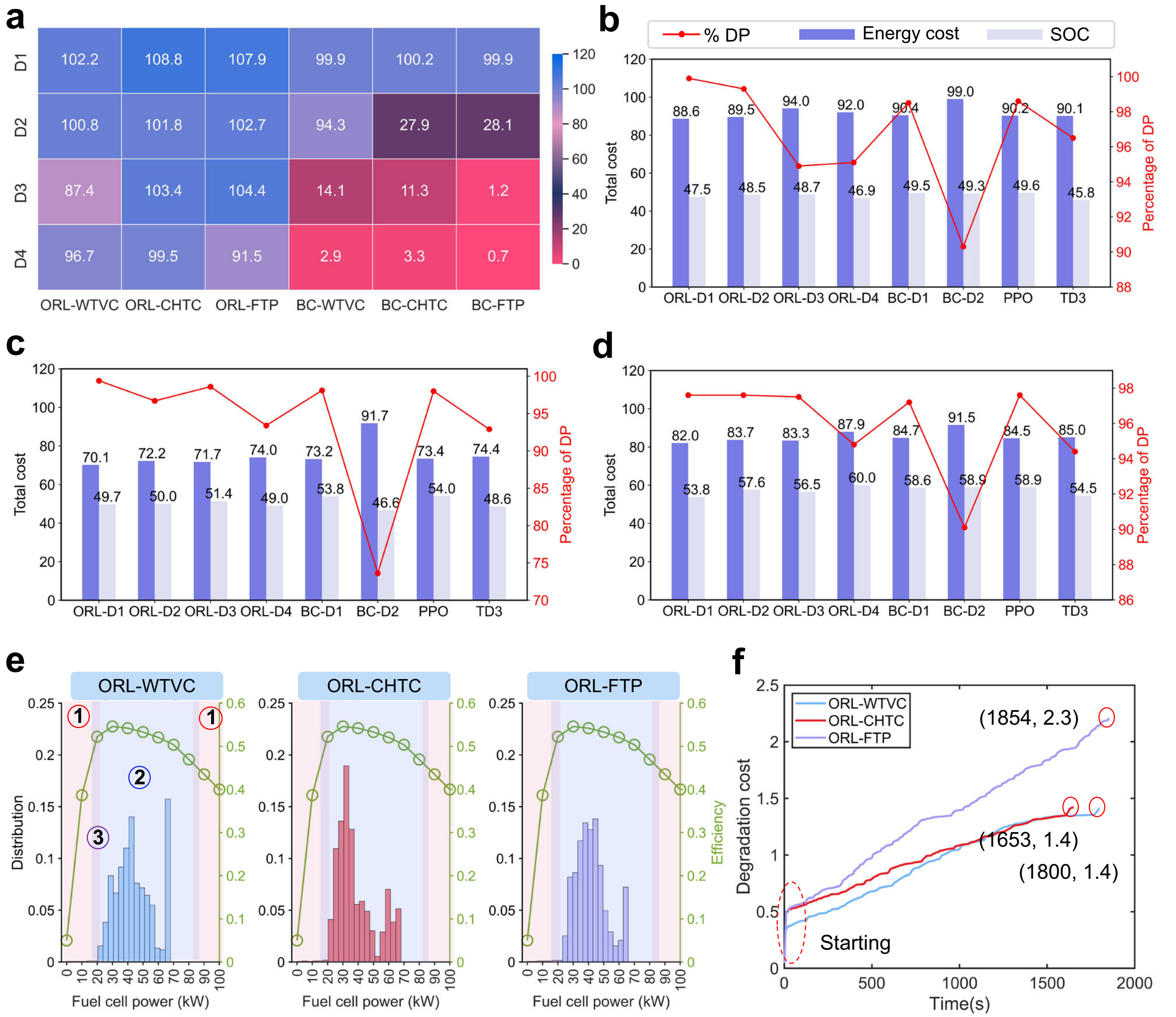

Fig. 5 | Performance analysis comparing different methods. a The comparison between two data-driven EMS methods: The matrix numbers represent the relative reward rates of ORL and BC compared to the expert EMS (PPO) under the same conditions, emphasizing the minimal influence of data quality on the ORL agent's performance. **b** Comprehensive performance of different algorithms under the WTVC condition, with DP representing the globally optimal EMS. **c** Comprehensive performance under the CHTC condition. **d** Comprehensive performance under the FTP condition. **e** The distribution of FC system power across the efficiency curve and degradation regions, where Region 1, Region 2, and Region 3 represent the high-degradation zones, the high-efficiency range, and the overlapping area, respectively. **f** FC degradation costs under three conditions.

In Fig. 5(a), we compare the testing rewards across the WTVC, CHTC, and FTP driving cycles, and calculate the percentage of ORL and BC costs relative to the expert EMS (PPO). In D1, both ORL and BC achieve favorable results, with ORL surpassing the original expert data by a maximum of 8.8%, while BC remains comparable to the expert. In D2, ORL maintains superiority over expert-based EMS, while BC experiences significant cost degradation (ranging from 6 to 70%). In D3 and D4, the ORL agent continues to outperform or closely match the expert, while BC, constrained by data quality, fails to learn an optimal EMS. This underscores ORL's ability to learn superior EMS from non-expert data, while imitation learning demonstrates poorer performance and struggles to learn favorable EMSs from non-expert data.

Figure 5 (b–d) presents detailed results of different methods under the WTVC, CHTC, and FTP conditions, with the red lines representing the percentage of cost compared to DP. It is evident that ORL learns an optimal EMS on the D1 dataset, achieving percentages close to 99.9, 99.4, and 97.6% of DP, respectively. In comparison, PPO, as a benchmark expert policy, yields cost results 98.6, 98.0, and 97.6% of DP, respectively. Thus, while BC learns a similar expert EMS in D1, its performance significantly deteriorates on suboptimal D2 data. Another online DRL method, TD3, also demonstrates satisfactory performance; however, its overall costs are higher than those of PPO and ORL.

In Fig. 5(e), the FC power distribution of the EMS learned by ORL on the D1 dataset is illustrated, mapped against its efficiency curve and degradation-prone regions. Power levels below 20% of the FC's maximum power (low-power operation) and above 80% of its maximum power (high-power operation) are identified as high-degradation zones (indicated as Region 1). In contrast, power levels corresponding to an efficiency exceeding 45% are classified within the high-efficiency region, representing optimal energy utilization (indicated as Region 2). Region 3 in the figure highlights the overlapping area between the high-degradation and the high-efficiency region. Under all three driving conditions, the power distribution guided by the ORL-based EMS is primarily concentrated in the high-efficiency region, with

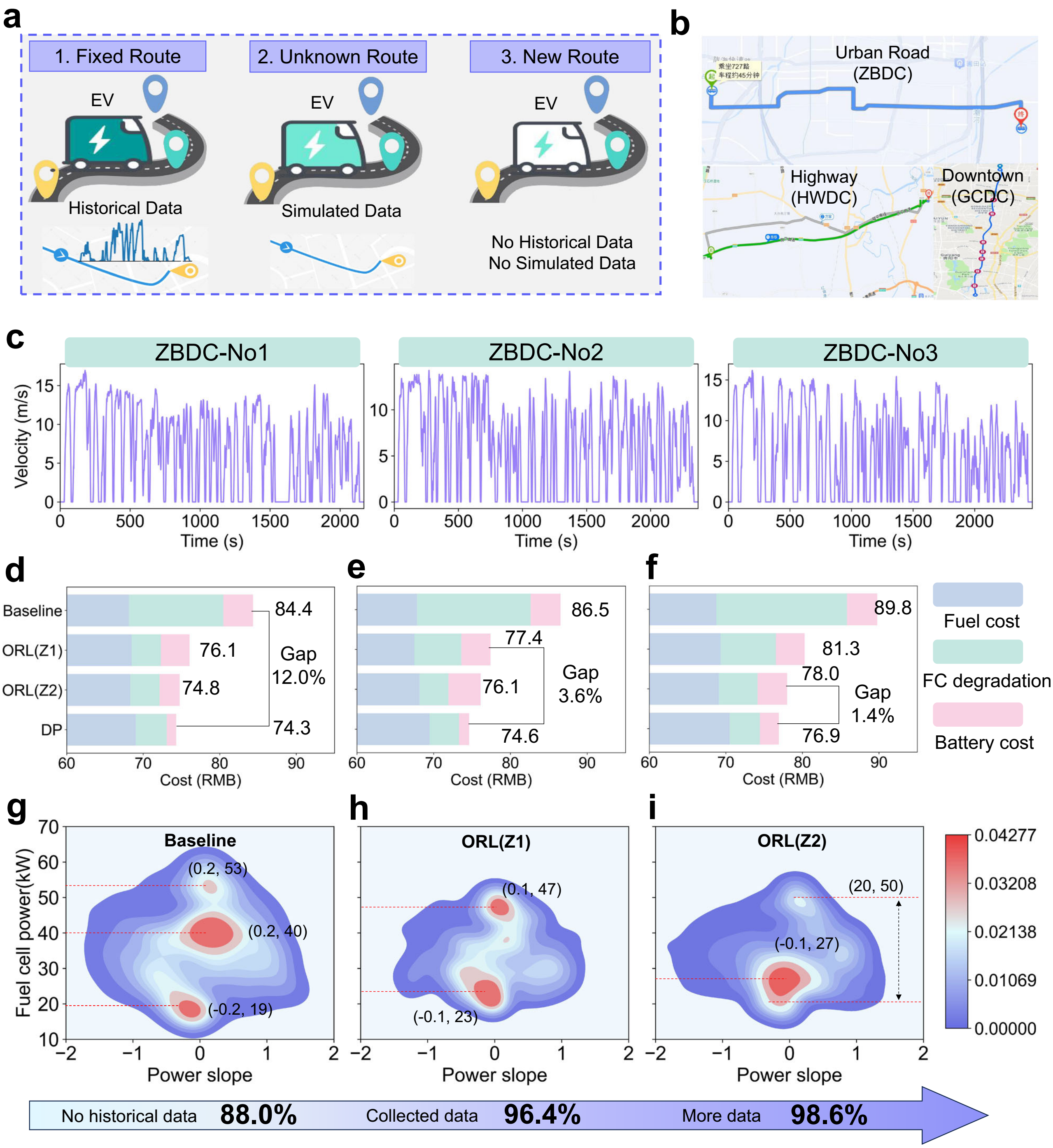

Fig. 6 | The ORL agent continuously learning and improving from data. a ORL for continuous learning from data in three scenarios. **b** Driving data collected from the real-world route. ZBDC: Zhengzhou Bus Driving Cycle; HWDC: Highway Driving Cycle; GCDC: Guiyang City Driving Cycle. **c** Speed trajectories of three ZBDC conditions. **d** Comparing the total cost under ZBDC-No. 1, which includes hydrogen consumption, battery cost, and cell degradation cost. **e** Comparing the total cost under ZBDC-No. 2. **f** Comparing the total cost under ZBDC-No. 3. **g** FC power distribution cloud chart of baseline EMS. **h** Cloud chart of FC power distribution following one data update. **i** FC power distribution cloud chart following two data updates.

minimal operation in the high-degradation zones. This underscores the effectiveness of the ORL in learning an optimized EMS from data, ensuring the FC system operates predominantly in the high-efficiency range, thereby reducing both hydrogen consumption and overall system costs. Additionally, a narrower power variation range, as shown in Fig. 5(d), minimizes FC degradation costs. As illustrated in Fig. 5(f), ORL incurs minimal FC degradation costs across the WTVC, CHTC, and FTP conditions, with costs of 2.3, 1.4, and 1.4, respectively. Furthermore, examination of Fig. 5(b–d) indicates that the battery SOC remains within a reasonable range. These findings collectively affirm that the ORL agent has successfully learned a superior EMS.

Continuous learning with growing data

We have demonstrated in previous experiments that the ORL can learn optimal EMS strategies from data and outperform other methods. In this section, we further showcase an ORL approach for continuous learning from data. We conduct experiments pertaining to three cases depicted in Fig. 6(a), collecting real-vehicle data in different driving

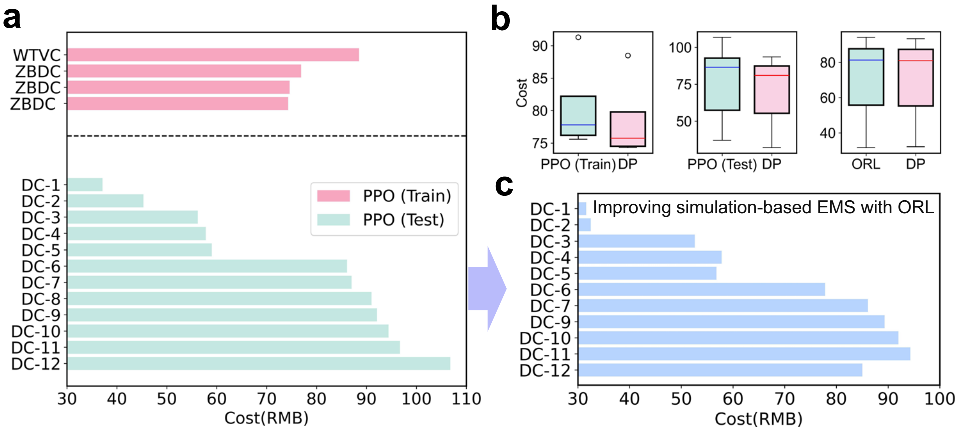

Fig. 7 | Evaluation of ORL performance enhancement using simulation-based EMS. a Performance of PPO trained on four datasets and tested on 12 testing datasets. **b** Comprehensive performance comparison of different EMS methods, revealing that ORL can significantly mitigate performance degradation observed in testing phase of PPO. **c** Performance of ORL across 12 testing datasets.

scenarios, including urban roads, highways, and downtown roads for the three cases (Fig. 6(b)). Notably, the training datasets used here differ from those (D1, D2, D3, D4) in the previous sections. The training data for Case 1 and Case 3 consists of EMS data generated by the augmented-reality EV platform, which utilizes real-world operational data as input. In Case 2, the data combines results from a simulation-based RL strategy applied to standard driving cycles, with a subset of real-world driving conditions.

Case 1: continuous learning from historical data. In Case 1, we illustrate the concept by using the example of driving a bus on fixed routes. Real electric bus driving data was collected in Zhengzhou, China, over three consecutive days. Figure 6(c) shows the speed trajectories of the bus for each of the three days, labeled ZBDC-No. 1, ZBDC-No. 2, and ZBDC-No. 3. Noticeable variations in the speed trajectories are observed along the same route over different days. Figure 6(d–f) shows the total cost of different EMS strategies across the three scenarios. These costs include hydrogen consumption, battery costs, and FC degradation. The baseline is the original EMS of the FCEV, which is used as a reference. For the first scenario, we use the baseline data from the ZBDC-No.1 driving cycle to train the ORL agent, yielding the ORL(Z1) strategy. This strategy is then applied to the new condition ZBDC-No. 2. Furthermore, a new ORL EMS, ORL(Z2), is trained using data from both the baseline data from ZBDC-No. 1 and the previously learned ORL(Z1) strategy from ZBDC-No. 2. This new strategy is then validated on the final driving cycle ZBDC-No. 3.

The baseline EMS demonstrates poor performance on the first day (ZBDC-No. 1), achieving 88.0% of the cost efficiency compared to DP. The corresponding FC power and power slope distributions are depicted in Fig. 6(g). On the second day, as illustrated in Fig. 6(e) and (h), the ORL(Z1) strategy significantly improves by learning from the previous data, achieving 96.4% of DP's cost efficiency under the ZBDC-No. 2 conditions. By the third day, after continuously learning from additional data, the ORL(Z2) further enhances cost efficiency, achieving 98.6% of DP's performance on ZBDC-No. 3, as shown in Fig. 6(f) and (i). A comparison of the power distributions across the three scenarios highlights that ORL(Z1) and ORL(Z2) allocate a greater proportion of FC output power to the high-efficiency range (Fig. 6(g), (h), and (i)). This adjustment not only reduces overall energy consumption but also minimizes the power slope, effectively lowering system degradation costs.

In conclusion, with continuous data updates, new information can be effectively utilized to train the ORL agent, enabling the development of progressively optimized EMS strategies. This demonstrates the capacity of ORL for continuous learning and improvement from historical data. Additionally, our approach integrates seamlessly with established EMS frameworks by leveraging real-time data from onboard controllers to enhance EMS performance. This integration ensures the preservation of baseline performance while facilitating further improvements through ORL, making it a valuable and adaptable extension to conventional EMS methodologies.

Case 2: improving from simulated data. Simulation-based EMS offers a low-cost and efficient approach to developing strategies derived from simulated EV models. However, as highlighted in the introduction, deploying these strategies in real-world scenarios often results in performance discrepancies, commonly referred to as the sim-to-real gap. This gap arises due to differences between simulated and real-world conditions, including variations in environmental dynamics, noise, and other uncertainties. Combining ORL with online RL presents a promising solution to address this challenge. In Case 2, we aim to experimentally demonstrate that the proposed ORL method effectively reduces the gap, enhancing the performance of simulation-based EMS in real-world environments. Specifically, we examine an online RL-based EMS method implemented using the PPO algorithm. Initially trained on limited data from a simulated environment, this algorithm is then deployed in a new environment characterized by altered vehicle parameters and unknown driving conditions.

As shown in Fig. 7(a), during the simulation phase, the PPO algorithm is trained on the standardized driving cycle (WTVC) and three specific driving conditions (ZBDC) (Fig. 6(c)) to derive an ideal EMS, denoted as PPO (Train). Subsequently, the resulting EMS is validated across 12 different local driving conditions, denoted as PPO (Test). To simulate the environmental discrepancies between PPO (Train) and PPO (Test), the vehicle mass is varied—set to 4500 kg during training and increased to 5000 kg during testing—emphasizing the differences in operational scenarios. As depicted in Fig. 7(b), the cost difference between PPO (Train) and DP across the four training conditions is minimal, averaging 3.16% (Fig. 7(b)). However, when tested on the 12 new conditions (DC-1 to DC-12), the average cost difference between PPO (Test) and DP considerably rises to 12.75%. This indicates a significant performance degradation of DRL-based methods when transitioning from simulation to real-world conditions.

To mitigate the sim-to-real problem, our proposed ORL method leverages data from PPO (Test) for further learning. As illustrated in Fig. 7(c), the ORL approach achieves substantially lower costs across the 12 local operating conditions compared to the PPO (Train) strategy. The average cost difference between ORL and DP is reduced to just 1.42% (Fig. 7(b)). This notable cost reduction underscores ORL's ability to refine the initial EMS strategy derived from simulation-based methods and adapt it effectively to real-world conditions. In summary,

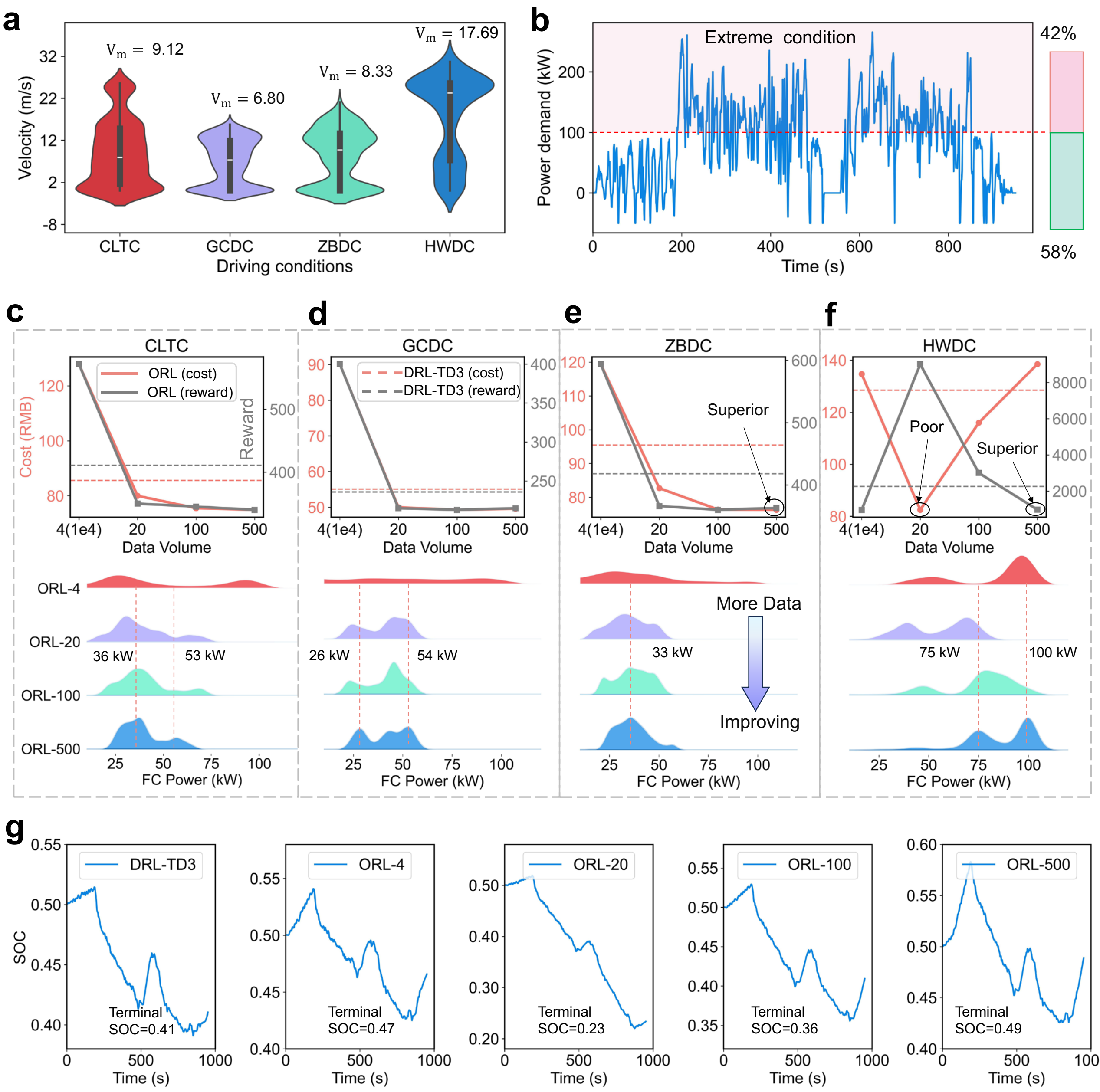

Fig. 8 | Performance with increased data. a Speed distribution of four conditions. **b** Demand power of HWDC, representing an extreme condition. The red dashed line indicates the maximum output power of the FC system. **c** Overall performance of the four cases as training data increases under the China Light-Duty Vehicle Test Cycle (CLTC). **d** Performance under the GCDC; **e** Performance under the ZNDC; **f** Performance under HWDC, indicating that the ORL agent can effectively learn a reasonable EMS even under extreme driving conditions. **g** Battery SOC trajectories for different EMS under the HWDC.

our experiments demonstrate that ORL can learn from simulated data to enhance the performance of the original EMS, providing a robust solution to the sim-to-real problem inherent in traditional simulation-based methods.

Case 3: learning a general EMS with large-scale data. To evaluate the generalization performance of the ORL model, the agent is trained on extensive data encompassing over 60 million kilometers and tested on four novel driving conditions. The training data, sourced from the augmented-reality EV platform, primarily comprises real-world driving conditions and a limited number of standard driving cycles, excluding the four test conditions reserved for validation. Four training datasets of varying scales are constructed, containing 4e4 (ORL-4), 20e4 (ORL-20), 100e4 (ORL-100), and 500e4 (ORL-500) samples. Figure 8(a) illustrates the speed distributions for these test conditions. Among these, the CLTC serves as the standard cycle, GCDC is based on data sourced from an EV operating in urban downtown areas, ZBDC represents a new condition recorded in Zhengzhou, and HWDC is derived from a fuel vehicle traveling on a highway. These conditions reflect diverse road, driver, and vehicle types, exhibiting significant variations in average speed V_m. Notably, under the HWDC condition, the FCEV experiences power demands exceeding 100 kW in over 42% of instances, with a peak demand surpassing 250 kW. As depicted in Fig. 8(b), this demonstrates that HWDC represents an extreme condition for the FCEV, as the power demand exceeds the 100 kW maximum output capacity of the FC system.

The results for the four validation conditions are depicted in Fig. 8(c–f). It is evident from Fig. 8(c), (d), and (e) that both the reward and cost exhibit a gradual decrease as the training data increases. With more training data, the ORL model consistently enhances its performance.

Notably, the rate of performance improvement diminishes after reaching the 100e4 sample mark, with minimal disparity observed between the ORL-100 and ORL-500. At approximately 20e4 samples, the ORL model outperforms the DRL-TD3 algorithm (as indicated by the dashed lines in the figures). Notably, under the extreme HWDC condition, while ORL-20 achieves the lowest cost (Fig. 8(f)), its reward absolute value is not the lowest. This phenomenon occurs because ORL-20 tends to prioritize battery consumption during periods of high power demand, exceeding the maximum output of the FC system. Consequently, the strategy fails to maintain the SOC within the desired range, rendering it ineffective as an EMS. Conversely, ORL-500 demonstrates superior performance, achieving both lower reward and cost while keeping the SOC within a reasonable range, as illustrated in Fig. 8(g). Overall, after training on five million data points (equivalent to over 60 million kilometers), the ORL agent successfully learns a general EMS that can adapt to unseen and even corner-case conditions.

This result highlights two key advantages of the ORL agent: First, its performance surpasses that of the original policy; second, it demonstrates that with increased data availability, learning performance improves. The ORL model effectively learns a general EMS from large-scale EV data, showcasing its adaptability and capability to enhance EMS performance as more data becomes available.

Discussion

In conclusion, we present a data-driven EMS for HESs in EVs. Our approach leverages an innovative ORL agent, which learns directly from EV data. To efficiently collect large-scale EMS datasets, we develop an augmented-reality EV platform that integrates real-world driving data from the EV monitoring and management system with a simulated FCEV powertrain model. Furthermore, we propose the AC-BPR, which incorporates BPR by combining BC with a discriminator. The BPR mechanism strikes a balance between conservatism and exploration, enabling AC-BPR to refine its policy even when learning from suboptimal or low-quality datasets. Experimental results demonstrate that the ORL agent not only learns optimal EMS strategies from expert data but also exhibits the ability to learn superior EMSs from datasets containing a mixture of expert and noisy data. The agent is also capable of achieving near-optimal strategies from entirely noisy datasets. Moreover, our approach demonstrates that, with increasing data availability, performance improves as the agent is trained with more data.

This approach offers three notable benefits. First, it is simple and data-driven, relying solely on collected data for automatic learning by the agent, unlike the traditional EMS development process, which often requires extensive expert knowledge and repeated measurements. Additionally, the data used in our approach are non-expert data that can be readily obtained from real vehicles. Second, our method ensures stable performance by seamlessly integrating with existing EMS, without altering the original baseline. Through data-driven enhancements, our approach continuously improves upon the baseline EMS, leveraging the strengths of both technologies. For example, to address the performance shortcomings in rule-based EMS, ORL enables incremental learning, allowing for continual enhancement of EMS performance using historical data. Similarly, ORL addresses the sim-to-real gap problem in simulation-based methods by refining pre trained EMS models, ensuring their effectiveness in real-world deployment scenarios. Finally, our approach demonstrates versatility: with sufficient data, it can learn a generalized EMS applicable across various EVs and operating conditions. This aligns with the current trends in artificial intelligence (AI) involving large-scale language models and similar approaches, where a single large model with large-scale data can be trained to perform well across diverse tasks and domains. Overall, we believe that ORL has the potential to serve as a foundational framework for data-driven EMS, with applications extending beyond EVs to include grid EMS, industrial energy management systems, and other vehicle control systems.

A limitation of this work is that the ORL agent requires significantly more data compared to traditional methods. In our experiments, over 60 million kilometers of EV driving data were utilized to develop a superior EMS. However, collecting such an extensive dataset from a single vehicle within a short timeframe is impractical. Leveraging data from large-scale vehicle fleets presents a more feasible solution. In China, for instance, a comprehensive EV monitoring and management system allows automakers to collect vast amounts of data via cloud-based platforms. This data can be utilized to enhance energy efficiency using the proposed ORL-based approach. However, current EV data standards, such as the GB/T 32960 protocol, are insufficient for this purpose and require more comprehensive and preprocessed datasets. Furthermore, ensuring the safety and reliability of AI-driven systems is paramount for real-world applications[40,41]. Integrating ORL with traditional EMS methods can provide a robust solution, where a baseline EMS ensures safety and a guaranteed minimum performance, while the ORL-based EMS optimizes energy efficiency using the available data. Further research is essential to develop more data-efficient algorithms and hybrid EMS frameworks that ensure safety, robustness, and adaptability for real-world large-scale vehicle applications.

Methods

EV powertrain model

In this work, we evaluate the EMS performance using the FCEV within a simulation environment. Figure 2(b) illustrates the schematic diagram of the FCEV and its components, which include the FC system, a hydrogen storage tank, an electric motor (EM), and a Lithium-ion battery (LIB) pack. The FC stack serves as the primary power source to meet the energy requirements of the vehicle. The diagram also depicts the energy flow from the hydrogen storage tank to the motor. The FC system converts hydrogen energy into electricity. This electricity then collaborates with the LIB via the high-voltage bus, powering a single electric motor, connected to the driving wheel via a fixed-ratio final gear. The main parameters of the FCEV model are listed in Table S4. According to the vehicle driving resistance equation, the driving power demand is determined by the speed and acceleration of the FCEV, and can be expressed as follows:

$$P_d = \frac{1}{3600 \cdot \eta_{\mathrm{me}}}\left(mgC_f v_t \cos(i) + m\delta v_t a_t + mgv_t \sin(\theta_s) + \frac{C_D A}{21.15} {v_t}^3\right) \tag{1}$$

where η_{me} is the efficiency of the vehicle drivetrain, m represents the vehicle mass, g denotes the gravitational constant, C_f is the rolling resistance coefficient, v_t indicates the longitudinal velocity at the time step t, a_t signifies the acceleration, δ refers to the rotational mass conversion coefficient, C_D represents the air resistance coefficient, A denotes the frontal area, and θ_s represents the angle of slope of the road. The power demand is provided by the FC system and the battery pack, with the power balance of the FCEV formulated as:

$$P_d = \left(P_{\mathrm{fc}} \cdot \eta_{\mathrm{DC/DC}} + P_{\mathrm{bat}}\right) \cdot \eta_{\mathrm{DC/AC}} \cdot \eta_{\mathrm{EM}} \tag{2}$$

where P_{fc} and P_{bat} respectively denote the output power of the FC system and the LIB pack; $\eta_{\mathrm{DC/DC}}$, $\eta_{\mathrm{DC/AC}}$, and η_{EM} represent the efficiency of the DC/DC converter, DC/AC inverter, and the electric motor, respectively. The battery pack is modeled using an equivalent circuit model, as shown in Equation (3):

$$\begin{cases} P_{\mathrm{bat}}(t) = V_{\mathrm{oc}}(t) - R_0 \cdot I^2(t) \\ I(t) = \frac{V_{\mathrm{oc}}(t) - \sqrt{V_{\mathrm{oc}}^2(t) - 4 \cdot R_0 \cdot P_{\mathrm{bat}}(t)}}{2R_0} \\ \mathrm{SOC}\,(t) = \frac{Q_0 - \int_0^t I(t)\,dt}{Q} \end{cases} \tag{3}$$

where SOC denotes the battery state of charge, V_{oc} is the open-circuit voltage, I_t represents the current at time t, R_0 indicates the internal resistance, P_{bat} refers to the output power in the charge-discharge cycles, Q_0 signifies the initial battery capacity, and Q is the nominal battery capacity.

According to the battery aging model in[16], the degradation rate of battery operation γ_{bat} is influenced by the charge/discharge rate (C_{rate}). The relationship between the battery aging correction factor and C_{rate} can be derived from experiment data:

$$\gamma_{bat} = \mu_1 |C_{rate}|^2 + \mu_2 |C_{rate}| + \mu_3 \tag{4}$$

where μ_1, μ_2, μ_3 are the curve-fitting coefficients. LIB can operate for about 5000 full cycles in a lifetime. The battery degradation cost $C_{bat,degr}$ can be calculated by:

$$C_{bat,\,degr} = \int_0^t \gamma_{bat}^{-1} P_{bat} dt \cdot PR_{bat} / (5000 \cdot 3600) \tag{5}$$

where PR_{bat} is the battery price per kWh that is 1500RMB/kWh.

The efficiency of the FC system under different power conditions is obtained from experiment data. Thus, the mass flow rate of the hydrogen consumption can be calculated by:

$$\dot{m}_{H_2} = P_{fcs} / \left(\eta_{fcs} \cdot \mathrm{LHV}_{H_2} \right) \tag{6}$$

where η_{fcs} is the FC system efficiency; P_{fcs} denotes the FC system output power; and LHV_{H_2} represents the hydrogen low calorific value. The FC hydrogen cost can be calculated by:

$$C_{fcs,H_2} = PR_{H_2} \cdot \int_0^t \dot{m}_{H_2} dt \tag{7}$$

where PR_{H_2} is the hydrogen price per kilogram(60RMB/kg).

The FC degrades rapidly under four typical conditions: load changing, start/stop, low power, and high power conditions. We assume that the FC system continues operating until the vehicle power system is shut down, thus the start/stop condition is not considered in the EMS. The degradation rate of the FC voltage, denoted as γ_{fcs}, can be calculated by:

$$\gamma_{fcs} = \kappa_{low} \cdot T_{low} + \kappa_{high} \cdot T_{high} + \kappa_{cha} \cdot \Delta P_{fcs} \tag{8}$$

where κ_{low} is the degradation rate under low power conditions; T_{low} denotes the duration of low power conditions; κ_{high} represents the degradation rate under high power conditions; T_{high} indicates the duration of high power conditions; κ_{cha} refers to the degradation rate under load-changing conditions; and ΔP_{fcs} is the FC power slope.

The FC is considered to reach the end of its life when it has lost 10% of voltage at rated power. The FC operation degradation cost can be calculated as:

$$C_{fcs,\,degr} = k_{fcs} \cdot \gamma_{fcs} \cdot P_{fcs,\,rate} \cdot PR_{fcs} / \left(V_{fcs,\,end} \cdot 1000 \right) \tag{9}$$

where k_{fcs} is the FC life correction factor; $V_{fcs,end}$ denotes the FC voltage drop at the end-of-life; $P_{fcs,rate}$ represents the rated power of the FC; and PR_{fcs} is the FC price per kilowatt(4000RMB/kW).

Problem modeling

In this work, the EMS of electric vehicles is modeled as a long-term sequential decision process objective to minimize total energy costs while maintaining battery SOC within reasonable limits. The optimization objective can be formulated as:

$$J_{EMS} = \min \sum_{t=0}^{T} \mathrm{cost}(t) + \alpha f_s(\mathrm{SOC}(t)) \tag{10}$$

where T is the total length of the driving cycle; cost(t) denotes the energy cost, including hydrogen consumption, battery costs, and FC degradation; f_s(SOC(t)) represents the SOC maintaining function; and α refers to the tradeoff between energy cost and SOC.

To tackle the sequential decision, the energy management system is formulated as an MDP, which provides a framework for learning the optimal EMS through interaction in order to minimize total energy costs. The MDP is defined by a tuple $(S, A, P, R, \rho_0, \gamma)$, where $\mathcal{S}$ denotes the state space, A represents the action space, $P(s'|s, a)$ is the transition distribution, $\rho_0(s)$ indicates the initial state distribution, $R(s, a)$ refers to the reward function, and $\gamma \in (0, 1)$ denotes the discount factor. The goal is to identify a policy $\pi(a|s)$ that maximizes the expected cumulative discounted rewards $J(\pi) = E_{\pi,P,\rho_0}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$. For the FCEV, the state space at time point t is defined as:

$$S = \left\{ v_t, a_t, \mathrm{SOC}_t, P_{fcs}^t \right\} \tag{11}$$

where v_t, a_t, P_{fcs}, SOC_t are the vehicle speed, acceleration, FC power, and battery SOC, respectively. The action represents the control variable, which involves allocating power to the energy sources of the vehicle. In the context of the FCEV, the action is defined as the FC power slope, denoted as ΔP_{fcs}. The continuous action can be described as follows:

$$A = \left\{ \Delta P_{fcs} = P_{fcs}^t - P_{fcs}^{t-1}, \Delta P_{fcs} \in [-10, \mathrm{kW}, 10, \mathrm{kW}] \right\} \tag{12}$$

The reward function R represents the reward $R(s_{t+1}; s_t; a_t)$ associated with transitioning from state s_t to state s_{t+1} using action a_t. The design of the reward function is crucial to the learning process. For the FCEV, multiple objectives are considered, such as hydrogen consumption, FC degradation, and battery-related costs (including electricity consumption and degradation). Additionally, it is essential to maintain the battery SOC. Therefore, the reward function is defined as the sum of energy costs, while ensuring that the battery charge-sustaining constraints are satisfied:

$$R = -\left\{ C_{fcs,H_2} + C_{fcs,\,degr} + C_{bat,eH_2} + C_{bat,\,degr} + \alpha \left[\mathrm{SOC}_{ref} - \mathrm{SOC}(t) \right]^2 \right\} \tag{13}$$

The battery electricity consumption C_{bat,eH_2} is calculated according to the battery charge/discharge efficiency and converted into price cost:

$$C_{bat,eH_2} = \int_0^t \left[P_{bat} / \left(\eta_{d/c} \eta_{DC/DC}, \mathrm{LHV}_{H_2} \right) \right] dt \cdot PR_{H_2} \tag{14}$$

where $\eta_{d/c}$ is the battery discharge/charge efficiency.

ORL algorithm

RL is a paradigm for learning optimal policies in a sequential decision-making problem. It involves an agent interacting with an environment, taking actions, and receiving feedback in the form of rewards. In this study, we apply the RL paradigm to address the MDP problem described earlier. The objective is to learn a policy $\pi \sim a_t$ (π: $S \rightarrow A$) that maximizes the expected sum of discounted rewards $J(\pi)$. Each policy π has a corresponding state-action value function (also known as the Q function), which denotes the expected return $Q(s, a)$ when following the policy π after taking an action a in state s.

$$Q(s,a)=\mathbb{E}\left[\sum_{i=t}^{\infty}\gamma^{i-t}R_i|s_t=s,a_t=a\right] \tag{15}$$

where $\mathbb{E}()$ denotes the mathematical expectation. Online RL is an interactive ML paradigm where the agent learns from continuous interactions with the environment. This allows the agent to gradually improve its policy over time. This approach benefits from exploration, where the agent attempts different actions to discover potentially superior policies. However, the main challenge in interactive learning is the need to recollect the dataset every time the policy changes, which can be costly and impractical in real-world scenarios[34]. To address this, online RL often relies on simulated training to avoid the expenses and risks associated with real-world interactions.

ORL is a data-driven extension of traditional RL, leveraging pre-existing datasets to refine policy training without requiring further interaction with the environment. This paradigm is particularly useful in scenarios where real-world interactions are costly, risky, or infeasible. By utilizing datasets collected under a behavioral policy (π_β), ORL aims to derive an optimized decision policy (π_{off}) while avoiding direct exploration of the environment. Given a static dataset of transitions $D=\left\{(s_t,a_t,s_{t+1},r_t)_i\right\}$ where i indexes a transition, the actions are sampled from the behavior policy $a_t\sim\pi_\beta(\cdot|s_t)$, the states are drawn from a distribution induced by the behavior policy $s_t\sim d^{\pi_\beta}(\cdot)$, the next state is determined by the transition dynamics $s_{t+1}\sim T(\cdot|s_t,a_t)$, and the reward is a function of state and action $r_t=r(s_t,a_t)$. The objective of ORL remains to identify a policy π_{off} that maximizes the expected return.

However, a critical challenge in ORL is the state-action distribution shift, where the learned policy π_{off} encounters states or actions that are not adequately represented in the dataset[42]. In other words, the ORL agent may face unfamiliar state-action regimes that were not covered by the offline EMS dataset, leading to inaccurate Q-value estimation and reduced policy performance. Additionally, offline EMS datasets are often diverse or of a poor-quality, as they contain suboptimal actions or noisy data due to imperfections in the behavior policy π_β or inconsistencies during data collection. This further complicates the learning process and makes identifying the optimal policy challenging. To address these issues and ensure sample-efficient learning in ORL, the AC-BPR algorithm is proposed. This algorithm introduces BPR, which integrates BC and a discriminator to balance conservatism with exploratory learning.

BC aims to ensure the learned policy remains aligned with the behavior policy that generated the offline dataset. By minimizing the divergence between the learned policy and the demonstrated actions, BC mitigates distribution shifts[37]. The BC objective is defined as:

$$\min_{\pi}\mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}}[-\log\pi(a|s)] \tag{16}$$

where $\pi(a|s)$ represents the likelihood of selecting action a in state s under the policy π. This objective encourages the learned policy to mimic the expert behavior encoded in the dataset. However, in scenarios with non-optimal data, BC alone may lead to overfitting to undesirable actions, limiting policy performance. To address this limitation, BPR introduces a discriminator. The discriminator is trained to distinguish whether a given state-action pair (s, a) belongs to the original dataset $\mathcal{D}$, or if it was generated by the learned policy π_θ. In this setup, the policy π_θ acts as the generator in a Generative Adversarial Network (GAN)[43] setup. The discriminator helps guide the policy by encouraging exploration of actions that might not be represented in the dataset but are still plausible according to its judgment[38,44].

$$\min_{D}\mathbb{E}[\log D(s_t,a_t)]+\mathbb{E}[\log(1-D(s_t,\pi(s_t)))] \tag{17}$$

$D(s_t,a_t)$ evaluates the probability that (s_t,a_t) originates from the offline dataset, and $\pi(s_t)$ denotes the action suggested by the learned policy. The trained discriminator $D(s,\pi(s))$ is integrated into the policy objective as a reward signal. This incentivizes the learned policy to explore diverse, high-quality actions while discouraging over-reliance on suboptimal dataset behaviors.

The BPR mechanism combines the strengths of BC and DR within the Actor-Critic framework. AC-BPR can be seamlessly integrated into any Actor-Critic algorithm, and in this study, we implement it using the TD3 framework. In this setup, the Actor-network is responsible for selecting actions based on the current policy $\pi(s)$, which is influenced by expert behavior through BC and exploration of high Q-value regions via DR. The Critic network estimates the Q-values for state-action pairs, providing feedback to refine the policy. The final optimization objective of AC-BPR balances Q-value maximization, BC regularization, and discriminator-based exploration. This balance between conservatism and exploration allows AC-BPR to learn effectively even with suboptimal datasets. The Actor network is updated by optimizing the BPR objective, which combines these components to guide the policy toward more effective and diverse actions, improving learning efficiency in offline settings.

$$\pi=\arg\max_{\pi}\mathbb{E}_{(s,a)\sim D}\left[\lambda Q(s,\pi(s))-(1-\beta)(\pi(s)-a)^2+\beta\log(D(s,\pi(s)))\right] \tag{18}$$

where $Q(s,\pi(s))$ represents the expected return for the policy action $\pi(s)$ in state s. The term $(\pi(s)-a)^2$ penalizes deviations from the dataset actions to ensure conservatism. $\log(D(s,\pi(s)))$ rewards exploration in high-reward regions identified by the discriminator. The parameter β (range of 0 to 1) adjusts the balance between BC and DR constraints. The λ is a normalization term based on the average absolute value of Q to control the balance between RL and imitation, defined as:

$$\lambda=\frac{\alpha}{\frac{1}{N}\sum_{(s_i,a_i)}|Q(s_i,a_i)|} \tag{19}$$

The parameter α is used to control the strength of the regularization, with a larger value of α causing the algorithm to lean more toward more RL. The parameter N represents the number of transitions in the dataset and is used to normalize the characteristics of each state within the provided dataset. Let s_i be the i-th feature of the state s in the dataset, with μ_i, σ_i being the mean and standard deviation (η is a constant value to avoid division by zero.):

$$s_i=\frac{s_i-\mu_i}{\sigma_i+\eta} \tag{20}$$

The Critic network is responsible for estimating the Q-values for state-action pairs, providing feedback to the Actor network during training. It plays a crucial role in evaluating the Actor's actions and guiding policy improvement. To address the issue of overestimation bias in Q-values, AC-BPR utilizes dual Critic networks, a strategy analogous to that employed in the TD3 algorithm. Each Critic network, $Q_1(s,a|\theta^{Q_1})$ and $Q_2(s,a|\theta^{Q_2})$, corresponds to a target network $Q_1'(s,a|\theta^{Q_1'})$ and $Q_2'(s,a|\theta^{Q_2'})$, respectively. The minimum Q-value among the two Critics is used as the target Q-value during training. The Critic network is updated by minimizing the loss function:

$$L(\theta^{Q_i})=\mathbb{E}\left[\left(y_t-Q_i(s_t,a_t|\theta^{Q_i})|a_t=\mu(s_t|\theta^{\mu})\right)^2\right] \tag{21}$$

where θ^Q denotes the weights of the Critic network. The target Q-value y is evaluated by taking the minimum of the estimates from the two Q-functions:

$$y_t = r(s_t, a_t) + \gamma \min_{i=1,2} Q'_i\left(s_{t+1}, a_{t+1} | \theta^{Q'_i}\right) \tag{22}$$

where $a_{t+1} \sim \pi_{\phi'}(s_{t+1}) + \epsilon, \quad \epsilon \sim \mathrm{clip}(\mathcal{N}(0, \tilde{\sigma}), -c, c)$ is the exploration noise. This noise helps smooth the value estimates and improves the robustness of the learned Q functions. The term r represents the instantaneous one-step reward, with γ denoting the discounting factor.

Baseline methods

We use a series of baseline EMS methods for comparatively evaluating the ORL method. The inputs and outputs of all baselines are the same as those of the proposed method.

Dynamic Programming (DP)[45]: In optimization control methods, the EMS problem is formulated as a nonlinearly constrained optimization problem, aiming to minimize the objective function presented in Equation (10). DP is an optimization control method that operates by seeking the shortest path backward in time. Its objective is to derive the minimum cost function for each grid at every stage in reverse chronological order. In our study, DP is used as the benchmark EMS, representing the global optimum and providing upper limits for comparison. It's important to recognize that DP requires future information as input to achieve the optimization objective.

Behavior Cloning (BC)[46]: BC, as a fundamental imitation learning approach, seeks to emulate the EMS by directly learning from the provided dataset, which is assumed to be generated by an expert policy or near-expert policy. It employs supervised learning techniques to train a model to map states to actions. Both BC and ORL involve learning from data for EMS applications. In this context, we establish BC as the benchmark and aim to showcase the superior performance of ORL.

Proximal Policy Optimization (PPO)[47]: PPO is a state-of-the-art online DRL algorithm, which has been extensively applied in various applications requiring sophisticated decision-making in dynamic environments. PPO offers a robust and efficient approach to training agents by leveraging on-policy learning, effective use of data through mini-batch updates, stability through policy clipping, and adaptive learning rates. Leveraging the strengths of PPO, we utilize it to generate the dataset necessary for ORL, with its policy serving as an expert (near-optimal) strategy for comparison purposes. We provide it to explore the superiority of ORL compared to the online DRL.

Twin Delayed Deep Deterministic Policy Gradient (TD3)[39]: TD3 is an advanced online DRL algorithm, stemming from the Actor-Critic framework. It has garnered significant attention due to its effectiveness in overcoming challenges associated with continuous action spaces and high-dimensional state spaces. TD3 employs twin critic networks to estimate the value of actions more accurately. By utilizing two critic networks, TD3 mitigates overestimation bias and enhances the robustness of value function estimation. We also provide it to explore the superiority of ORL compared to the online DRL.

Data availability

All data generated in this study are provided in the Supplementary Information/Source Data file. The raw datasets used for modeling the driving conditions are sourced from the EV monitoring and management system in China. Source data are provided in this paper. Source data are provided with this paper.

Code availability

The code for data analysis can be obtained from the Source Data file. The DRL algorithms and EV powertrain model used in this study, LearningEMS[48], are publicly available at https://doi.org/10.5281/zenodo.14848553, which links to the GitHub repository: https://github.com/wangjail/LearningEMS. All other codes used in this study are available from the corresponding authors upon request.

References

1. Borlaug, B. et al. Heavy-duty truck electrification and the impacts of depot charging on electricity distribution systems. *Nat. Energy* **6**, 673–682 (2021).
2. Zhao, Y., Wang, Z., Shen, Z.-J. M. & Sun, F. Assessment of battery utilization and energy consumption in the large-scale development of urban electric vehicles. *Proc. Natl. Acad. Sci.* **118**, e2017318118 (2021).
3. He, H. et al. Deep reinforcement learning based energy management strategies for electrified vehicles: recent advances and perspectives. *Renew. Sustain. Energy Rev.* **192**, 114248 (2024).
4. Liu, R. et al. A cross-scale framework for evaluating flexibility values of battery and fuel cell electric vehicles. *Nat. Commun.* **15**, 280 (2024).
5. Li, Y. et al. Deep reinforcement learning for intelligent energy management systems of hybrid-electric powertrains: Recent advances, open issues, and prospects. *IEEE Transactions on Transportation Electrification* (2024).
6. Han, Q. et al. Hierarchical coordinated optimization and energy management control for plug-in hybrid electric heavy-duty truck platoon in coal mine transportation system. *IEEE Transactions on Vehicular Technology* (2024).
7. Ganesh, A. H. & Xu, B. A review of reinforcement learning based energy management systems for electrified powertrains: progress, challenge, and potential solution. *Renew. Sustain. Energy Rev.* **154**, 111833 (2022).
8. Zhang, F. et al. Comparative study of energy management in parallel hybrid electric vehicles considering battery ageing. *Energy* **264**, 123219 (2023).
9. Wu, J., Huang, C., He, H. & Huang, H. Confidence-aware reinforcement learning for energy management of electrified vehicles. *Renew. Sustain. Energy Rev.* **191**, 114154 (2024).
10. Wang, J., Kang, L. & Liu, Y. Optimal scheduling for electric bus fleets based on dynamic programming approach by considering battery capacity fade. *Renew. Sustain. Energy Rev.* **130**, 109978 (2020).
11. Quan, S. et al. Customized energy management for fuel cell electric vehicle based on deep reinforcement learning-model predictive control self-regulation framework. *IEEE Transactions on Industrial Informatics* (2024).
12. Tang, X. et al. Naturalistic data-driven predictive energy management for plug-in hybrid electric vehicles. *IEEE Trans. Transport. Electrif.* **7**, 497–508 (2020).
13. Wang, Y., Wu, Y., Tang, Y., Li, Q. & He, H. Cooperative energy management and eco-driving of plug-in hybrid electric vehicle via multi-agent reinforcement learning. *Appl. Energy* **332**, 120563 (2023).
14. Hu, X., Liu, T., Qi, X. & Barth, M. Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: Recent advances and prospects. *IEEE Ind. Electron. Mag.* **13**, 16–25 (2019).
15. Wang, Y., Tan, H., Wu, Y. & Peng, J. Hybrid electric vehicle energy management with computer vision and deep reinforcement learning. *IEEE Trans. Ind. Inform.* **17**, 3857–3868 (2020).
16. Chen, W. et al. Health-considered energy management strategy for fuel cell hybrid electric vehicle based on improved soft actor critic algorithm adopted with beta policy. *Energy Convers. Manag.* **292**, 117362 (2023).
17. Yuankai, W., Renzong, L., Yong, W. & Yi, L. Benchmarking deep reinforcement learning based energy management systems for hybrid electric vehicles. In *CAAI International Conference on Artificial Intelligence*, 613–625 (Springer, 2022).
18. Wang, Y. et al. LearningEMS: A unified framework and open-source benchmark for learning-based energy management of electric

vehicles. *Engineering*. https://www.sciencedirect.com/science/article/pii/S2095809924007136 (2024).
19. Li, J. et al. When data geometry meets deep function: Generalizing offline reinforcement learning. In *The Eleventh International Conference on Learning Representations* (2023).
20. Wang, Y., Lian, R., He, H., Betz, J. & Wei, H. Auto-tuning dynamics parameters of intelligent electric vehicles via Bayesian optimization. *IEEE Transactions on Transportation Electrification* (2023).
21. Pozzato, G. et al. Analysis and key findings from real-world electric vehicle field data. *Joule* **7**, 2035–2053 (2023).
22. Deng, F. et al. A big data approach to improving the vehicle emission inventory in China. *Nat. Commun.* **11**, 2801 (2020).
23. Peng, J. et al. Enhancing lithium-ion battery monitoring: a critical review of diverse sensing approaches. *eTransportation* **22**, 100360 (2024).
24. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
25. Li, W., Zhu, J., Xia, Y., Gorji, M. B. & Wierzbicki, T. Data-driven safety envelope of lithium-ion batteries for electric vehicles. *Joule* **3**, 2703–2715 (2019).
26. Zhang, J. et al. Realistic fault detection of li-ion battery via dynamical deep learning. *Nat. Commun.* **14**, 5940 (2023).
27. Roman, D., Saxena, S., Robu, V., Pecht, M. & Flynn, D. Machine learning pipeline for battery state-of-health estimation. *Nat. Mach. Intell.* **3**, 447–456 (2021).
28. Lu, J., Xiong, R., Tian, J., Wang, C. & Sun, F. Deep learning to estimate lithium-ion battery state of health without additional degradation experiments. *Nat. Commun.* **14**, 2760 (2023).
29. Li, Q. et al. A hybrid physics-data driven approach for vehicle dynamics state estimation. *Mech. Syst. Signal Process.* **225**, 112249 (2025).
30. Munoz, P. M. et al. Energy management control design for fuel cell hybrid electric vehicles using neural networks. *Int. J. Hydrog. Energy* **42**, 28932–28944 (2017).
31. Millo, F., Rolando, L., Tresca, L. & Pulvirenti, L. Development of a neural network-based energy management system for a plug-in hybrid electric vehicle. *Transport. Eng.* **11**, 100156 (2023).
32. Liu, B., Wei, X., Sun, C., Wang, B. & Huo, W. A controllable neural network-based method for optimal energy management of fuel cell hybrid electric vehicles. *Int. J. Hydrog. Energy* **55**, 1371–1382 (2024).
33. Rajaraman, N., Yang, L., Jiao, J. & Ramchandran, K. Toward the fundamental limits of imitation learning. *Adv. Neural Inf. Process. Syst.* **33**, 2914–2924 (2020).
34. Prudencio, R. F., Maximo, M. R. & Colombini, E. L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
35. He, H., Niu, Z., Wang, Y., Huang, R. & Shou, Y. Energy management optimization for connected hybrid electric vehicle using offline reinforcement learning. *J. Energy Storage* **72**, 108517 (2023).
36. Hu, B., Liu, B. & Zhang, S. A data-driven reinforcement learning based energy management strategy via bridging offline initialization and online fine-tuning for a hybrid electric vehicle. *IEEE Trans. Ind. Electron.* **71**, 12869–12878 (2024).
37. Fujimoto, S. & Gu, S. S. A minimalist approach to offline reinforcement learning. *Adv. neural Inf. Process. Syst.* **34**, 20132–20145 (2021).
38. Xu, H., Zhan, X., Yin, H. & Qin, H. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *International Conference on Machine Learning*, 24725–24742 (PMLR, 2022).
39. Fujimoto, S., Hoof, H. & Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, 1587–1596 (PMLR, 2018).
40. Feng, S. et al. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* **615**, 620–627 (2023).
41. Cao, Z. et al. Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning. *Nat. Mach. Intell.* **5**, 145–158 (2023).
42. Ran, Y., Li, Y.-C., Zhang, F., Zhang, Z. & Yu, Y. Policy regularization with dataset constraint for offline reinforcement learning. In *International Conference on Machine Learning*, 28701–28717 (PMLR, 2023).
43. Creswell, A. et al. Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* **35**, 53–65 (2018).
44. Kidera, S., Shintani, K., Tsuneda, T. & Yamane, S. Combined constraint on behavior cloning and discriminator in offline reinforcement learning. *IEEE Access* (2024).
45. Saiteja, P. & Ashok, B. Critical review on structural architecture, energy control strategies and development process towards optimal energy management in hybrid vehicles. *Renew. Sustain. Energy Rev.* **157**, 112038 (2022).
46. Block, A., Jadbabaie, A., Pfrommer, D., Simchowitz, M. & Tedrake, R. Provable guarantees for generative behavior cloning: Bridging low-level stability and high-level behavior. *Advances in Neural Information Processing Systems* **36** (2024).
47. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
48. Wang, Y., Wu, J., He, H., Wei, Z. & Sun, F. Learningems: Code for 'data-driven energy management for electric vehicles' https://doi.org/10.5281/zenodo.14848553 (2025).
49. He, H. et al. China's battery electric vehicles lead the world: achievements in technology system architecture and technological breakthroughs. *Green. Energy Intell. Transport.* **1**, 100020 (2022).

Acknowledgements

H.H. was supported by National Natural Science Foundation of China (Grant No. 52172377) and Beijing-Tianjin-Hebei Basic Cooperation Special Project (Grant No. F2021203118).

Author contributions

Y.W. designed the study and methodology; Y.W., J.Wu., and H.H. collected and analyzed data; Y.W. generated the figures; Y.W., J.Wu., and W.Z. wrote the manuscript; H.H., W.Z., and F.S. reviewed and edited the manuscript. H.H. and F.S. planned and supervised the project. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58192-9.

Correspondence and requests for materials should be addressed to Hongwen He.

Peer review information *Nature Communications* thanks Emanuele de Santis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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