



ARTICLE

Scalable Deep Reinforcement Learning for Multi-Home Battery Dispatch Optimization in Smart Grids

Abdullah Alshammari^{1,*}, Ammar Ahmed E. Elhadi¹ and Ashraf Osman Ibrahim²

¹Department of Computer Science and Engineering, College of Computer Science and Engineering, University of Hafr Al Batin, Hafar Al-Batin, Saudi Arabia

²Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

*Corresponding Author: Abdullah Alshammari. Email: dr.abdullah@uhb.edu.sa

Received: 08 April 2026; Accepted: 22 June 2026; Published: 21 October 2026

ABSTRACT: Rising penetration of distributed renewables and residential battery storage is reshaping demand-side management in smart grids, yet coordinating many household batteries to reduce grid dependence remains an open challenge. Conventional rule-based and model-based controllers scale poorly and require accurate forecasts, motivating data-driven control. This study evaluates the capacity of Deep Reinforcement Learning (DRL) to reduce grid electricity consumption across a network of 200 smart homes using a realistic, synthetic, year-long, daily demand dataset. We develop a custom OpenAI Gymnasium environment that models household-level battery dispatch, assuming a 5 kWh battery per home with daily renewable recharge. Four widely used DRL algorithms from Stable-Baselines3, Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), Soft Actor-Critic (SAC), and Twin-Delayed DDPG (TD3) are trained under identical conditions to minimize total annual grid draw. Results show substantial savings relative to a no-control baseline, with A2C achieving 48.2% annual energy savings, TD3 44.3%, SAC 43.8%, and PPO 0.2%. Analysis of learning dynamics, savings distributions, and sample operational periods suggests that policy-gradient and off-policy continuous-control methods can learn effective battery scheduling policies in high-dimensional action spaces, whereas PPO requires additional tuning to achieve comparable performance. The work contributes a reproducible DRL benchmarking framework for realistic energy datasets and surfaces practical design considerations, normalization, rollout length, and model capacity that materially affect convergence and performance in multi-home energy management.

KEYWORDS: Smart home energy management; deep reinforcement learning; proximal policy optimization; advantage actor-critic; soft actor-critic; twin-delayed DDPG; battery dispatch optimization

1 Introduction

A contemporary smart grid must be capable of accommodating large volumes of distributed energy resources (DER) and battery storage devices in a reliable, cost-effective manner. The combination of DER and storage devices has the potential to decrease the cost of generation, balance generation and loss, and improve reliability in service provision [1]. However, this also introduces time-varying and uncertain dynamics that must be carefully modeled and simulated [2,3]. The literature on smart grids highlights several technical challenges in integrating dispersed renewable resources and energy storage, including interoperability between multiple nodes, handling large amounts of data, navigating socio-economic regulations, and meeting capital investment requirements [4,5]. Effective storage solutions and common standards are crucial for avoiding load shedding and ensuring grid stability [6]. Additionally, distributed batteries can help mitigate the

intermittency of renewable generation by charging during periods of low demand and feeding back during peak times [1,3]. However, recent review articles have pointed out that the real two-time control and dispatch of these assets remain emerging research challenges [7,8].

One potential architecture currently in development is the multi-home battery dispatch system. This system involves equipping each home in a community with a small battery and potentially rooftop photovoltaics. When properly coordinated, these batteries can be utilized to reduce peak grid demand, increase the use of renewable energy, and facilitate peer-to-peer energy trading. However, there is limited data available on the effectiveness of this system. The Energy Consumption Dataset for Smart Homes provides synthetic daily consumption and production profiles for 20, 50, 100, and 200 homes, including a simulated dataset of 50 homes in northern Saudi Arabia [9]. This dataset allows researchers to model multi-home dispatch and analyze how individual batteries contribute to overall aggregate behavior [9].

Most deployed energy-management systems use rule-based or heuristic strategies due to their low computational demands and ease of interpretation [10,11]. However, these strategies require designers to predetermine rules for all possible situations and are unable to adapt to changing circumstances [12]. While rule-based approaches may be transparent, they often struggle to balance accuracy and computational complexity, resulting in poor performance. Recent surveys on smart-grid algorithms also highlight the need for accurate information on costs, prices, and renewable outputs in model-based optimization and deterministic scheduling [13]. Stochastic programming can handle uncertainty, but it becomes difficult to scale when dealing with a large number of homes and daily decisions [14,15]. As a result, researchers have turned to model-free methods [16].

DRL has garnered significant attention in the field of energy management due to its ability to learn optimal control policies through interactions with the environment, operate in high-dimensional state spaces and uncertain environments, and generalize to novel conditions [17,18]. In contrast, traditional rule-based or proportional–integral–derivative (PID) controllers used in building energy management systems are static, struggle with multi-objective trade-offs, and cannot be scaled. However, DRL has the potential to learn and optimize multiple objectives in real-time and generalize across buildings [19,20]. For instance, algorithms based on reinforcement learning (RL) have been proposed for scheduling battery energy storage in energy markets, and have been shown to learn charging/discharging policies without requiring accurate environmental models, making them robust to uncertainties [21]. In another study, a battery model and an electricity price forecasting model are integrated into an RL-enhanced scheduling framework, where a profit-maximizing agent is trained through reinforcement learning to consider physical constraints [22]. Similarly, a study using soft actor-critic (SAC) for community battery dispatch in the presence of uncertain solar generation and energy prices found that SAC outperformed proximal policy optimization (PPO) and deep deterministic policy gradient (DDPG) on a community battery scheduling problem [23]. Furthermore, DRL control of co-located batteries has been shown to achieve 61.96% of the theoretical optimal operation in grid-scale applications and is more effective at shifting solar energy to high-demand times compared to rule-based control [24]. Additionally, DRL dispatch introduces diversity in battery actions, minimizing ramping problems caused by synchronized behavior [25].

Although DRL has shown promise, the majority of studies focus on single buildings or small microgrids, often using simulated data or simplified environments [26]. Few works examine multi-home battery scheduling at the community scale or compare multiple DRL algorithms on the same dataset. To address this gap, our study makes four contributions. Our first contribution is optimizing multi-home battery dispatch in smart grids. We use a synthetic dataset of daily energy consumption and production from 200 homes, with a variety of household loads and renewable contributions [9]. Managing a large number of distributed batteries is crucial in reducing peak grid demand and energy costs. To simulate this problem, we design a

custom OpenAI Gym environment for multi-home battery dispatch. Each episode represents one year of daily decisions, and the state space includes per-home energy demand and a time index. The environment allows for continuous actions to indicate the proportion of battery capacity used in each home, and rewards are based on the amount of grid energy consumed.

We train and compare four representative DRL algorithms, PPO, A2C, SAC, and TD3, using the same environment and training budget. Recent literature indicates that SAC often outperforms other RL methods in battery scheduling, while A2C and PPO have shown promise in other energy-management contexts [23]. Our results reveal how these methods perform in a high-dimensional multi-home setting [26]. This paper is a proposal of a scalable DRL framework to optimize battery dispatch in a smart-grid environment with 200 smart homes. We model a community using a real-world energy consumption dataset, with each home having a 5 kWh battery and possibly local renewable generation. An environment based on OpenAI Gymnasium simulates day-to-day household demand, and the DRL agents PPO, A2C, SAC, and TD3 can learn charging and discharging policies that minimize the energy drawn on the grid. The methodology models the dynamic interaction of several independent batteries with the grid and allows coordinated actions to decrease peak demand, increase self-consumption of renewable energy, and maximize energy efficiency, as shown in Fig. 1.

In the remainder of this paper, Section 2 reviews existing research on smart grid energy management and DRL applications in distributed battery scheduling. Section 3 details the proposed multi-home battery dispatch environment, dataset, and DRL training setup. Section 4 presents a comparative evaluation of PPO, A2C, SAC, and TD3 on real-world consumption data. Section 5 interprets the results and highlights practical implications for smart grid integration. Finally, Section 6 summarizes the key findings and outlines directions for future research.

2 Related Work

The integration of distributed energy storage systems (DESS) into smart grids has been widely recognized as a promising approach to enhance grid flexibility, support renewable integration, and reduce operational costs [27,28]. Battery energy storage systems (BESS) can perform peak shaving, load shifting, and frequency regulation, thus alleviating grid congestion and improving power quality [29,30]. While early works primarily relied on deterministic or rule-based scheduling strategies [31,32], these approaches often fail to adapt to stochastic demand profiles and renewable generation uncertainties.

Model-based optimization techniques, such as mixed-integer linear programming (MILP) and dynamic programming (DP), have been widely applied to battery dispatch problems [33]. However, their computational complexity grows rapidly with problem size and time horizon, making them unsuitable for large-scale residential deployments. Furthermore, such approaches typically require precise forecasts of load and generation, which may be unreliable in synthetic contexts [34]. As a result, heuristic and metaheuristic methods, including genetic algorithms [35] and particle swarm optimization [36], have been explored, yet these methods can struggle with convergence and scalability.

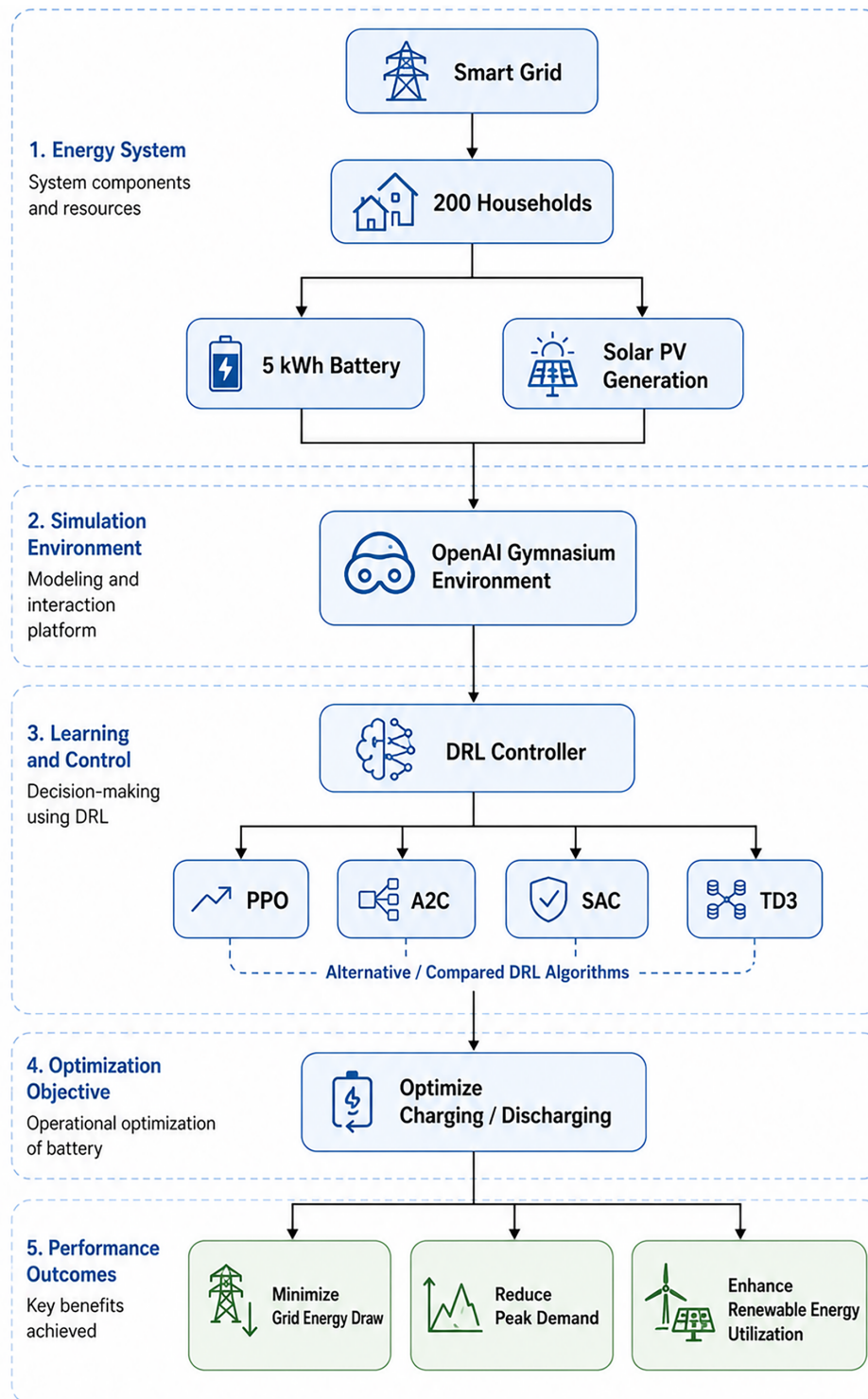


Figure 1: Scalable DRL framework for optimizing battery dispatch across 200 smart homes.

In recent years, deep reinforcement learning (DRL) has emerged as a powerful alternative for sequential decision-making in energy systems [37,38]. DRL combines reinforcement learning's adaptability with

deep learning's representational capacity, enabling agents to learn optimal control strategies from high-dimensional state–action spaces without explicit modelling of the system dynamics [39]. Notably, DRL has been applied to building energy management [10,12], microgrid scheduling [13,40], and electric vehicle charging coordination [41].

Among DRL algorithms, policy gradient methods such as Proximal Policy optimization (PPO) [42], Advantage Actor–Critic (A2C) [43], Soft Actor–Critic (SAC) [44], and Twin Delayed Deep Deterministic Policy Gradient (TD3) [45] have shown strong performance in continuous control tasks relevant to energy systems. PPO has been used in residential demand response scenarios with promising convergence stability [46], while A2C has demonstrated superior learning efficiency in stochastic environments [47]. Off-policy algorithms such as SAC and TD3 can leverage replay buffers for improved sample efficiency, making them attractive for scenarios with limited interaction opportunities [48].

Recent studies have explored multi-home and community-level battery optimization using DRL. For instance, study [49] proposed a multi-agent DRL framework for coordinated charging/discharging in community microgrids, while study [50] combined DRL with model predictive control (MPC) to improve operational stability.

More recent multi-agent DRL (MADRL) frameworks explicitly target the coordination and scalability issues of many-home settings: study [51] propose a centralized-training/decentralized-execution MADDPG variant that shares only strategic information to address single-point failures, communication overhead, and privacy in smart communities, while study [52] uses federated RL across multiple smart homes with distributed energy resources, and study [53] demonstrates scalable privacy-preserving MADRL for large-scale peer-to-peer energy trading. Relative to these, our study contributes a controlled single-environment benchmark of four actor–critic algorithms; we position a decentralized MADRL reformulation as the natural extension (Section 5).

integrated DRL with model predictive control for improved operational stability. However, most of these studies rely on simulated datasets or simplified load profiles, limiting their applicability to real-world scenarios. The use of large-scale, realistic synthetic residential datasets, as in this work, remains comparatively underexplored and poses additional challenges in handling variability and demand diversity [54].

In this study, we build upon these advances by developing a scalable DRL benchmarking framework using synthetic data from 200 smart homes. Our work differs from prior studies in three key aspects: (i) the use of a large-scale, synthetic residential dataset; (ii) the implementation of a custom, high-dimensional Gymnasium environment to capture the complexity of multi-home battery scheduling; and (iii) a comparative evaluation of four state-of-the-art DRL algorithms under identical training and evaluation conditions, providing reproducible baselines for future research.

2.1 Related Formulations and Our Model

Rule-based and heuristic schemes remain common in residential energy management due to simplicity, but they lack adaptability and provide no optimality guarantees under uncertainty [38]. Optimization and model predictive control (MPC) formulations explicitly minimize energy/cost subject to battery dynamics (SOC) and power constraints, but they require accurate forecasts and scale poorly for large communities and long horizons [12,13]. DRL methods learn policies directly from interaction, handling high-dimensional states and uncertainties; actor–critic families (A2C, SAC, TD3, PPO) are widely adopted for continuous control in energy systems [37,42–45].

We cast multi-home battery dispatch as an episodic MDP over a one-year horizon ($T = 365$ days) with $N = 200$ homes. The *state* at day t is

$$s_t = [d_{1,t}, \dots, d_{N,t}, \tau_t] \in \mathbb{R}^{N+1}, \quad (1)$$

where $d_{i,t} \geq 0$ is home- i demand (Wh) and $\tau_t \in [0, 1]$ is the normalized day index.

The *action* is a continuous discharge fraction per home,

$$a_t = [a_{1,t}, \dots, a_{N,t}] \in [0, 1]^N, \quad (2)$$

interpreted as a proposal to use $a_{i,t} C_{\text{bat}}$ Wh from home- i 's battery ($C_{\text{bat}} = 5000$ Wh).

The *feasible discharge* is clipped by demand and available energy:

$$\Delta E_{i,t} = \min \{a_{i,t} C_{\text{bat}}, d_{i,t}\}, \quad (3)$$

and the *grid energy* drawn that day is

$$E_t^{\text{grid}} = \sum_{i=1}^N (d_{i,t} - \Delta E_{i,t}). \quad (4)$$

Because the environment (per code) *fully recharges* each battery daily, inter-day SOC coupling is removed, yielding independent daily dispatch decisions.¹

The *reward* follows your implementation:

$$r_t = -E_t^{\text{grid}}, \quad (5)$$

So maximizing return

$$\sum_{t=1}^T \gamma^{t-1} r_t \quad (6)$$

(we use $\gamma \approx 1$) is equivalent to minimizing annual grid energy

$$\sum_{t=1}^T E_t^{\text{grid}}. \quad (7)$$

The *baseline* and a tight *theoretical bound* are

$$E^{\text{base}} = \sum_{t=1}^T \sum_{i=1}^N d_{i,t}, \quad (8)$$

$$E_{\min}^{\text{theory}} = \sum_{t=1}^T \max \left\{ \sum_{i=1}^N d_{i,t} - N C_{\text{bat}}, 0 \right\}, \quad (9)$$

and reported savings are

$$100 \times \left(1 - \frac{\sum_t E_t^{\text{grid}}}{E^{\text{base}}} \right) \%. \quad (10)$$

¹If SOC carryover is desired, introduce $E_{i,t}$ and dynamics $E_{i,t+1} = E_{i,t} + \eta_c P_{i,t}^{\text{ch}} \Delta t - \frac{1}{\eta_d} P_{i,t}^{\text{dis}} \Delta t$, with bounds; cf. MPC/DP formulations [33,55].

We train PPO, A2C, SAC, and TD3 (all continuous-control actor–critic methods [42–45]) within this environment; literature indicates these families balance stability, sample efficiency, and exploration for energy scheduling tasks [37,38].

3 Methodology

This section details the proposed framework for scalable deep reinforcement learning (DRL)–based battery dispatch optimization across 200 smart homes using synthetic energy consumption dataset consumption data. The methodology is organized into four main components: (i) dataset acquisition and preprocessing, (ii) custom Gymnasium environment design, (iii) DRL algorithm selection and tuning, and (iv) experimental setup and evaluation procedure. The overall workflow is illustrated in Fig. 1.

Fig. 2 summarizes this formulation as an agent–environment loop, linking the state, action, and reward definitions above to the data pipeline and evaluation.

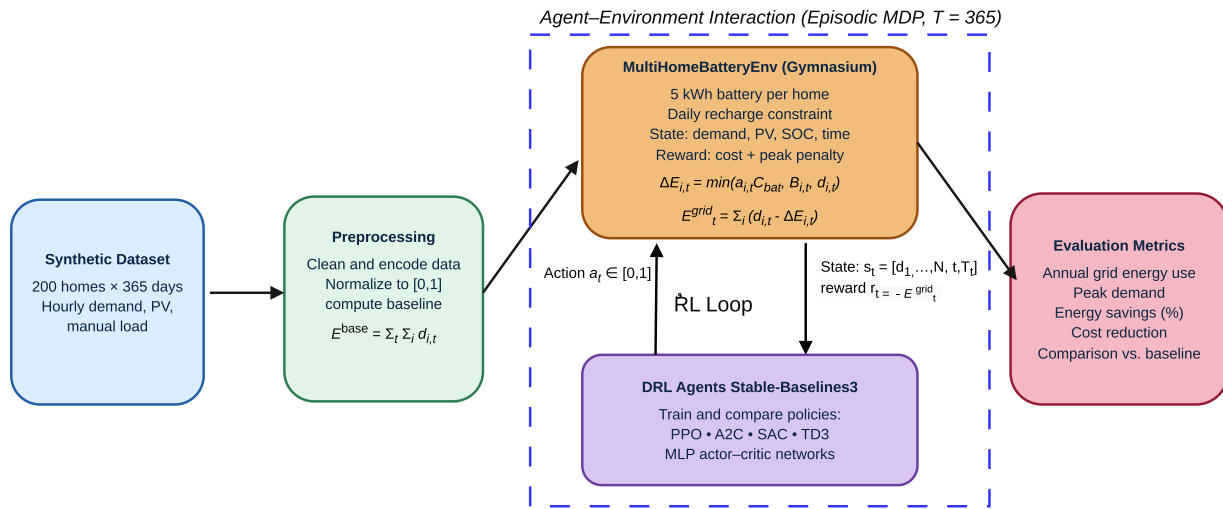


Figure 2: Episodic MDP formulation of multi-home battery dispatch: agent–environment loop with state s_t , action $a_t \in [0,1]^{200}$, and reward $r_t = -E^{\text{grid}}_t$.

3.1 Data Acquisition and Preprocessing

The data employed in this paper consists of synthetic daily energy-consumption profiles for 200 residential units, generated by Ramadan [9] to emulate realistic household demand, including a 50-home subset simulated using northern Saudi Arabian weather conditions. Each row represents a home, and each column represents a day, with values expressed in watt-hours (Wh). Missing or non-numeric values are forced to zero. The reference to calculate savings percentages is a baseline annual grid consumption (E^{base}) calculated as the sum of daily loads without battery control. The data preprocessing algorithm is formalized in Algorithm 1 and Fig. 3a.

Algorithm 1: Data acquisition and preprocessing

Data:

- **FILE_ID:** Google Drive file ID for CSV of daily consumption.
- **CSV_PATH:** Local path to save/load the dataset.

Result:

- Cleaned matrix $\mathbf{D} \in \mathbb{R}^{N_{\text{homes}} \times N_{\text{days}}}$.
-

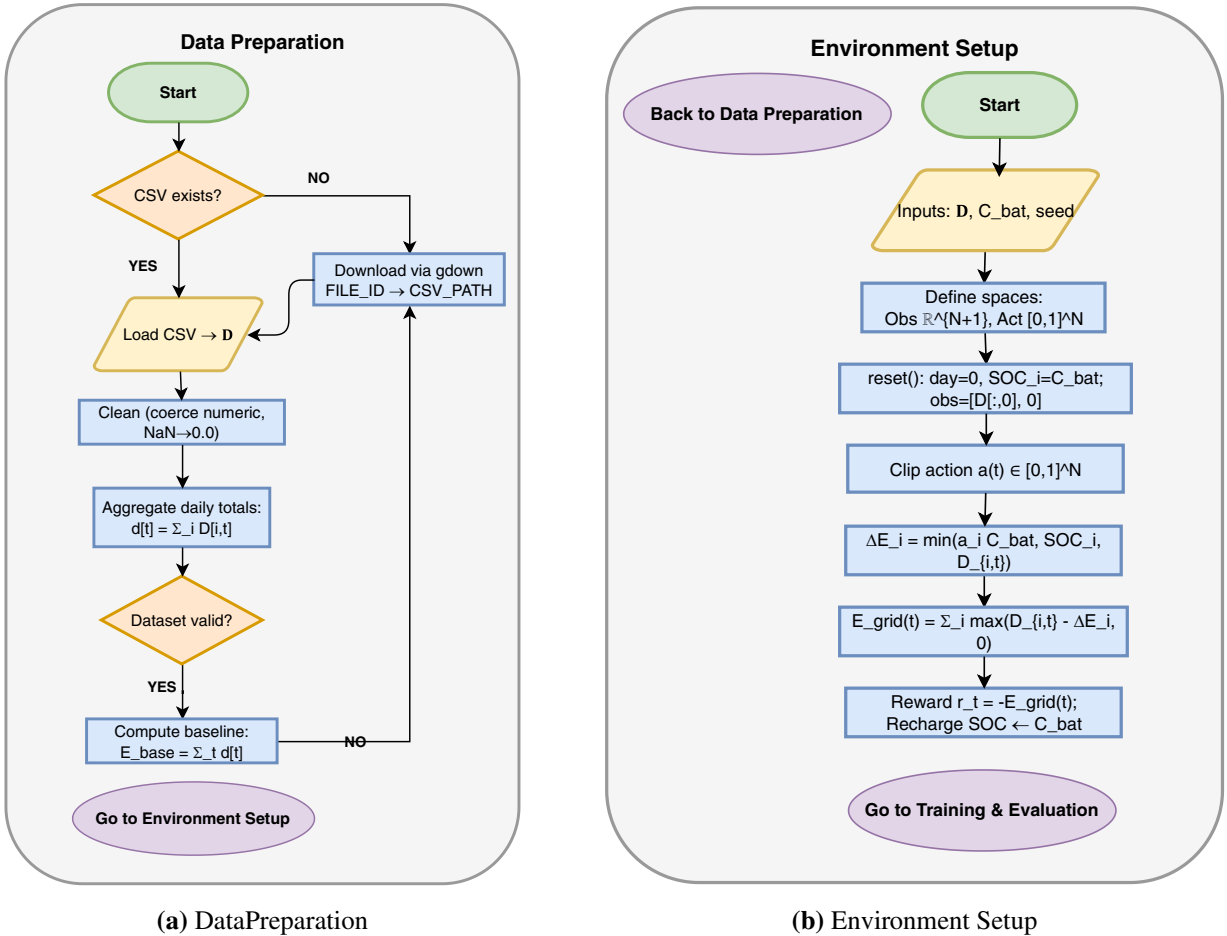
(Continued)

Algorithm 1 (continued)

```

    • Baseline annual grid energy  $E^{base}$ .
    /*Step 1: Download dataset if missing */
1  if  $\neg exists(CSV\_PATH)$  then
2      gdown.download(id=FILE_ID, output=CSV_PATH)
3  end
    /*Step 2: Load and clean dataset */
4   $D \leftarrow read\_csv(CSV\_PATH)$ 
5  Coerce non-numeric entries to numeric and replace NaN with 0.0
    /*Step 3: Calculate baseline demand */
6   $d \leftarrow sum\_rows\_by\_day(D)$ 
7   $E^{base} \leftarrow \sum_{t=1}^{N_{days}} d[t]$ 
8  return  $D, d, E^{base}$ 

```

**Figure 3:** Data preparation and environment setup diagrams.**3.2 Environment Design**

We developed a custom OpenAI Gymnasium environment, MultiHomeBatteryEnv, was developed to simulate the operational dynamics of distributed residential battery storage. Each home is equipped with an

identical battery of capacity C_{bat} Wh. The environment's state vector at each step includes the current demand for all homes and the day index. The action vector specifies the proportion of battery discharge for each home, bounded between 0 and 1. The reward is defined as the negative total grid draw for that day, incentivizing load reduction from the grid. A full daily recharge of state-of-charge (SOC) is assumed, reflecting a regime in which each home's battery is sized for diurnal cycling and is replenished each day by local renewable (e.g., rooftop PV) generation, a common operating point for residential PV-plus-storage systems. This assumption of decoupling consecutive days reduces the problem to a tractable, high-dimensional *daily allocation* task (how to distribute discharge across 200 homes to minimize grid draw) and isolates it from multi-day arbitrage. We adopt this assumption for three reasons: (i) it yields a clean, reproducible benchmark in which all four algorithms are compared on the same well-posed sub-problem; (ii) it matches the daily-resolution granularity of the available dataset, which records one demand value per home per day and thus does not expose the intra-day charge/discharge cycles needed to identify inter-day SOC trajectories; and (iii) it provides a conservative lower bound on achievable savings, since reintroducing inter-day storage can only expand the feasible set and therefore cannot reduce the savings reported here. We acknowledge that this abstraction removes inter-day coupling and seasonal storage effects that are central to some dispatch settings; the environment retains the structure needed to relax it (Section 2.1, footnote); however, we treat full SOC carryover as a direct extension (Section 5). The environment formulation is given in Algorithm 2, and a schematic overview of the environment is shown in Fig. 3b.

Algorithm 2: MULTIHOMEBATTERYENV: multi-home battery energy environment

Data:

- **D:** Daily demand per home (Wh).
- C_{bat} : Battery capacity per home (Wh).
- **seed:** Random seed.

Result:

- Gym environment with states, actions, rewards for battery dispatch.

```

/*Initialization                                     */
1 Set num_homes  $\leftarrow N_{\text{homes}}$ , num_days  $\leftarrow N_{\text{days}}$ 
2 Define observation space  $\mathbb{R}^{N_{\text{homes}}+1}$ 
3 Define action space  $[0,1]^{N_{\text{homes}}}$ 
/*Reset function                                     */
4 Day  $\leftarrow 0$ ; soc[i]  $\leftarrow C_{\text{bat}}$  for all  $i$ 
5 Return initial observation  $[\mathbf{D}[:,0], 0]$ 
/*Step function                                       */
6 Clip action  $\mathbf{a}(t)$  to  $[0,1]$ 
7  $\Delta E \leftarrow \min(\mathbf{a}(t) \cdot C_{\text{bat}}, \mathbf{soc}, \mathbf{d}_{\text{today}})$ 
8  $E^{\text{grid}} \leftarrow \sum_i (d_i(t) - \Delta E_i)$ 
9 Reward  $r_t \leftarrow -E^{\text{grid}}$ 
10 Recharge soc  $\leftarrow C_{\text{bat}}$  for next day
11 Return new observation,  $r_t$ , done flag, and info

```

3.3 DRL Algorithms

We benchmarked four widely used DRL algorithms: PPO [42], A2C [43], SAC [44], and TD3 [45]. These algorithms represent a mix of on-policy and off-policy actor-critic methods with distinct exploration and stability characteristics:

- **PPO** employs a clipped surrogate objective for stable policy updates and is known for robustness in high-variance environments.
- **A2C** uses synchronous policy and value updates across multiple environments, improving sample efficiency for temporally correlated tasks.
- **SAC** optimizes a maximum-entropy objective, encouraging exploration and robustness to stochasticity.
- **TD3** incorporates twin critics and delayed policy updates to mitigate overestimation bias in continuous action spaces.

Each algorithm was trained using the `stable_baselines3` library, with hyperparameters selected through manual search over a small set of candidate values, guided by convergence stability and energy-savings performance on preliminary runs (Table 1). We did not employ automated hyperparameter optimization; a systematic search (e.g., grid search or Bayesian optimization), particularly for PPO (Section 5), is left to future work. The full optimization loop is shown in Algorithm 3, while algorithm-specific diagrams will be provided in Figs. 3 and 4.

Table 1: Hyperparameters for DRL algorithms.

Parameter	PPO	A2C	SAC	TD3
Learning rate	3×10^{-4}	7×10^{-4}	3×10^{-4}	1×10^{-3}
Discount factor γ	0.99	0.99	0.99	0.99
Batch size	64	64	256	256
Policy network	$2 \times [64]$	$2 \times [64]$	$2 \times [256]$	$2 \times [256]$
Train timesteps	200,000	200,000	200,000	200,000
Entropy coeff.	0.0	0.0	auto	N/A
Seed	42	42	42	42

Algorithm 3: DRL-based multi-home battery energy optimization

Data:

- Env from Algorithm 2.
- $\mathcal{A} = \{\text{PPO, A2C, SAC, TD3}\}$: DRL algorithms.
- `TOTAL_TIMESTEPS`: training budget per algorithm.
- Baseline annual energy E^{base} from Algorithm 1.
- C_{bat} : battery capacity per home.
- N_{homes} : number of homes.

Result:

- Best-performing algorithm a^* .
- Annual grid energy $E_{a^*}^{\text{grid}}$ and savings percentage.
- Theoretical minimum grid draw $E_{\text{min}}^{\text{theory}}$ and maximum savings.

/*Step 1: Training each algorithm

*/

```

1 foreach  $a \in \mathcal{A}$  do
2   Create log directory for  $a$ 
3   Initialize  $M_a \leftarrow a(\text{MlpPolicy, env, seed})$ 
4   Train  $M_a$  with learn(TOTAL_TIMESTEPS)
5   Save  $M_a$  and training logs

```

(Continued)

Algorithm 3 (continued)

```

6  end
   /*Step 2: Evaluation of trained models */
7  foreach  $a \in \mathcal{A}$  do
8      Reset eval env; set total_grid  $\leftarrow 0$ 
9      while episode not done do
10         Predict action  $\mathbf{u}$  from  $M_a$  (deterministic)
11         Step env with  $\mathbf{u}$  to get info["total_grid"]
12         Accumulate into total_grid
13     end
14      $E_a^{\text{grid}} \leftarrow \text{total\_grid}$ 
15      $\text{savings\%}_a \leftarrow \left(1 - \frac{E_a^{\text{grid}}}{E_{\text{base}}}\right) \times 100$ 
16 end
   /*Step 3: Select best-performing algorithm */
17  $a^* \leftarrow \arg \max_a \text{savings\%}_a$ 
18  $E_{a^*}^{\text{grid}} \leftarrow E_{a^*}^{\text{grid}}$ 
19  $\text{savings\%}_{a^*} \leftarrow \max_a \text{savings\%}_a$ 
   /*Step 4: Compute theoretical upper bound */
20 for  $t \leftarrow 1$  to  $N_{\text{days}}$  do
21      $E_t^{\text{theory}} \leftarrow \max(d_t - N_{\text{homes}} \cdot C_{\text{bat}}, 0)$ 
22 end
23  $E_{\text{min}}^{\text{theory}} \leftarrow \sum_t E_t^{\text{theory}}$ 
24  $\text{savings\%}_{\text{max}} \leftarrow \left(1 - \frac{E_{\text{min}}^{\text{theory}}}{E_{\text{base}}}\right) \times 100$ 
25 return  $a^*, E_{a^*}^{\text{grid}}, \text{savings\%}_{a^*}, E_{\text{min}}^{\text{theory}}, \text{savings\%}_{\text{max}}$ 

```

3.4 Hyperparameter Settings

Table 1 summarizes the final hyperparameter configuration for each algorithm. Values were selected based on convergence stability, training time constraints, and energy savings performance.

3.5 Experimental Setup

The experiments were conducted on Google Colab Pro, utilizing a paid GPU instance equipped with an NVIDIA A100 GPU. The runtime environment was configured with Python 3.10 and stable_baselines3 1.8.0. Table 2 lists the software stack.

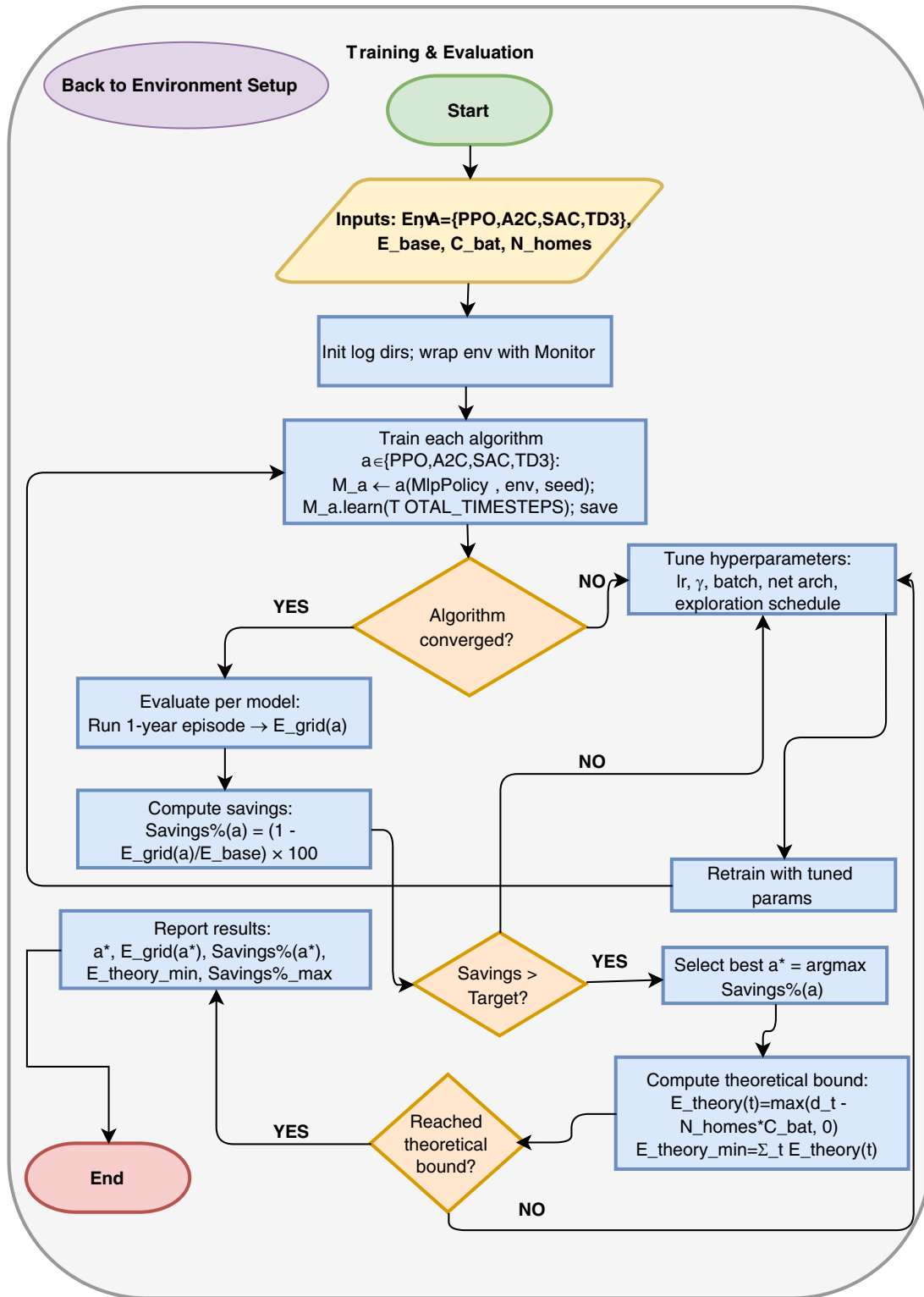
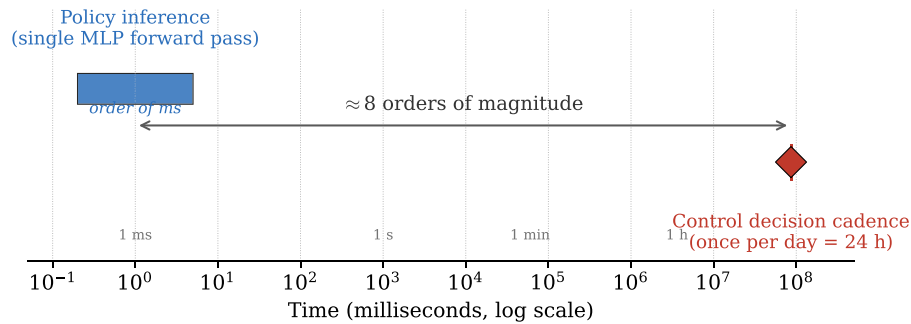


Figure 4: Training and evaluation of RL algorithms.

Table 2: Technologies and libraries used.

Tool/Library	Version
Python	3.10
Pandas	1.5.3
NumPy	1.24.2
Gymnasium	0.28.1
Stable Baselines3	1.8.0
Matplotlib	3.7.1
Seaborn	0.12.2

We distinguish training cost from inference cost. For deployment, inference is the key metric and is non-limiting: dispatch is performed once per day, whereas the learned MLP policy requires only one forward pass, typically on the order of milliseconds on commodity hardware (Fig. 5). Thus, a trained policy could run on an edge controller without GPU acceleration. Training cost varies by algorithm: A2C is the lightest, PPO adds repeated rollout optimization, and SAC/TD3 are more expensive due to replay buffers, larger networks, and twin-critic updates. All algorithms used the same 200,000-timestep budget on a single NVIDIA A100. Detailed wall-clock and memory profiling was not performed and is left for future benchmarking.

**Figure 5:** Policy-inference latency vs. daily control cadence.

Each algorithm was trained for 200,000 timesteps using the same environment and seed for reproducibility.

4 Results

This section contains the empirical analysis of the proposed multi-home battery dispatch framework based on synthetic daily consumption profiles for 200 smart homes. Four deep reinforcement learning algorithms, PPO, A2C, SAC, and TD3, were trained and benchmarked in the custom Gymnasium environment. Performance was measured by annual grid energy consumption and percent savings compared to a no-control baseline, training convergence patterns, daily savings distributions, and exemplary operational time series. This is aimed at measuring the relative effectiveness, stability, and operating behaviour of each algorithm under realistic demand and renewable generation variability.

Table 3 summarizes the annual grid energy consumption and the corresponding percentage savings achieved by each algorithm relative to the baseline. all DRL methods achieve substantial grid-energy reductions, whereas PPO with default parameters barely improves upon the baseline (0.2%). The most effective algorithm under the tested single-seed configuration is A2C, reducing annual grid consumption

from 2.92×10^8 Wh to 1.52×10^8 Wh, a 48.2% savings. TD3 (44.3%) and SAC (43.8%) follow closely, demonstrating that off-policy actor-critic methods are well-suited to high-dimensional continuous control. These results corroborate findings from previous studies that DRL can learn coordinated dispatch strategies that outperform heuristic or rule-based controllers, and that DRL control of distributed batteries can achieve a large fraction of the theoretical optimum.

Table 3: Grid consumption and savings percentage for each algorithm.

Algorithm	Grid_Wh	Savings_%
A2C	1.517934×10^8	48.2
TD3	1.631083×10^8	44.3
SAC	1.647358×10^8	43.8
PPO	2.922627×10^8	0.2
Theoretical optimum (pooled, Eq. (9))	≈ 0 (most days)	$\rightarrow$ near 100

The pooled theoretical bound of Eq. (9) is loose for this dataset: the mean daily aggregate demand ($E^{\text{base}}/T \approx 8.0 \times 10^5$ Wh) is below the total installed daily battery capacity ($NC_{\text{bat}} = 1.0 \times 10^6$ Wh), so $\sum_i d_{i,t} - NC_{\text{bat}} \leq 0$ on most days, and the bound permits near complete grid-energy elimination.

Fig. 6 plots the total grid energy per episode a full year against training episodes. A2C exhibits a steady reduction in grid consumption over the first 50–100 episodes before oscillating around a stable value; SAC and TD3 converge quickly to their final performance within a few dozen episodes; PPO remains flat, indicating negligible learning. The ability of A2C, SAC, and TD3 to adapt to the environment contrasts sharply with static rule-based methods and underscores the capacity of DRL to handle complex, stochastic energy systems. The near-flat trajectories of SAC and TD3 reflect rapid convergence rather than an initialization artifact. Both are off-policy methods that perform a gradient update at essentially every environment step using the replay buffer; within the first 365-step episode, they therefore accumulate hundreds of updates, so most of their learning is completed before the first per-episode point is logged. The apparent flatness is consequently a resolution effect of the per-episode horizontal axis, not an absence of learning. That these curves settle well below PPO's untrained baseline ($\sim 2.0 \times 10^8$ Wh) and evaluate deterministically to 43.8% (SAC) and 44.3% (TD3) savings confirms that the flat lines correspond to genuinely learned low-grid-energy policies. By contrast, the on-policy A2C learns from short fresh rollouts and its improvement is visibly distributed across the first ~ 50 episodes. A finer-grained, per-timestep logging (or a logarithmic episode axis) would resolve the early transient of the off-policy methods explicitly.

Fig. 7 visualizes the savings reported in Table 3. It shows that A2C consistently attains the highest savings, approaching half of the baseline consumption, followed by TD3 and SAC. The poor performance of PPO is consistent with our observation that PPO does not explore the action space effectively under default hyperparameters in this high-dimensional setting. These comparative results mirror trends observed in other energy management tasks, where actor-critic DRL algorithms such as SAC and TD3 often outperform PPO and deterministic policy gradients.

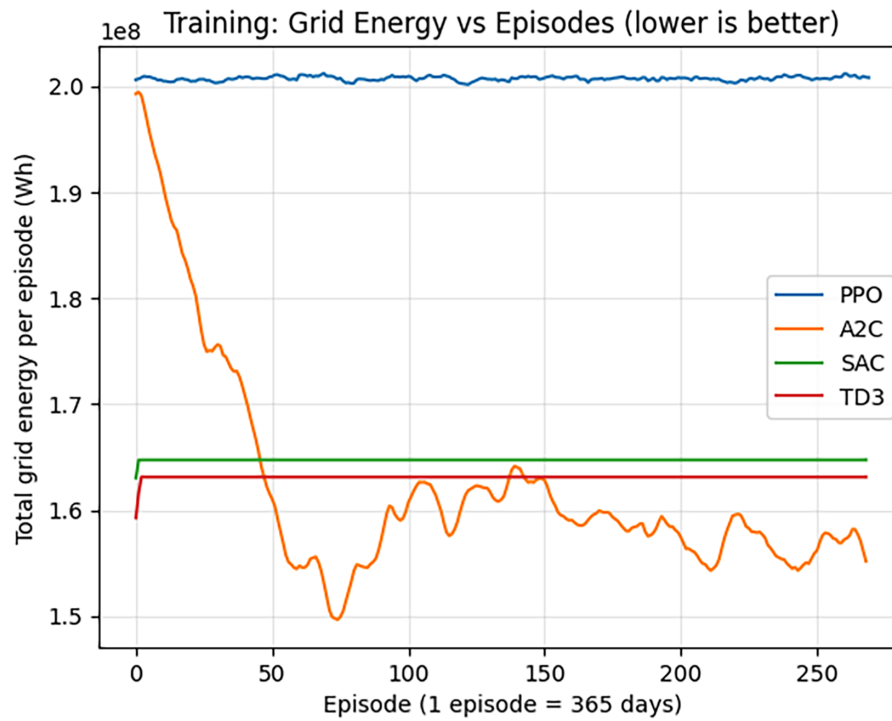


Figure 6: Total grid energy per episode (year) vs. training episodes for each algorithm.

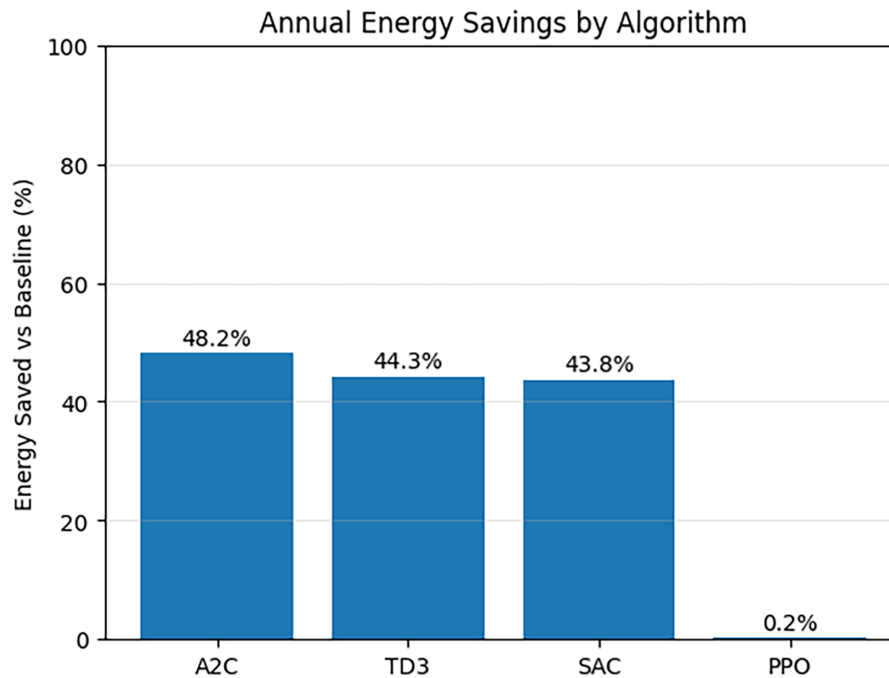


Figure 7: Annual energy savings achieved by each algorithm relative to the no-control baseline.

Fig. 8 shows the distribution of daily savings for the A2C policy. Most days achieve between 44% and 52% reduction, with a mean around 48%. The relatively narrow spread indicates that the policy is robust

across different demand profiles and seasons. Occasional outliers near 44% correspond to days with unusually high demand or limited renewable generation, where the batteries exhaust their capacity early. Such stability is desirable for long-term smart-grid operation, as it implies predictable peak-shaving performance.

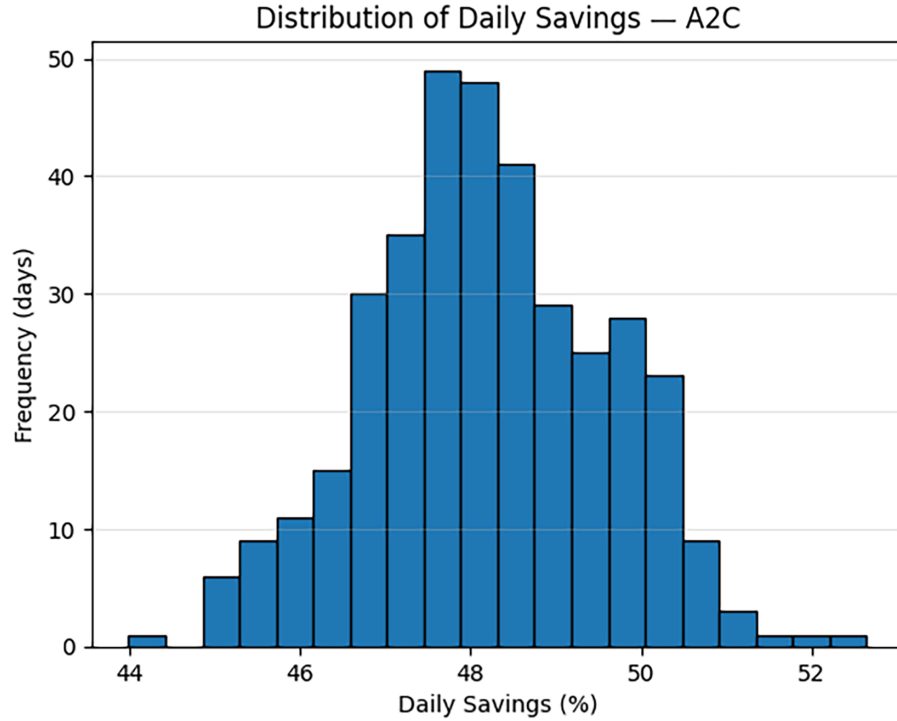


Figure 8: Distribution of daily energy savings achieved by the A2C policy.

[Fig. 9](#) plots a 30-day segment of the baseline aggregated demand vs. the grid energy after applying the A2C controller. The shaded region represents energy supplied by the batteries. The learned policy systematically discharges batteries during peak demand periods, resulting in a significant reduction of grid draw on most days. This behaviour aligns with reported advantages of DRL in shifting energy consumption away from high-cost periods and diversifying battery actions. By contrast, rule-based schemes often discharge too early or too late, leading to suboptimal peak shaving.

The results demonstrate that DRL can learn effective battery dispatch strategies that significantly reduce grid consumption and that actor–critic algorithms, particularly A2C and TD3, outperform PPO in this high-dimensional multi-home setting. These findings support the growing consensus that DRL is a promising approach for distributed energy management and highlight the importance of algorithm selection and hyperparameter tuning for practical deployments.

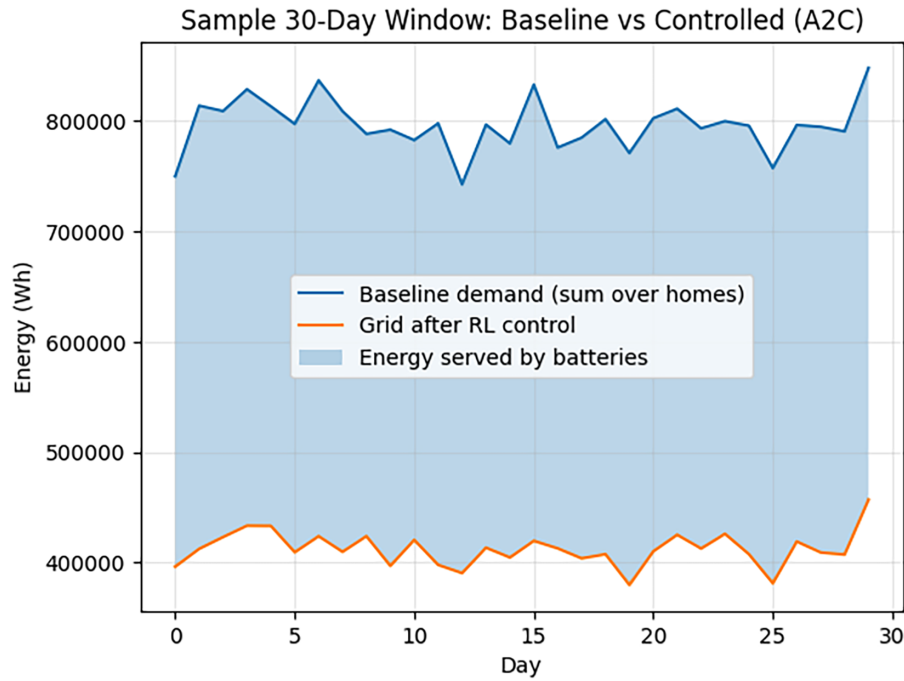


Figure 9: Sample 30-day window for the A2C policy: baseline demand vs. grid energy after control.

5 Discussion

The comparative results in [Section 4](#) highlight clear performance differences among the four evaluated DRL algorithms. Consistent with prior work in continuous control domains, the on-policy A2C and the off-policy TD3 emerged as the most effective controllers for multi-home battery dispatch, achieving 48.2% and 44.3% annual savings, respectively. Several factors may explain this outcome. First, A2C's synchronous policy updates with multiple parallel environments promote stable gradient estimates and more efficient exploration in environments with high temporal correlations, such as daily and seasonal demand cycles [\[43,56\]](#). This aligns with observations from other energy system studies, where A2C's ability to learn temporally consistent policies led to superior load-shifting performance [\[37,57\]](#).

TD3's strong performance can be attributed to its twin critic networks and delayed policy updates, which mitigate overestimation bias and improve stability in high-dimensional action spaces [\[45\]](#). These design features are particularly relevant in battery scheduling tasks, where control actions are continuous and must balance short-term grid relief with long-term state-of-charge preservation [\[58\]](#). The ability of TD3 to learn nuanced control rules without excessive variance suggests it is a promising candidate for deployment in distributed energy resource management systems (DERMS) [\[59\]](#).

SAC's performance, while slightly lower than TD3, is consistent with findings that its maximum entropy objective can sometimes trade off deterministic exploitation for exploration [\[44\]](#). This trade-off is beneficial in highly stochastic environments, but in our case where daily load patterns are partly predictable, the additional randomness may have led to less aggressive peak shaving compared to A2C and TD3. Nevertheless, SAC's 43.8% savings still represent a substantial improvement over the baseline, reaffirming that entropy-regularised methods remain competitive in demand-side management [\[38,41\]](#).

PPO's near-zero savings (0.2%) contrast with its strong performance in many continuous-control tasks [\[42\]](#), and warrant careful interpretation. We emphasize that all algorithms were run under a single, fixed configuration to enable a controlled comparison ([Table 1](#)); The reported PPO result, therefore, reflects

a *default-hyperparameter baseline* rather than a tuned upper bound. Notably, A2C attains 48.2% using an identical 2×256 neural-network architecture. $2 \times [64]$ policy network, which rules out raw model capacity as the cause and points instead to PPO's update dynamics in this specific regime. We hypothesize that, in a 200-dimensional continuous action space with a long (365-step) horizon, PPO's clipped surrogate objective combined with the zero entropy coefficient used here [Table 1](#) suppresses exploration early in training, allowing the Gaussian policy to collapse toward a near-inactive (low-discharge) solution that the clipping range then locks in—yielding behavior close to the no-control baseline. This is consistent with reports that PPO's effective performance is highly sensitive to code-level and hyperparameter choices, and can degrade sharply in high-dimensional action spaces without targeted tuning [\[60,61\]](#). We stress that this explanation is a hypothesis grounded in the literature and in the observed learning curves ([Fig. 6](#), where PPO remains flat) rather than a claim validated by ablation; isolating the individual contributions of clipping, entropy regularization, rollout length, and learning rate would require a dedicated sensitivity study, which we identify as a limitation of the present work.

The environment exposes a 200-dimensional continuous action space one discharge fraction per home, which raises two questions: whether learning converges at this scale, and whether centralized control is realistic. First, our results provide direct evidence: A2C, SAC, and TD3 all converge to stable policies within tens of episodes ([Fig. 6](#)), indicating that actor-critic methods with factored Gaussian (A2C/SAC) or deterministic (TD3) policies scale to this dimensionality without prohibitive sample cost. This is consistent with these algorithms' use in high-dimensional continuous control elsewhere, and is aided here by the problem's structure: because batteries recharge daily, the per-home decisions are weakly coupled (interacting only through the shared grid-draw objective), so the effective decision complexity is far lower than the raw 200-D action space suggests. PPO's failure to converge (0.2%) is therefore best read as an algorithm-specific exploration issue ([Section 5](#)) rather than a fundamental barrier imposed by dimensionality, since three of four algorithms succeed in the same space.

On the second question, we treat the single centralized agent as a modeling abstraction for establishing achievable savings and a reproducible benchmark, not as a literal deployment proposal. A monolithic 200-output controller raises practical concerns single point of failure, communication overhead, privacy of per-home demand, and poor elasticity as homes join or leave. Realistic large-scale deployment would more naturally adopt a decentralized or multi-agent architecture (e.g., per-home or per-cluster agents with parameter sharing, or a hierarchical aggregator coordinating local controllers), which preserves a low per-agent action dimension while scaling to arbitrary community sizes [\[25,49\]](#). We therefore position the centralized formulation as an upper-reference benchmark and identify a multi-agent reformulation as the natural route to deployment-scale and beyond-200-home settings.

The robustness of the A2C policy, demonstrated by the narrow spread in daily savings [Fig. 8](#), is particularly noteworthy. High day-to-day consistency is critical for grid operators, as it ensures predictable load-shaving benefits and supports reliable integration with wholesale market bidding strategies [\[62\]](#). Furthermore, the operational patterns observed in [Fig. 9](#) confirm that DRL agents can learn context-aware strategies charging batteries during low-cost, low-demand hours and discharging during peaks, mirroring strategies developed in model-based optimization frameworks [\[63\]](#) but without requiring explicit demand forecasting.

From a broader perspective, these findings reinforce the growing body of evidence that DRL offers a scalable, data-driven alternative to heuristic and rule-based demand-side management methods [\[64\]](#). Unlike fixed strategies, DRL agents can adapt to evolving grid conditions, making them suitable for future energy systems characterised by increasing penetration of distributed renewables, electric vehicles, and flexible

loads [65,66]. Our reproducible benchmarking framework also enables systematic comparison of algorithms under consistent conditions, addressing the lack of standardisation noted in previous reviews [37,67].

The reward in Eq. (5) targets a single physical objective minimizing total grid energy draw and deliberately omits monetary terms such as battery degradation, time-of-use tariffs, and demand charges. This scoping is intentional: energy minimization is a well-defined, price-agnostic objective that isolates the dispatch-learning problem from market and tariff assumptions, yielding a transparent and reproducible benchmark that is not tied to a particular pricing regime or geography. It also provides a conservative reference, since policies that minimize aggregate grid energy already perform the peak-shaving behavior (Fig. 9) that underlies most economic benefit.

We emphasize, however, that the framework is modular with respect to the objective: the environment dynamics (Eqs. (3) and (4)) are unchanged if the scalar reward is replaced by a weighted economic objective, for example

$$r_t = - \left(\underbrace{\pi_t E_t^{\text{grid}}}_{\text{energy/tariff cost}} + \lambda_{\text{pk}} \pi^{\text{pk}} \max_{\tau \leq t} E_{\tau}^{\text{grid}} + \lambda_{\text{deg}} \sum_{i=1}^N c_{\text{deg}} |\Delta E_{i,t}| \right), \quad (11)$$

where π_t is the (possibly time-of-use) energy price, π^{pk} a demand charge applied to the running peak, c_{deg} a per-Wh throughput degradation cost, and $\lambda_{\text{pk}}, \lambda_{\text{deg}}$ trade-off weights. Setting $\pi_t \equiv 1$ and $\lambda_{\text{pk}} = \lambda_{\text{deg}} = 0$ recovers the energy-only reward used in this study. Because these terms enter only through the reward signal and not the state-transition logic, the same agents can be retrained against Eq. (11) without modifying the environment. We identify this economic extension together with the calibration of degradation and tariff parameters to a specific market as an important direction for future work.

The experimental results suggest that actor-critic DRL algorithms, particularly A2C and TD3, are well-suited for large-scale battery dispatch in smart grids. Future research should explore hybrid approaches that combine the stability of A2C with the precision of TD3, investigate transfer learning for faster adaptation across different regions, and evaluate the integration of market price signals to enhance economic benefits in addition to energy savings.

A primary limitation is the daily SOC reset, which removes inter-day coupling. Relaxing it requires sub-daily demand and generation data to identify intra-day charge/discharge cycles; with the SOC dynamics already specified in the footnote of Section 2.1, extending the environment to carry SOC across days—and quantifying its impact on the reported savings is a clear next step that would also enable multi-day arbitrage and seasonal storage strategies.

This study reports results from one fixed seed (seed = 42) to ensure exact reproducibility; therefore, the values are treated as point estimates rather than statistically averaged results. Since evaluation is deterministic over the fixed 365-day episode, variability arises only from training. The large performance gap between A2C (48.2%) and PPO (0.2%) supports the qualitative ranking, although small differences among A2C, SAC, and TD3 should be interpreted cautiously. Multi-seed evaluation with mean $\pm$ standard deviation is left for future confirmatory work.

6 Conclusion

This paper presents a scalable deep reinforcement learning algorithm for optimizing distributed battery dispatch to 200 smart homes using a realistic energy consumption dataset. To accurately reflect the operational limits and energy flows of a multi-home environment, an OpenAI Gymnasium environment was created. Four state-of-the-art DRL algorithms (PPO, A2C, SAC, and TD3) were trained and tested in

the same environment, and their performance was compared in terms of annual grid energy draw and percentage savings compared to a no-storage baseline. The experimental results showed that the A2C agent achieved the highest savings (48.2%), followed by TD3 (44.3%), SAC (43.8%), and PPO (0.2%). These findings demonstrate the effectiveness of off-policy and synchronous actor-critic approaches for large-scale residential battery coordination, while also highlighting the limitations of some on-policy methods in this area. Furthermore, the use of a repeatable benchmarking environment and dataset makes this work a valuable baseline for future research in DRL-based energy optimization.

Building directly on the limitations identified in this study, we propose several concrete extensions. Firstly, we suggest relaxing the daily state of charge (SOC) reset in Section 3 by implementing the inter-day storage dynamics mentioned in the footnote of Section 2.1. This would require sub-daily demand and generation data and would allow for multi-day arbitrage and seasonal storage. Secondly, we recommend replacing the energy-only reward (Eq. (5)) with the economic objective of Eq. (11), which would involve jointly optimizing grid energy, time-of-use tariffs, demand charges, and throughput-based battery degradation. Thirdly, we propose reformulating the centralized 200-dimensional controller in Section 5 as a decentralized or hierarchical multi-agent system with per-home or per-cluster agents and parameter sharing. This would eliminate the single point of failure and allow for scaling beyond 200 homes. Fourthly, we suggest conducting a PPO-specific sensitivity analysis (entropy coefficient, clip range, rollout length, learning rate) to test the exploration-collapse hypothesis proposed in Section 5 and its resulting 0.2% result. Fifthly, we recommend performing a multi-seed statistical validation with reported mean $\pm$ standard deviation and paired significance tests to formally confirm the algorithm ranking. Finally, we propose a controlled computational benchmark (per-algorithm training wall-clock time, memory, and inference latency across GPU and edge hardware) to complement the deployability analysis in the study.

Acknowledgement: The authors would like to thank the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for supporting this research through project number (0054-1446-S).

Funding Statement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work through the project number (0054-1446-S).

Author Contributions: The authors confirm the contribution to the paper as follows: study conception and design: Abdullah Alshammari, Ammar Ahmed E. Elhadi, Ashraf Osman Ibrahim; methodology development and experimental design: Abdullah Alshammari, Ammar Ahmed E. Elhadi, Ashraf Osman Ibrahim; data analysis and interpretation: Abdullah Alshammari, Ammar Ahmed E. Elhadi, Ashraf Osman Ibrahim; manuscript preparation and writing: Abdullah Alshammari, Ammar Ahmed E. Elhadi, Ashraf Osman Ibrahim; review and editing: Abdullah Alshammari, Ammar Ahmed E. Elhadi, Ashraf Osman Ibrahim. All authors reviewed and approved the final version of the paper.

Availability of Data and Materials: The synthetic daily energy-consumption dataset supporting the findings of this study is the *Energy Consumption Dataset For Smart Homes* by Ramadan [9], publicly available on Mendeley Data under the CC BY 4.0 license. The dataset can be accessed at <https://data.mendeley.com/datasets/5h86vnxw68/1> or via DOI <https://doi.org/10.17632/5h86vnxw68.1>. The specific subset used in this paper is the 200-home daily consumption profile (200_consumption_data.csv). The custom OpenAI Gymnasium environment and training scripts developed for this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu G, Yu W, Griffith D, Golmie N, Moulema P. Toward integrating distributed energy resources and storage devices in smart grid. *IEEE Internet Things J.* 2016;4(1):192–204. doi:10.1109/jiot.2016.2640563.
2. Madjovski D, Dumancic I, Tranchita C. Dynamic modeling of distribution power systems with renewable generation for stability analysis. *Energies.* 2024;17(20):5178. doi:10.3390/en17205178.
3. Shamarova N, Suslov K, Ilyushin P, Shushpanov I. Review of battery energy storage systems modeling in microgrids with renewables considering battery degradation. *Energies.* 2022;15(19):6967.
4. Ethirajan V, Mangaiyarkarasi S. An in-depth survey of latest progress in smart grids: paving the way for a sustainable future through renewable energy resources. *J Electr Sys Inform Technol.* 2025;12(1):9. doi:10.1186/s43067-025-00195-z.
5. Moreno Escobar JJ, Morales Matamoros O, Tejeida Padilla R, Lina Reyes I, Quintana Espinosa H. A comprehensive review on smart grids: challenges and opportunities. *Sensors.* 2021;21(21):6978. doi:10.3390/s21216978.
6. Stecca M, Elizondo LR, Soeiro TB, Bauer P, Palensky P. A comprehensive review of the integration of battery energy storage systems into distribution networks. *IEEE Open J Ind Electron Soc.* 2020;1:46–65. doi:10.1109/ojies.2020.2981832.
7. Mahmood M, Chowdhury P, Yeassin R, Hasan M, Ahmad T, Chowdhury NUR. Impacts of digitalization on smart grids, renewable energy, and demand response: an updated review of current applications. *Energy Convers Manag X.* 2024;24:100790.
8. Cui F, An D, Zhang G. A game strategy for demand response based on load monitoring in smart grid. *Front Energy Res.* 2023;11:1240542. doi:10.3389/fenrg.2023.1240542.
9. Arslan S, Zekun G. Smart Home Energy Consumption Dataset for AI-Driven Residential Load Forecasting and Demand Response. *IEEE Dataport.* 2026. doi:10.21227/7c3j-xs34.
10. Gomes I, Bot K, Ruano MG, Ruano A. Recent techniques used in home energy management systems: a review. *Energies.* 2022;15(8):2866. doi:10.3390/en15082866.
11. Ruelens F, Claessens BJ, Vandael S, De Schutter B, Babuška R, Belmans R. Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Trans Smart Grid.* 2016;8(5):2149–59. doi:10.1109/tsg.2016.2517211.
12. Azuatalam D, Paridari K, Ma Y, Förstl M, Chapman AC, Verbič G. Energy management of small-scale PV-battery systems: a systematic review considering practical implementation, computational requirements, quality of input data and battery degradation. *Renew Sustain Energy Rev.* 2019;112:555–70.
13. Pamulapati T, Cavus M, Odigwe I, Allahham A, Walker S, Giaouris D. A review of microgrid energy management strategies from the energy trilemma perspective. *Energies.* 2022;16(1):289.
14. Schwarz H, Bertsch V, Fichtner W. Two-stage stochastic, large-scale optimization of a decentralized energy system: a case study focusing on solar PV, heat pumps and storage in a residential quarter. *OR Spectr.* 2018;40(1):265–310. doi:10.1007/s00291-017-0500-4.
15. Ruddick J, Camargo LR, Putratama MA, Messagie M, Coosemans T. TreeC: a method to generate interpretable energy management systems using a metaheuristic algorithm. *Knowl Based Syst.* 2025;309:112756.
16. Saroha P, Singh G, Lilhore UK, Simaiya S, Khan M, Alroobaea R, et al. Dynamic appliance scheduling and energy management in smart homes using adaptive reinforcement learning techniques. *Sci Rep.* 2025;15(1):24594. doi:10.1038/s41598-025-08125-9.
17. Yu L, Qin S, Zhang M, Shen C, Jiang T, Guan X. Deep reinforcement learning for smart building energy management: a survey. *arXiv:2008.05074.* 2020.
18. Wei T, Wang Y, Zhu Q. Deep reinforcement learning for building HVAC control. In: *Proceedings of the 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC); 2017 Jun 18–22; Austin, TX, USA.* p. 1–6.
19. Yu L, Qin S, Xu Z, Guan X, Shen C, Yue D. Optimal operation of a hydrogen-based building multi-energy system based on deep reinforcement learning. *arXiv:2109.10754.* 2021.
20. Hou H, He Z, Xiang M, Lu Y, Yang J, Huang L, et al. Interval multi-objective optimization for low-carbon building energy management system upon deep reinforcement learning. *IEEE Trans Ind Appl.* 2025;61(2):2193–202.

21. Liang T, Zhang X, Tan J, Jing Y, Liangnian L. Deep reinforcement learning-based optimal scheduling of integrated energy systems for electricity, heat, and hydrogen storage. *Elect Power Syst Res.* 2024;233(4):110480. doi:10.1016/j.epsr.2024.110480.
22. Liu Y, Lu Q, Yu Z, Chen Y, Yang Y. Reinforcement learning-enhanced adaptive scheduling of battery energy storage systems in energy markets. *Energies.* 2024;17(21):5425. doi:10.3390/en17215425.
23. Fan J, Wang H. Deep reinforcement learning for community battery scheduling under uncertainties of load, PV generation, and energy prices. In: *Proceedings of the 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2)*; 2023 Dec 15–18; Hangzhou, China. p. 4871–6.
24. Sage M, Zhao YF. Deep reinforcement learning for economic battery dispatch: a comprehensive comparison of algorithms and experiment design choices. *J Energy Storage.* 2025;115:115428.
25. Chen J, Zhao Y, Wang M, Yang K, Ge Y, Wang K, et al. Multi-timescale reward-based DRL energy management for regenerative braking energy storage system. *IEEE Trans Transp Electrification.* 2025;11(3):7488–500. doi:10.1109/tte.2025.3528255.
26. Shaqour A, Hagishima A. Systematic review on deep reinforcement learning-based energy management for different building types. *Energies.* 2022;15(22):8663. doi:10.3390/en15228663.
27. Fang X, Misra S, Xue G, Yang D. Smart grid—the new and improved power grid: a survey. *IEEE Commun Surv Tutor.* 2011;14(4):944–80.
28. Palensky P, Dietrich D. Demand side management: demand response, intelligent energy systems, and smart loads. *IEEE Trans Ind Inform.* 2011;7(3):381–8.
29. Zakeri B, Syri S. Electrical energy storage systems: a comparative life cycle cost analysis. *Renew Sustain Energy Rev.* 2015;42:569–96.
30. Hannan M, Hoque MM, Mohamed A, Ayob A. Review of energy storage systems for electric vehicle applications: issues and challenges. *Renew Sustain Energy Rev.* 2017;69(7):771–89. doi:10.1016/j.rser.2016.11.171.
31. Becchi L, Belloni E, Bindi M, Intravaia M, Grasso F, Lozito GM, et al. A computationally efficient rule-based scheduling algorithm for battery energy storage systems. *Sustainability.* 2024;16(23):10313. doi:10.3390/su162310313.
32. Keshtkar A, Arzanpour S, Keshtkar F. Adaptive residential demand-side management using rule-based techniques in smart grid environments. *Energy Build.* 2016;133(12):281–94. doi:10.1016/j.enbuild.2016.09.070.
33. Tang R, Wang S. Model predictive control for thermal energy storage and thermal comfort optimization of building demand response in smart grids. *Appl Energy.* 2019;242:873–82. doi:10.1016/j.apenergy.2019.03.038.
34. Li Y, Peng J, Jia H, Zou B, Hao B, Ma T, et al. Optimal battery schedule for grid-connected photovoltaic-battery systems of office buildings based on a dynamic programming algorithm. *J Energy Storage.* 2022;50(6):104557. doi:10.1016/j.est.2022.104557.
35. Ricardo CH, Adriana LH, Nelson DA. Energy management supported on genetic algorithms for the equalization of battery energy storage systems in microgrid systems. *J Energy Storage.* 2023;72:108510.
36. Radosavljević J, Jevtić M, Klimenta D. Energy and operation management of a microgrid using particle swarm optimization. *Eng Optim.* 2016;48(5):811–30.
37. Glavic M, Javadi MS, Le Ny J. Reinforcement learning for electric power system decision and control: past considerations and perspectives. *Annu Rev Control.* 2021;52:118–42.
38. Vázquez-Canteli JR, Nagpal Z. Reinforcement learning for demand response: a review of algorithms and modeling techniques. *Appl Energy.* 2019;235:1072–89.
39. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature.* 2015;518(7540):529–33. doi:10.1038/nature14236.
40. Hosseini E, Horrillo-Quintero P, Carrasco-Gonzalez D, Garcia-Trivino P, Sarrias-Mena R, Garcia-Vazquez CA, et al. Reinforcement learning-based energy management system for lithium-ion battery storage in multilevel microgrid. *J Energy Storage.* 2025;109(2):115114. doi:10.1016/j.est.2024.115114.
41. Huo W, Zhao T, Yang F, Chen Y. An improved soft actor-critic based energy management strategy of fuel cell hybrid electric vehicle. *J Energy Storage.* 2023;72(27):108243. doi:10.1016/j.est.2023.108243.
42. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. *arXiv:1707.06347.* 2017.

43. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 19–24; New York, NY, USA. p. 1928–37.
44. Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 1861–70.
45. Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 1582–91.
46. Alonso M, Amaris H, Martin D, de la Escalera A. Proximal policy optimization for energy management of electric vehicles and PV storage units. *Energies*. 2023;16(15):5689. doi:10.3390/en16155689.
47. Ji Y, Wang J, Xu J, Fang X, Zhang H. Real-time energy management of a microgrid using deep reinforcement learning. *Energies*. 2019;12(12):2291. doi:10.3390/en12122291.
48. Xu J, Li K, Abusara M. Preference based multi-objective reinforcement learning for multi-microgrid system optimization problem in smart grid. *Memetic Comput*. 2022;14(2):225–35. doi:10.1007/s12293-022-00357-w.
49. Lu Y, Xiang Y, Huang Y, Yu B, Weng L, Liu J. Deep reinforcement learning based optimal scheduling of active distribution system considering distributed generation, energy storage and flexible load. *Energy*. 2023;271(4):127087. doi:10.1016/j.energy.2023.127087.
50. Cui Y, Xu Y, Li Y, Wang Y, Zou X. Deep reinforcement learning based optimal energy management of multi-energy microgrids with uncertainties. *CSEE J Power Energy Syst*. 2026;12(2):1002–14.
51. Wilk P, Wang N, Li J. Multi-agent reinforcement learning for smart community energy management. *Energies*. 2024;17(20):5211. doi:10.3390/en17205211.
52. Lee S, Choi DH. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. *IEEE Trans Industr Inform*. 2022;18(1):488–97.
53. Ye Y, Tang Y, Wang H, Zhang XP, Strbac G. A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading. *IEEE Trans Smart Grid*. 2021;12(6):5185–200. doi:10.1109/tsg.2021.3103917.
54. Zhou X, Lin W, Kumar R, Cui P, Ma Z. A data-driven strategy using long short term memory models and reinforcement learning to predict building electricity consumption. *Appl Energy*. 2022;306:118078.
55. Nassereddine K, Turzynski M, Biellokha H, Strzelecki R. Simulation of energy management system using model predictive control in AC/DC microgrid. *Sci Rep*. 2025;15(1):5388. doi:10.1038/s41598-025-89036-7.
56. Wang Z, Bapst V, Heess N, Murdock T, Pritzel A, de Freitas N, et al. Sample efficient actor-critic with experience replay. *arXiv:1611.01224*. 2017.
57. Ruelens F, Claessens BJ, Vandael S, De Schutter B, Babuska R, Belmans R. Reinforcement learning applied to an electric water heater: from theory to practice. *IEEE Trans Smart Grid*. 2017;9(4):3792–800.
58. Zhang L, Yang F, Yan D, Qian G, Li J, Shi X, et al. Multi-agent deep reinforcement learning-based distributed voltage control of flexible distribution networks with soft open points. *Energies*. 2024;17(21):5244.
59. Li M, Mour N, Smith L. Machine learning based on reinforcement learning for smart grids: predictive analytics in renewable energy management. *Sustain Cities Soc*. 2024;109:105510.
60. Engstrom L, Ilyas A, Santurkar S, Tsipras D, Janoos F, Rudolph L, et al. Implementation matters in deep policy gradients: a case study on PPO and TRPO. *arXiv:2005.12729*. 2020.
61. Metwly MY, Ahmed M, Hamad MS, Abdel-Khalik AS, Hamdan E, Elmalhy NA. Power management optimization of electric vehicles for grid frequency regulation: comparative study. *Alex Eng J*. 2023;65(5):749–60. doi:10.1016/j.aej.2022.10.030.
62. Parag Y, Sovacool BK. Electricity market design for the prosumer era. *Nat Energy*. 2016;1:16032.
63. Inaolaji A, Wu X, Roychowdhury R, Smith R. Optimal allocation of battery energy storage systems for peak shaving and reliability enhancement in distribution systems. *J Energy Storage*. 2024;95:112305.

64. Kaewdornhan N, Srithapon C, Liemthong R, Chatthaworn R. Real-time multi-home energy management with EV charging scheduling using multi-agent deep reinforcement learning optimization. *Energies*. 2023;16(5):2357. doi:10.3390/en16052357.
65. Zhang B, Hu W, Cao D, Huang Q, Chen Z, Blaabjerg F. Deep reinforcement learning-based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy. *Energy Convers Manag*. 2019;202(6321):112199. doi:10.1016/j.enconman.2019.112199.
66. Gautam M. Deep reinforcement learning for resilient power and energy systems: progress, prospects, and future avenues. *Electricity*. 2023;4(4):336–80. doi:10.3390/electricity4040020.
67. Latoń D, Grela J, Ożadowicz A. Applications of deep reinforcement learning for home energy management systems: a review. *Energies*. 2024;17(24):6420. doi:10.20944/preprints202411.0814.v1.