

ProHome: Recurrent proximal policy optimization-driven advanced home energy management

Mahmoud Sallam^{a,*}, Kuljeet Kaur^{a,b,c}

^a Department of Electrical Engineering, École de Technologie Supérieure, Montreal, H3C 1K3, QC, Canada

^b Department of Electrical Engineering, Canadian University Dubai, Dubai, United Arab Emirates

^c Centre for Research Impact and Outcome, Chitkara University, Rajpura, 140401, Punjab, India

ARTICLE INFO

Keywords:

Appliance scheduling
Deep reinforcement learning
Demand response
Dynamic pricing
Electricity trading
Home energy management system
Recurrent proximal policy optimization

ABSTRACT

Advanced Home Energy Management Systems (HEMS) are essential for integrating Renewable Energy Sources (RES) and Electric Vehicles (EVs) into smart homes; however, many existing solutions overlook the temporal dependencies in energy dynamics and are evaluated under overly simplified or fully known price signals. This work proposes ProHome, a HEMS framework based on Recurrent Proximal Policy Optimization (RPPO), where a Long Short-Term Memory (LSTM)-enabled policy learns memory-dependent control strategies from sequential observations. ProHome coordinates a comprehensive set of household assets—namely, heating, ventilation, and air conditioning (HVAC), electric water heater, deferrable appliances, photovoltaic (PV) generation, an energy storage system, and an EV with bidirectional power exchange—under time-varying electricity prices without assuming access to future prices. A realistic smart home simulation environment driven by historical data was developed for training and testing. Evaluated over a comprehensive 1089-day dataset (roughly 3 years with a 75/25 train-test split), ProHome performs competitively against Deep Reinforcement Learning (DRL) baselines. While operating costs are comparable to standard Proximal Policy Optimization (PPO) (reducing the mean daily cost by a modest 1%), the proposed recurrent agent provides more reliable EV readiness by increasing the state of energy at departure by 2.1% toward the user-defined target, all while maintaining near-zero comfort violations for HVAC and hot-water temperature.

1. Introduction

Residential electricity systems are undergoing rapid change due to increasing penetration of distributed renewable energy sources (RES) and electrified loads such as electric vehicles (EVs). Under dynamic and time-varying electricity pricing, households can reduce costs by shifting flexible consumption, coordinating local generation and storage, and trading with the grid. However, practical operation is challenging due to RES intermittency, uncertain EV availability and charging needs, and the need to coordinate heterogeneous devices (thermal loads, deferrable appliances, storage) under volatile price signals [1,2].

Home Energy Management Systems (HEMS) aim to automate these decisions by scheduling controllable loads and storage while respecting operational constraints and user preferences. Classical optimization-based approaches have been widely studied [3,4], but their performance can degrade when key inputs are uncertain and when repeated re-optimization is required. More recently, Deep Reinforcement Learning

(DRL) has been increasingly explored as a model-free alternative that can learn adaptive policies through interaction with the environment and can react online to observed system conditions [5,6].

However, important knowledge gaps remain that limit whole-home, real-time DRL control under realistic constraints. These bottlenecks include: (i) multi-agent and hybrid DRL approaches often introduce training non-stationarity, credit assignment conflicts, and coordination overheads, which complicate deployment at scale [7–11]; (ii) continuous-only DRL (e.g., Soft Actor-Critic (SAC)) requires heuristic rounding to manage binary appliances, which can violate non-interruptibility and disrupt optimization gradients, while discrete-only DRL reduces control precision for thermal devices [12,13]; and (iii) fragmented or tabular frameworks suffer from memory scaling issues and stability problems in whole-home settings [14]. ProHome uniquely fills these gaps by integrating composite action representations (exact appliance on/off decisions without rounding errors, combined with precise continuous control signals) and Long Short-Term Memory (LSTM)-based

* Corresponding author.

Email addresses: mahmoud.sallam@mail.mcgill.ca (M. Sallam), kuljeet.kaur@etsmtl.ca (K. Kaur).

recurrence (helping the agent avoid short-sighted decisions despite having no access to future information) within a unified, stable single-agent framework that avoids both multi-agent coordination overhead and the memory-scaling, and stability issues of fragmented or tabular architectures.

To address these gaps, the objective of this work is to develop and evaluate a forecast-free DRL-based HEMS controller that (i) explicitly accounts for non-anticipative control settings, (ii) supports mixed discrete/continuous control with non-interruptible deferrable loads, and (iii) is assessed using historical, time-varying price trajectories.

To bridge the above gaps, this paper proposes ProHome, a comprehensive HEMS framework built around a recurrent variant of PPO (RPPO). This journal paper extends our earlier conference version [15] by introducing recurrent (LSTM-based) policies to address the lack of forecast information, using *daily-varying* real-world price profiles (instead of a repeated daily curve), and incorporating perfect-information optimization (PIO) benchmarking for stronger validation.

The contributions of this work are:

- A forecast-free recurrent control formulation for HEMS using an LSTM-based RPPO policy that operates without access to future prices.
- A unified composite (mixed) action-space design that combines continuous control (HVAC, ESS, EV, EWH) with discrete start decisions for non-interruptible deferrable loads, together with an environment implementation that enforces appliance execution.
- A fully specified reward design (including weights and deferrable-load waiting penalties) to support reproducibility and interpretation of learned behaviors.
- An evaluation against multiple baselines under the same data and constraints, including PPO, SAC, and a perfect-information MILP benchmark.

1.1. Organization

The remainder of this paper is structured as follows. [Section 2](#) presents a dedicated literature review. [Section 3](#) provides the system model and problem formulation for ProHome. [Section 4](#) introduces the Markov Decision Process (MDP) framework underlying ProHome. [Section 5](#) presents the RPPO-based agent design. The experimental setup and results are discussed in [Section 6](#), followed by conclusions in [Section 7](#).

2. Literature review

Research on HEMS for smart homes and smart grids has explored various methodologies, broadly classifiable as model-based and model-free approaches. Model-based methods construct explicit mathematical models of the system and environment to derive optimal control strategies, whereas model-free methods learn policies through direct interaction with the environment.

2.1. Model-based approaches

Traditional HEMS often rely on model-based optimization techniques. These formulate the energy management task as an optimization problem, typically aiming to minimize energy costs or peak load subject to operational constraints and user preferences, often for day-ahead scheduling. Common techniques include linear programming [16], Mixed-Integer Programming (MIP) [3,4], and convex programming [17]. For instance, [3] utilized MIP via the CPLEX solver for scheduling four appliances, while [4] targeted cost reduction, peak shaving, and efficiency using a similar approach. Nature-inspired heuristic and metaheuristic algorithms, such as genetic algorithms [18], particle swarm optimization (for EV charging) [19], and grey wolf optimization [20], have also been employed to find near-optimal solutions, particularly for complex, non-convex problems.

While offering theoretical robustness under ideal conditions, model-based methods face significant practical limitations. Their performance heavily depends on the accuracy of forecasts for uncertain variables like electricity prices, RES generation (e.g., Photovoltaic (PV) output), and load demands [21]. Obtaining accurate, long-term forecasts is notoriously difficult. Furthermore, these methods often struggle with the inherent randomness of user behavior and unscheduled appliance usage. Scalability can be an issue as the number of appliances and system complexity increases, leading to high computational burdens [20,22,23]. The need for iterative recomputation when conditions change makes many traditional optimization methods unsuitable for real-time adaptive control [7].

2.2. Model-free approaches: reinforcement learning

To overcome the limitations of model-based techniques, particularly the reliance on explicit models and forecasts, model-free approaches based on RL have gained significant momentum. RL agents learn optimal decision-making policies by interacting with the environment and receiving feedback in the form of rewards, making them inherently adaptive to dynamic and uncertain conditions.

Early applications utilized tabular RL methods like Q-learning [5] for tasks such as real-time scheduling of basic appliance categories. However, the reliance on a Q-table restricts Q-learning's applicability to problems with small, discrete state-action spaces, rendering it inefficient for complex HEMS involving numerous appliances and continuous variables (such as battery charge levels or HVAC power).

2.3. Deep reinforcement learning in HEMS

DRL overcomes the limitations of tabular RL by employing Deep Neural Networks (DNNs) as function approximators for the policy and/or value functions, enabling operation in large, continuous state-action spaces. Recent surveys confirm the rapid expansion of DRL-based HEMS across demand response, load scheduling, and RES/EV integration [12]. Various DRL algorithms have been applied to HEMS:

- **Value-Based Methods (DQN):** Deep Q-Networks (DQN) have been used, for example, to optimize community microgrids involving ESS and HVAC control [23] and to enforce operational constraints in systems with renewables using MIP-DQN formulations [24]. Multi-agent DQN has also been explored for coordinating multiple devices or homes [9]. While effective, DQN-based approaches are primarily suited for discrete action spaces and can struggle with the continuous control required for devices like ESS, EVs, and HVAC. Many implementations also rely on simplified system models or static pricing.
- **Actor-Critic Methods (DDPG, SAC, PPO, MAPPO):** Actor-critic algorithms are well-suited for continuous control problems. DDPG was adopted in [25] for controlling a residential system with a modulating heat pump, PV, and storage, and was benchmarked against MPC and rule-based control. SAC has been applied to facilitate inter-household electricity trading [7] and optimize appliance scheduling considering user satisfaction [26]. PPO was utilized in [21] for optimizing power management considering ESS and EVs under varying conditions. Multi-Agent PPO (MAPPO) has been explored for scalable control; for example, [8] applied multi-agent PPO for demand-side scheduling, and [10] proposed a MAPPO-based framework with imitation learning to improve convergence and to enhance PV self-consumption while maintaining thermal comfort in building energy systems.
- **Hierarchical and Hybrid DRL Frameworks:** Hierarchical DRL frameworks have also been proposed, for instance, using SAC or other algorithms to manage different layers of control (e.g., appliance scheduling vs. ESS/EV control) [27].

A complementary direction focuses on bridging the simulation-to-reality gap and incorporating explicit safety/constraint handling. For

Table 1

Illustrative comparison between the proposed ProHome and existing state-of-the-art HEM-based solutions.

Reference	Deferrable Loads	Fixed Loads	EWB	ESS	EV	HVAC	PV	Real-time pricing	DRL model	Contribution
[23]	✓	X	X	✓	X	AC	X	X	DQN	Cost reduction and user satisfaction for a single household.
[7]	✓	X	X	✓	✓	✓	✓	X	SAC	Detailed electricity trading scheme among households and an operator to optimize profit.
[8]	✓	✓	X	X	X	X	X	X	Multi-Agent PPO	Cost and a trading scheme among households to optimize profit.
[24]	X	✓	X	✓	X	X	✓	X	MIP-DQN	MIP-DQN formulation, optimal scheduling of renewable-based distributed energy resources (DERs).
[9]	✓	✓	X	X	X	X	✓	X	Multi-agent DQN	Tunable preference vector to avoid retraining, cost reduction and user satisfaction.
[26]	✓	X	✓	X	✓	AC	X	X	SAC	Cost reduction and user satisfaction.
[27]	✓	✓	X	✓	✓	AC	✓	X	SAC	Cost reduction and user satisfaction.
[21]	✓	✓	X	✓	✓	AC	✓	X	PPO	Reactive power optimization, cost reduction and user satisfaction.
[29]	✓	✓	X	X	✓	X	✓	✓	LSTM + PPO	Utilizing LSTM for pricing feature extraction. Optimize appliance scheduling under real-time pricing.
[25]	X	X	✓	✓	X	✓	✓	X	DDPG	HEMS control for a PV-equipped home with a modulating heat pump and storage; benchmarked against MPC and rule-based control.
[10]	X	X	X	✓	X	✓	✓	X	MAPPO (PPO-based)	Scalable multi-agent energy management with imitation learning; improved PV self-consumption and maintained thermal comfort in building energy systems.
[28]	X	X	X	✓	✓	X	✓	X	Safe RL / MPC / DT	Real-world validation comparing safe RL, MPC, and decision-tree control on residential emulators with PV, battery, and controllable EV charging.
ProHome	✓	✓	✓	✓	✓	✓	✓	✓	Recurrent PPO	Consideration of a real-time daily price curve, integration of LSTM with PPO (RPPO), comprehensive range of appliances and loads, cost reduction and user satisfaction.

example, Ruddick et al. conducted real-world experiments comparing safe RL, MPC, and decision-tree-based control on multiple residential emulators equipped with PV, battery storage, and controllable EV charging [28].

2.4. Limitations of existing DRL-based HEMS

Despite their promise, current DRL approaches for HEMS often exhibit limitations hindering practical deployment. Many rely on overly simplified environmental models, particularly static or predictable electricity pricing, failing to capture real-world market volatility. Furthermore, the prevalent use of standard MLP architectures neglects crucial temporal dependencies in energy usage, prices, and system states, which recurrent architectures like LSTMs can model more effectively. Additionally, the scope is frequently restricted to a narrow subset of appliances, overlooking the complex interplay within a complete household energy ecosystem. Finally, simplified reward functions may not adequately balance conflicting objectives like cost, comfort, and grid stability.

The work closest to the ProHome approach is [26], which combined PPO with LSTM but suffered from a limited appliance set, simplified models, and used LSTM primarily for pre-processing price data rather than integrating it into a recurrent policy network operating on the full state [26]. Table 1 summarizes and contrasts the features of relevant HEMS studies, highlighting the specific gaps ProHome aims to fill.

3. ProHome system model and problem formulation

The smart home environment considered in this study, illustrated in Fig. 1, is equipped with the ProHome HEMS and interfaces with the utility grid through a smart meter. The system architecture comprises three domains: infrastructure, communication, and management.

The infrastructure domain contains the physical energy assets, including PV generation, an energy storage system (ESS), and household electrical loads (including an EV). The communication domain, enabled by the smart meter, supports bidirectional data exchange with the utility (e.g., receiving real-time prices and transmitting metering information). The management domain is implemented by the ProHome agent, which coordinates controllable assets based on the system state, user preferences, and grid signals to optimize energy operations.

Electrical loads within the home are categorized and modeled as follows:

1. **Fixed loads:** Devices with predefined operating patterns that are not directly controlled by the HEMS. This category includes the refrigerator, microwave, oven, kettle, TV, computer, lighting, vacuum cleaner, toaster, iron, security cameras, and water pump. These loads contribute to the aggregate demand but are not scheduled by ProHome.
2. **Deferrable loads:** Appliances that require a fixed duration of uninterrupted operation within a flexible time window. ProHome schedules their start times to optimize the selected objectives (e.g., cost). This category includes the dishwasher, washing machine, and clothes dryer.
3. **Modulatable/interruptible loads:** Devices whose power can be continuously adjusted and/or whose operation can be interrupted and resumed, subject to device constraints. This category includes:
 - Heating, Ventilation, and Air Conditioning (HVAC) system: Controlled to maintain the indoor temperature within user-defined comfort bounds.
 - Electric Water Heater (EWH): Controlled to maintain the water temperature within acceptable bounds.
 - Electric Vehicle (EV): Modeled as a flexible load and (optionally) an additional storage resource. Its charging (and potential discharging) power is controlled by the HEMS to manage energy costs, satisfy travel readiness, and potentially provide grid support.

These loads provide flexibility for demand response and energy optimization.

The ESS (distinct from the EV) is also controlled by the HEMS, providing dedicated storage for energy arbitrage and self-consumption. Coordinated management of deferrable loads, modulatable/interruptible loads, and the ESS enables efficient energy utilization and grid interaction.

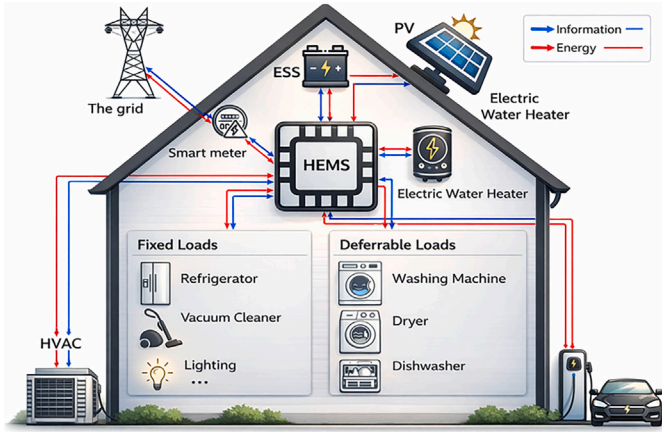


Fig. 1. A typical HEMS integrating different loads and power generation sources.

The operational dynamics and constraints of the key controllable components managed by ProHome are detailed in the following subsections.

3.1. HVAC system model

The HVAC system's thermal behavior and operational constraints are modeled based on principles adapted from [7,30]. The primary goal is to maintain the indoor air temperature T^{in} within a user-defined comfort range.

The evolution of the indoor temperature over a discrete time step Δt is governed by the thermal dynamics equation. We start from the standard first-order RC thermal balance model:

$$\frac{dT^{\text{in}}}{dt} = -\frac{1}{C^{\text{HVAC}} R^{\text{HVAC}}} (T^{\text{in}} - T^{\text{out}} + \eta^{\text{HVAC}} R^{\text{HVAC}} p^{\text{HVAC}}) \quad (1)$$

and apply an explicit Euler discretization with time step Δt to obtain

$$T_{t+1}^{\text{in}} = T_t^{\text{in}} - \frac{\Delta t (T_t^{\text{in}} - T_t^{\text{out}} + \eta^{\text{HVAC}} R^{\text{HVAC}} p_t^{\text{HVAC}})}{C^{\text{HVAC}} R^{\text{HVAC}}} \quad (2)$$

Eq. (2) describes the indoor temperature evolution as a function of the indoor-outdoor temperature difference and the thermal energy added/removed by the HVAC. Eqs. (2)–(3d) follow the first-order RC building thermal model used in HVAC flexibility studies [7,30]. Here, T_t^{in} and T_{t+1}^{in} denote the indoor temperatures at time steps t and $t+1$, respectively, and T_t^{out} is the outdoor temperature at time t . The parameters C^{HVAC} and R^{HVAC} represent the building thermal capacitance and thermal resistance. The term η^{HVAC} denotes the HVAC efficiency (COP), and p_t^{HVAC} is the signed HVAC power command (positive for cooling and negative for heating).

The operation of the HVAC is subject to several constraints: The system can only operate in one mode (heating or cooling) or be idle at any given time step:

$$b_{\text{heat},t}^{\text{HVAC}} + b_{\text{cool},t}^{\text{HVAC}} \leq 1 \quad (3a)$$

where $b_{\text{heat},t}^{\text{HVAC}}$ and $b_{\text{cool},t}^{\text{HVAC}}$ are binary variables indicating operation in heating mode (1 if active, 0 otherwise) and cooling mode (1 if active, 0 otherwise), respectively.

The electrical power consumed in heating mode, $p_{\text{heat},t}^{\text{HVAC}}$, and cooling mode, $p_{\text{cool},t}^{\text{HVAC}}$, is limited by the maximum power capacity of the system, $p_{\text{max}}^{\text{HVAC}}$:

$$0 \leq p_{\text{heat},t}^{\text{HVAC}} \leq b_{\text{heat},t}^{\text{HVAC}} p_{\text{max}}^{\text{HVAC}} \quad (3b)$$

$$0 \leq p_{\text{cool},t}^{\text{HVAC}} \leq b_{\text{cool},t}^{\text{HVAC}} p_{\text{max}}^{\text{HVAC}} \quad (3c)$$

These variables represent the magnitude of power consumed when the respective mode is active.

The net thermal power p_t^{HVAC} in Eq. (2), representing heat removal (cooling) or heat injection (heating), is related to the electrical power consumption (assuming η^{HVAC} accounts for the conversion):

$$p_t^{\text{HVAC}} = p_{\text{cool},t}^{\text{HVAC}} - p_{\text{heat},t}^{\text{HVAC}} \quad (3d)$$

Here, we adopt the convention that cooling power ($p_{\text{cool},t}^{\text{HVAC}}$) contributes positively to p_t^{HVAC} and heating power ($p_{\text{heat},t}^{\text{HVAC}}$) contributes negatively. With this sign convention, a positive p_t^{HVAC} decreases T^{in} (cooling) and a negative p_t^{HVAC} increases T^{in} (heating), consistent with (2). Note: The actual electrical power drawn from the grid would be $p_{\text{cool},t}^{\text{HVAC}}$ or $p_{\text{heat},t}^{\text{HVAC}}$ depending on the active mode. The HEMS controls these power levels.

Finally, the indoor temperature must be maintained within the user-specified comfort bounds:

$$T_{\min}^{\text{HVAC}} \leq T_t^{\text{in}} \leq T_{\max}^{\text{HVAC}} \quad (3e)$$

where $T_{\min}^{\text{HVAC}}$ and $T_{\max}^{\text{HVAC}}$ define the lower and upper limits of the desired temperature range, respectively.

The ProHome HEMS controls the operational mode ($b_{\text{heat},t}^{\text{HVAC}}, b_{\text{cool},t}^{\text{HVAC}}$) and the power levels ($p_{\text{heat},t}^{\text{HVAC}}, p_{\text{cool},t}^{\text{HVAC}}$) subject to constraints (3) to satisfy the comfort requirement (3e) while optimizing broader objectives like energy cost.

3.2. Electric water heater (EWH) model

The thermal behavior of the Electric Water Heater (EWH) tank is modeled following approaches in [26,31,32], capturing heat loss to the surroundings, energy added by the heating element, and the effects of hot water usage.

The water temperature inside the EWH tank at the next time step, T_{t+1}^{EWH} , is determined by the current temperature T_t^{EWH} , ambient temperature T_t^{amb} , inlet tap water temperature T^{tap} , hot water draw rate F_t , and the electrical power consumed by the heating element p_t^{EWH} :

$$T_{t+1}^{\text{EWH}} = \beta T_t^{\text{EWH}} + (1 - \beta) (GT_t^{\text{amb}} + k_t T^{\text{tap}} + Q_t) R' \quad (4)$$

This formulation is consistent with the aggregate EWH thermal models commonly used in residential demand-response studies [26,31,32]. In this study, the ambient temperature T_t^{amb} affecting the EWH is assumed to be equal to the indoor air temperature T_t^{in} . This assumption is based on the common placement of water heaters within the conditioned space of a home (e.g., in a utility closet or basement), making the indoor temperature the most direct environmental influence.

The parameters in Eq. (4) are defined as follows: The inertia factor β depends on the time step Δt and the tank's thermal characteristics:

$$\beta = e^{-\frac{\Delta t}{Z R'}} \quad (5a)$$

where Z represents the thermal capacitance of the water in the tank:

$$Z = C_p V \rho_{\text{water}} \quad (5b)$$

Here, C_p is the specific heat capacity of water, V is the tank volume, and ρ_{water} is the density of water.

The term G relates to the heat loss rate to the ambient environment:

$$G = \frac{SA}{TR} \quad (5c)$$

where SA is the surface area of the tank and TR is its insulation thermal resistance.

The term k_t represents the effect of water flow:

$$k_t = C_p F_t \rho_{\text{water}} \quad (5d)$$

where F_t is the volumetric hot water flow rate (usage) at time t .

The term Q_t represents the thermal input from the heating element, related to the electrical power consumption p_t^{EWH} :

$$Q_t = 3600 p_t^{\text{EWH}} \quad (5e)$$

The factor 3600 converts kW to kJ/h. In the implemented update Eq. (4), the effect of the time step is captured by β (via the exponential term), so Δt does not appear explicitly in Q_t .

The equivalent thermal resistance R' is defined as:

$$R' = \frac{1}{G + k_t} \quad (5f)$$

The operation of the EWH is subject to constraints on its power consumption and the resulting water temperature: The electrical power consumption p_t^{EWH} is limited by the maximum rated power of the heating element, $p_{\text{max}}^{\text{EWH}}$:

$$0 \leq p_t^{\text{EWH}} \leq p_{\text{max}}^{\text{EWH}} \quad (6a)$$

The water temperature must be maintained within the user-specified acceptable range:

$$T_{\text{min}}^{\text{EWH}} \leq T_t^{\text{EWH}} \leq T_{\text{max}}^{\text{EWH}} \quad (6b)$$

where $T_{\text{min}}^{\text{EWH}}$ and $T_{\text{max}}^{\text{EWH}}$ are the lower and upper temperature bounds, respectively, ensuring user comfort.

The ProHome HEMS controls the electrical power consumption p_t^{EWH} at each time step, respecting the constraints (6), to meet the temperature requirements while contributing to overall energy optimization goals.

3.3. Energy storage system (ESS) model

The dedicated Energy Storage System (ESS), typically a battery, plays a crucial role in enhancing energy efficiency by storing surplus energy (e.g., from PV generation) for later use, reducing reliance on the grid during high-price periods, and mitigating peak demand. Its operation is modeled based on [7,21].

The state of energy (SoE) stored in the ESS at the next time step, E_{t+1}^{ESS} , is determined by the current state E_t^{ESS} , the charging power $p_{c,t}^{\text{ESS}}$, the discharging power $p_{d,t}^{\text{ESS}}$, and the respective efficiencies η_c^{ESS} and η_d^{ESS} over the time interval Δt :

$$E_{t+1}^{\text{ESS}} = E_t^{\text{ESS}} + p_{c,t}^{\text{ESS}} \Delta t \eta_c^{\text{ESS}} - \frac{p_{d,t}^{\text{ESS}} \Delta t}{\eta_d^{\text{ESS}}} \quad (7)$$

This representation of ESS SoE dynamics and operational constraints is standard in HEMS battery storage modeling [7,21]. Here, $\eta_c^{\text{ESS}} \in (0, 1]$ is the charging efficiency (energy stored per unit energy drawn from the source) and $\eta_d^{\text{ESS}} \in (0, 1]$ is the discharging efficiency (energy delivered per unit energy drawn from storage).

The ESS operation is subject to the following constraints: The ESS cannot simultaneously charge and discharge:

$$b_{c,t}^{\text{ESS}} + b_{d,t}^{\text{ESS}} \leq 1 \quad (8a)$$

where $b_{c,t}^{\text{ESS}}$ and $b_{d,t}^{\text{ESS}}$ are binary variables indicating charging (1 if active, 0 otherwise) and discharging (1 if active, 0 otherwise), respectively.

The charging and discharging power magnitudes are limited by the ESS's maximum power rating, $p_{\text{max}}^{\text{ESS}}$:

$$0 \leq p_{c,t}^{\text{ESS}} \leq b_{c,t}^{\text{ESS}} p_{\text{max}}^{\text{ESS}} \quad (8b)$$

$$0 \leq p_{d,t}^{\text{ESS}} \leq b_{d,t}^{\text{ESS}} p_{\text{max}}^{\text{ESS}} \quad (8c)$$

These define the rates at which energy can be transferred into or out of the ESS.

The net power exchanged between the ESS and the rest of the home system at time t , denoted as p_t^{ESS} , is defined as:

$$p_t^{\text{ESS}} = p_{d,t}^{\text{ESS}} - p_{c,t}^{\text{ESS}} \quad (8d)$$

Following the convention that power drawn from the system (charging) is negative and power supplied to the system (discharging) is positive. Note that p_t^{ESS} represents the power flow at the ESS terminals *within* the home; the actual power drawn from/supplied to the grid also accounts for efficiencies.

The stored energy level must remain within the operational bounds of the ESS:

$$E_{\text{min}}^{\text{ESS}} \leq E_t^{\text{ESS}} \leq E_{\text{max}}^{\text{ESS}} \quad (8e)$$

where $E_{\text{min}}^{\text{ESS}}$ and $E_{\text{max}}^{\text{ESS}}$ represent the minimum and maximum allowable energy levels (often related to depth-of-discharge limits and total capacity), respectively.

ProHome controls the charging and discharging power levels, $p_{c,t}^{\text{ESS}}$ and $p_{d,t}^{\text{ESS}}$ (and implicitly the modes $b_{c,t}^{\text{ESS}}$, $b_{d,t}^{\text{ESS}}$), subject to constraints (8), to optimize overall system objectives based on energy prices, PV generation, load demand, and the current ESS state E_t^{ESS} .

3.4. Electric vehicle (EV) model

The EV serves a dual role in the ProHome system: it is a significant flexible load when charging and can act as an auxiliary energy storage resource, potentially enabling Vehicle-to-Grid (V2G) interactions. Its operational model is adapted from [7,21].

Similar to the ESS, the EV's state of energy (SoE) at time $t + 1$, E_{t+1}^{EV} , evolves based on its current state E_t^{EV} and charging/discharging activities:

$$E_{t+1}^{\text{EV}} = E_t^{\text{EV}} + p_{c,t}^{\text{EV}} \Delta t \eta_c^{\text{EV}} - \frac{p_{d,t}^{\text{EV}} \Delta t}{\eta_d^{\text{EV}}} \quad (9)$$

The same modeling approach is applied to the EV SoE dynamics and constraints, consistent with prior residential energy management studies [7,21], where $p_{c,t}^{\text{EV}}$ and $p_{d,t}^{\text{EV}}$ are the charging and discharging powers, respectively, and $\eta_c^{\text{EV}} \in (0, 1]$, $\eta_d^{\text{EV}} \in (0, 1]$ are the corresponding efficiencies. Δt is the time step duration.

The EV's operation is subject to several constraints, reflecting its usage pattern and physical limits: Charging and discharging cannot occur simultaneously:

$$b_{c,t}^{\text{EV}} + b_{d,t}^{\text{EV}} \leq 1 \quad (10a)$$

where $b_{c,t}^{\text{EV}}$ and $b_{d,t}^{\text{EV}}$ are binary variables indicating the charging and discharging modes, respectively (1 if active, 0 otherwise).

The charging and discharging power magnitudes are limited by the EV's maximum charging/discharging rates, $p_{c,\text{max}}^{\text{EV}}$ and $p_{d,\text{max}}^{\text{EV}}$:

$$0 \leq p_{c,t}^{\text{EV}} \leq b_{c,t}^{\text{EV}} p_{c,\text{max}}^{\text{EV}} \quad (10b)$$

$$0 \leq p_{d,t}^{\text{EV}} \leq b_{d,t}^{\text{EV}} p_{d,\text{max}}^{\text{EV}} \quad (10c)$$

Note that $p_{d,\text{max}}^{\text{EV}}$ might be zero if V2G capability is not assumed or enabled.

The net power exchanged with the home system, p_t^{EV} , is:

$$p_t^{\text{EV}} = p_{d,t}^{\text{EV}} - p_{c,t}^{\text{EV}} \quad (10d)$$

(Positive for discharge, negative for charge, consistent with the ESS convention).

The EV is assumed to be unavailable for charging or discharging when it is away from home. This is modeled by setting both modes inactive during the departure period:

$$b_{c,t}^{\text{EV}} = b_{d,t}^{\text{EV}} = 0, \quad \forall t \in [t_{\text{departure}}, t_{\text{arrival}}) \quad (10e)$$

where $t_{\text{departure}}$ is the daily departure time and t_{arrival} is the return time, which may vary.

The EV's battery energy level must remain within its physical limits:

$$E_{\min}^{\text{EV}} \leq E_t^{\text{EV}} \leq E_{\max}^{\text{EV}} \quad (10f)$$

where $E_{\min}^{\text{EV}}$ and $E_{\max}^{\text{EV}}$ are the minimum and maximum energy levels.

Crucially, the EV must have sufficient charge for the anticipated travel needs by the time of departure:

$$E_t^{\text{EV}} \geq E_{\text{tr}}^{\text{EV}}, \quad \text{at } t = t_{\text{departure}} \quad (10g)$$

where $E_{\text{tr}}^{\text{EV}}$ is the target or required energy level for travel, specified by the user or estimated based on typical usage.

ProHome manages the EV's charging/discharging schedule ($p_{c,t}^{\text{EV}}, p_{d,t}^{\text{EV}}$) when the vehicle is available (plugged in at home), respecting constraints (10), particularly the travel requirement (10g), while optimizing costs and potentially utilizing the EV for grid services or self-consumption enhancement.

3.5. Deferrable appliances model

Deferrable appliances, such as dishwashers or washing machines ($\forall d \in D$, where D is the set of deferrable appliances), require a fixed duration, t_{dur}^d , of continuous operation to complete their task cycle. They offer flexibility in their start time. It is assumed that the user indicates their desired operation by setting an initial time t_{ini}^d . The ProHome HEMS then determines the actual start time t_{start}^d , where $t_{\text{start}}^d \geq t_{\text{ini}}^d$. The appliance effectively waits for the period $t_{\text{start}}^d - t_{\text{ini}}^d$ before commencing operation. Once started, the operation cannot be interrupted.

The operational state of appliance d at time t is represented by a binary variable $o_d(t)$:

$$o_d(t) = \begin{cases} 1, & \text{if } t \in [t_{\text{start}}^d, t_{\text{start}}^d + t_{\text{dur}}^d) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

The start time t_{start}^d must be chosen such that the operation finishes within the scheduling horizon (e.g., by the end of the day, t_{end}), i.e., $t_{\text{start}}^d + t_{\text{dur}}^d \leq t_{\text{end}}$.

The power consumption of appliance d at time t , p_t^d , is determined by its operational state and its rated power P_t^d :

$$p_t^d = o_d(t) \cdot P_t^d \quad (12)$$

To track the completion status, a variable B_t^d representing the fraction of the task completed by time t can be defined:

$$B_t^d = \frac{1}{t_{\text{dur}}^d} \sum_{k=0}^t o_d(k) \cdot \Delta t \quad (13)$$

The constraint that the task must be completed ensures that $B_{t_{\text{end}}}^d = 1$.

ProHome's control action for each deferrable appliance d is to select the optimal start time t_{start}^d that satisfies user constraints and minimizes overall costs or achieves other performance objectives.

3.6. Fixed loads model

Fixed loads encompass appliances like refrigerators, lighting, or entertainment systems whose operation is primarily dictated by user needs or internal controls, and thus are not actively scheduled by the ProHome HEMS. Their operation patterns are considered exogenous inputs to the system. The set of fixed load appliances is denoted by F .

Although not controlled, their power consumption must be accounted for in the overall energy balance. The operational state of a fixed load $f \in F$ at time t can be represented by a binary variable b_t^f :

$$b_t^f = \begin{cases} 1, & \text{if appliance } f \text{ is operating at time } t \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

The start times t_{start}^f and durations t_{dur}^f (or more generally, the entire operational schedule represented by b_t^f for all t) are assumed to follow typical, potentially random or pre-defined, usage patterns.

The power consumption of fixed load f at time t , p_t^f , is given by its rated power P_t^f when active:

$$p_t^f = b_t^f \cdot P_t^f \quad (15)$$

The total power consumption from all fixed loads at time t is $P_t^F = \sum_{f \in F} P_t^f$. This consumption contributes to the overall household demand and affects the net grid exchange, but it is not directly controlled by the HEMS.

3.7. Net power consumption and cost calculation

The total net electrical power consumption of the household at time t , P_t^{total} , represents the balance between consumption by all loads/storage and generation from PV panels. It determines the power exchanged with the utility grid:

$$P_t^{\text{total}} = p_{\text{elec},t}^{\text{HVAC}} + p_t^{\text{ESS}} + p_t^{\text{EV}} + p_t^{\text{EWH}} + \sum_{d \in D} p_t^d + \sum_{f \in F} p_t^f - p_t^{\text{PV}} \quad (16)$$

where:

- $p_{\text{elec},t}^{\text{HVAC}}$ is the actual electrical power consumed by the HVAC (either $p_{\text{heat},t}^{\text{HVAC}}$ or $p_{\text{cool},t}^{\text{HVAC}}$ depending on the active mode).
- p_t^{ESS} is the net power from/to the ESS (negative for charging, positive for discharging).
- p_t^{EV} is the net power from/to the EV (negative for charging, positive for discharging).
- p_t^{EWH} is the power consumed by the EWH heating element.
- $\sum_{d \in D} P_t^d$ is the total power consumed by active deferrable loads.
- $\sum_{f \in F} P_t^f$ is the total power consumed by active fixed loads.
- p_t^{PV} is the power generated by the PV system at time t .

A positive P_t^{total} signifies net power import from the grid, while a negative value indicates net power export to the grid (surplus generation).

The instantaneous electricity cost (or revenue) incurred at time t , denoted C_t , depends on the net power exchange P_t^{total} and the time-varying electricity prices:

$$C_t = \begin{cases} P_t^{\text{total}} \cdot \lambda_t^+ \cdot \Delta t, & \text{if } P_t^{\text{total}} > 0 \quad (\text{Importing}) \\ P_t^{\text{total}} \cdot \lambda_t^- \cdot \Delta t, & \text{if } P_t^{\text{total}} < 0 \quad (\text{Exporting}) \\ 0, & \text{if } P_t^{\text{total}} = 0 \end{cases} \quad (17)$$

where λ_t^+ is the electricity purchase price from the grid at time t , and λ_t^- is the selling price (feed-in tariff) for exporting surplus energy to the grid at time t . Δt converts power to energy for the time step duration. In the base case, the export price is modeled as a fixed fraction of the purchase price, $\lambda_t^- = \alpha \lambda_t^+$ with $\alpha = 0.5$, following the tariff assumption adopted from [21]. This represents a conservative self-consumption-oriented tariff in which exported energy is compensated below the retail purchase price. The sensitivity of results to α is examined in Section 6.1. The primary objective of the ProHome HEMS is to minimize the total cost over a given period, $\sum_t C_t$, while satisfying all operational and comfort constraints.

4. Markov decision process formulation

The ProHome energy management problem is formulated as a Markov Decision Process (MDP), providing a standard framework for sequential decision-making under uncertainty [33,34]. An MDP is defined by a tuple $(S, \mathcal{A}, P, R, \gamma)$, where S is the set of states, $\mathcal{A}$ is the set of actions, P is the state transition probability function, R is the reward function, and γ is the discount factor. In this context, the ProHome HEMS acts as the agent learning a policy $\pi : S \rightarrow \mathcal{A}$ to maximize the expected discounted cumulative reward. The components specific to this formulation are detailed below.

4.1. State space (S)

The state vector $s_t \in S$ summarizes the information available to the ProHome agent at time step t for decision-making. Importantly, this observation is forecast-free (no access to future price or environmental forecasts), meaning the HEMS operates under real-time constraints without lookahead. The state is defined as:

$$s_t = \{t, \{t_{\text{ini}}^d, t_{\text{start}}^d, B_t^d \forall d \in D\}, T_t^{\text{in}}, T_t^{\text{out}}, p_t^{\text{PV}}, E_t^{\text{ESS}}, E_t^{\text{EV}}, P_t^F, T_t^{\text{EWH}}, \lambda_t^+\} \quad (18)$$

4.1.1. Temporal and appliance states

Current time step within the 24-hour cycle: t

For each deferrable appliance $d \in D$ (Dishwasher, Washing Machine, Clothes Dryer):

- t_{ini}^d : User-defined preferred start time
- t_{start}^d : Actual start time (0 if not yet started)
- B_t^d : Normalized execution progress of appliance d (0 = not started; increases to 1 when completed)

Once an appliance is started (i.e., $t \geq t_{\text{start}}^d$), the environment enforces the non-interruptible constraint internally by overriding scheduling actions as needed until the task is completed ($B_t^d = 1$).

4.1.2. Environmental and energy states

Temperature and weather conditions:

- T_t^{in} : Indoor temperature
- T_t^{out} : Outdoor temperature
- p_t^{PV} : Current PV generation

Energy storage and consumption:

- E_t^{ESS} : State of Energy (SoE) of the Energy Storage System
- E_t^{EV} : State of Energy (SoE) of the Electric Vehicle
- P_t^F : Aggregate power consumption of all fixed loads ($\sum_{f \in F} p_t^f$)
- T_t^{EWH} : Water temperature in the Electric Water Heater

Market conditions:

- λ_t^+ : Electricity purchase price (selling price λ_t^- derived from λ_t^+)

The state vector provides the agent with a complete snapshot of the home's energy status, environmental conditions, and operational constraints at each decision point.

4.2. Action space (A)

The action space A defines the set of control commands issued by the ProHome agent at each time step t . The action vector $\mathbf{a}_t$ has 7 dimensions: 4 continuous controls and 3 binary start decisions for deferrable appliances:

$$\mathbf{a}_t = [a_t^{\text{HVAC}}, a_t^{\text{ESS}}, a_t^{\text{EV}}, a_t^{\text{EWH}}, a_t^{d_1}, a_t^{d_2}, a_t^{d_3}] \quad (19)$$

where indices 0–3 correspond to continuous actions (HVAC, ESS, EV, EWH) and indices 4–6 correspond to the three deferrable appliances (Dishwasher, Washing Machine, Clothes Dryer).

4.2.1. Continuous control actions

The following actions provide continuous control over power consumption and energy storage:

HVAC Control ($a_t^{\text{HVAC}} \in [-1, 1]$)

- Positive values activate cooling: $p_{\text{cool},t}^{\text{HVAC}} = a_t^{\text{HVAC}} \cdot p_{\text{max}}^{\text{HVAC}}$
- Negative values activate heating: $p_{\text{heat},t}^{\text{HVAC}} = |a_t^{\text{HVAC}}| \cdot p_{\text{max}}^{\text{HVAC}}$
- Zero indicates an idle state

Energy Storage System ($a_t^{\text{ESS}} \in [-1, 1]$)

- Positive values discharge: $p_{\text{d},t}^{\text{ESS}} = a_t^{\text{ESS}} \cdot p_{\text{max}}^{\text{ESS}}$
- Negative values charge: $p_{\text{c},t}^{\text{ESS}} = |a_t^{\text{ESS}}| \cdot p_{\text{max}}^{\text{ESS}}$

Electric Vehicle ($a_t^{\text{EV}} \in [-1, 1]$)

- Similar to ESS control for charging/discharging power ($p_{\text{c},t}^{\text{EV}}, p_{\text{d},t}^{\text{EV}}$).
- Active only when EV is available (not traveling).

Electric Water Heater ($a_t^{\text{EWH}} \in [-1, 1]$)

- Scaled to heating power: $p_t^{\text{EWH}} = \frac{a_t^{\text{EWH}} + 1}{2} \cdot p_{\text{max}}^{\text{EWH}}$
- Maps continuous action to range $[0, p_{\text{max}}^{\text{EWH}}]$

4.2.2. Discrete control actions

Deferrable Appliances ($a_t^d \in \{0, 1\}$ for each $d \in D$)

- Binary action for each deferrable appliance (Dishwasher, Washing Machine, Clothes Dryer).
- $a_t^d = 1$: Start appliance d at time t (sets $t_{\text{start}}^d = t$).
- $a_t^d = 0$: Do not start appliance d at time t .

In the proposed composite policy, each pending deferrable appliance is parameterized by a Bernoulli distribution; sampling yields a valid $\{0, 1\}$ command, and the environment enforces non-interruptible execution once started. The waiting-time penalty in Eq. (23) discourages excessive delays and induces an explicit trade-off between cost minimization and user convenience.

Collectively, these actions enable the HEMS to coordinate all controllable devices while respecting operational constraints.

4.3. Reward function (R)

The reward function $R(s_t, a_t, s_{t+1})$ provides feedback to the agent. It is designed as a weighted sum of components reflecting operational costs and user satisfaction/comfort. The immediate reward r_t received after taking action $\mathbf{a}_t$ in state s_t is calculated as follows:

4.3.1. Comfort violation penalties

Penalties are incurred if thermal comfort bounds are violated:

- HVAC Reward (r_t^{HVAC}): Penalizes deviation of indoor temperature T_{t+1}^{in} from the bounds $[T_{\text{min}}^{\text{HVAC}}, T_{\text{max}}^{\text{HVAC}}]$.

$$r_t^{\text{HVAC}} = - \left([T_{t+1}^{\text{in}} - T_{\text{max}}^{\text{HVAC}}]^+ + [T_{\text{min}}^{\text{HVAC}} - T_{t+1}^{\text{in}}]^+ \right) \quad (20)$$

where $[x]^+ = \max(x, 0)$. The reward is zero if the temperature is within bounds, otherwise it is negative proportional to the deviation.

- EWH Reward (r_t^{EWH}): Similarly penalizes deviation of water temperature T_{t+1}^{EWH} from its bounds $[T_{\text{min}}^{\text{EWH}}, T_{\text{max}}^{\text{EWH}}]$.

$$r_t^{\text{EWH}} = - \left([T_{t+1}^{\text{EWH}} - T_{\text{max}}^{\text{EWH}}]^+ + [T_{\text{min}}^{\text{EWH}} - T_{t+1}^{\text{EWH}}]^+ \right) \quad (21)$$

4.3.2. EV travel readiness reward (r_t^{EV})

A penalty is applied only at the departure time $t_{\text{departure}}$ if the EV's charge $E_{t_{\text{departure}}}^{\text{EV}}$ is below the required level $E_{\text{tr}}^{\text{EV}}$:

$$r_t^{\text{EV}} = \begin{cases} -[E_{\text{tr}}^{\text{EV}} - E_t^{\text{EV}}]^+ & \text{if } t = t_{\text{departure}} \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

4.3.3. Deferrable appliance waiting time reward (r_t^d)

For each deferrable appliance $d \in D$, a penalty accumulates for each time step it waits after being initialized by the user (t_{ini}^d) but before being started by the HEMS (t_{start}^d):

$$r_t^d = \begin{cases} -(t - t_{\text{ini}}^d) \cdot \Delta t & \text{if } t_{\text{ini}}^d \leq t < t_{\text{start}}^d \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

Because the instantaneous penalty magnitude grows with the waiting duration $(t - t_{\text{ini}}^d)$, the *cumulative* waiting penalty over multiple time steps grows approximately quadratically with total delay. We intentionally use this “progressive urgency” design to discourage the agent from deferring an appliance indefinitely for marginal cost savings, while still allowing short delays to exploit low-price periods.

(Note: The initial assignment of negative values for $t_{\text{ini}}^d, t_{\text{start}}^d$ before they are known is an implementation detail handled by the environment simulation.)

4.3.4. Total reward (r_t)

The total immediate reward r_t is a weighted combination of (i) economic performance (negative cost) and (ii) user satisfaction terms:

$$r_t = -\omega_{\text{cost}} C_t + \omega_{\text{HVAC}} r_t^{\text{HVAC}} + \omega_{\text{EWH}} r_t^{\text{EWH}} + \omega_{\text{EV}} r_t^{\text{EV}} + \sum_{d \in D} \omega_d r_t^d \quad (24)$$

Here, C_t is the electricity cost from (17), and the penalty terms are defined in (20)–(23). All comfort and travel-related variables (e.g., $T_{\text{max}}^{\text{HVAC}}$, $E_{\text{tr}}^{\text{EV}}$) are user-defined parameters specified in Section 3. The weights $\omega_{\text{cost}}, \omega_{\text{HVAC}}, \omega_{\text{EWH}}, \omega_{\text{EV}}, \omega_d$ are non-negative coefficients that balance the trade-off between minimizing cost and maintaining user satisfaction; the values used in our experiments are reported in Table 6.

5. ProHome agent: recurrent deep reinforcement learning

This section presents the ProHome control agent and its training procedure. We adopt Recurrent Proximal Policy Optimization (RPPO), i.e., a recurrent extension of PPO obtained by augmenting the policy network with Long Short-Term Memory (LSTM) [35]. Recurrence is particularly important for HEMS because the agent’s observations are forecast-free (Section 4): at time t it sees the current state (including the current price) but not future prices or future user-driven events. The LSTM hidden state therefore provides compact memory of recent dynamics (prices, SoE evolution, thermal trajectories), enabling anticipatory behavior under strictly forecast-free observations.

5.1. Proximal policy optimization (PPO) background

PPO is an actor-critic DRL algorithm known for its stability, sample efficiency, and ease of implementation compared to earlier policy gradient methods [36]. The actor network learns the policy $\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$, which maps states to a probability distribution over actions, while the critic network learns the state-value function $V_\phi(\mathbf{s}_t)$, estimating the expected cumulative reward from state $\mathbf{s}_t$ under policy π_θ .

PPO optimizes a surrogate objective function that constrains the change in the policy between updates, preventing large, destabilizing steps. The commonly used clipped objective function is:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t [\min(\delta_t(\theta) \hat{A}_t, \text{clip}(\delta_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)], \quad (25)$$

where $\mathbb{E}_t$ denotes the expectation over a batch of transitions. $\delta_t(\theta) = \frac{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta_{\text{old}}}(\mathbf{a}_t | \mathbf{s}_t)}$ is the probability ratio between the current policy π_θ and the policy before the update $\pi_{\theta_{\text{old}}}$. $\hat{A}_t$ is an estimate of the advantage function at time t , quantifying how much better the chosen action $\mathbf{a}_t$ was compared to the average action from state $\mathbf{s}_t$. ϵ is a small hyperparameter defining the clipping range (e.g., 0.1 or 0.2). The $\text{clip}(\delta_t(\theta), 1 - \epsilon, 1 + \epsilon)$ function restricts the ratio $\delta_t(\theta)$ to the interval $[1 - \epsilon, 1 + \epsilon]$, thereby limiting the policy update magnitude.

The advantage term $\hat{A}_t$ is commonly computed using Generalized Advantage Estimation (GAE) [37]. For clarity, the one-step TD form used as a building block is:

$$\hat{A}_t = r_t + \gamma V_\phi(\mathbf{s}_{t+1}) - V_\phi(\mathbf{s}_t), \quad (26)$$

where r_t is the immediate reward after applying $\mathbf{a}_t$ in $\mathbf{s}_t$, and $\gamma \in [0, 1]$ is the discount factor. In our implementation, advantage estimation is

computed consistently with PPO/RPPO training; the sign of $\hat{A}_t$ determines whether the sampled action should become more or less likely under the updated policy.

The critic network ϕ is trained by minimizing a loss function based on the squared error between its predictions $V_\phi(\mathbf{s}_t)$ and target values (e.g., the discounted return $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ or bootstrapped estimates like $r_t + \gamma V_\phi(\mathbf{s}_{t+1})$):

$$L^{\text{VF}}(\phi) = \mathbb{E}_t [(V_\phi(\mathbf{s}_t) - V_t^{\text{targ}})^2] \quad (27)$$

Typically, the PPO loss function combines the policy surrogate loss, the value function loss, and potentially an entropy bonus term to encourage exploration.

A key challenge in this work is handling the mixed action space, which contains both continuous and discrete components. In our implementation, the policy is represented using a composite action distribution (TorchRL CompositeDistribution) that factorizes the action into two parts:

- Continuous control (4 dims): A Normal distribution for $[a_t^{\text{HVAC}}, a_t^{\text{ESS}}, a_t^{\text{EV}}, a_t^{\text{EWH}}]$ parameterized by learned location and scale.
- Discrete scheduling (3 dims): A Bernoulli distribution for the three deferrable appliances, parameterized by learned logits.

At each time step, the action vector is sampled from this composite distribution. The environment receives explicitly sampled binary decisions $\{0, 1\}$ for the deferrable loads, while the continuous actions are normalized to $[-1, 1]$ (HVAC/ESS/EV) and the EWH action is mapped via $\frac{a_t^{\text{EWH}} + 1}{2}$ to the physical range $[0, p_{\text{max}}^{\text{EWH}}]$ as in Section 4.

5.2. Recurrent proximal policy optimization (RPPO)

The core advantage of RPPO over standard PPO in the HEMS context lies in its state representation. While a standard PPO agent’s policy, $\pi(\mathbf{a}_t | \mathbf{s}_t)$, is conditioned only on the current (forecast-free) state, the RPPO agent’s composite policy is parameterized as $\pi_\theta(\mathbf{a}_t | \mathbf{s}_t, h_t, c_t)$. In the LSTM actor, h_t denotes the hidden state and c_t denotes the cell state. Together, (h_t, c_t) serve as recurrent memory states updated sequentially. They do not store future lookaheads, but summarize recent observed history (e.g., price trajectories, indoor temperature gradients, and appliance execution progress). This acts as a compressed summary of the historical trajectory, providing the necessary context for effective long-term planning. This allows RPPO to outperform PPO in environments where the optimal action depends on a sequence of past events, which is precisely the case in energy management with its fluctuating prices and accumulating resource states (like battery charge).

The RPPO implementation used in ProHome employs an LSTM in the actor (policy) network, while the critic remains a feedforward MLP that estimates $V_\phi(\mathbf{s}_t)$ from the current observation.

At each environment step, the inference workflow proceeds as follows: the current observation is normalized using running statistics computed during an initial random-policy warm-up; the normalized state along with the LSTM hidden and cell states (h_t, c_t) is passed through the LSTM actor followed by the MLP layers, producing 11 output units split into 4 Normal distribution parameters (continuous actions for HVAC, ESS, EV, EWH) and 3 Bernoulli logits (deferrable appliance start decisions); actions are sampled from this composite distribution and submitted to the environment, which enforces physical bounds and non-interruptible appliance execution. The resulting transition $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}, h_t, c_t)$ is stored in the on-policy rollout buffer. When the buffer contains sufficient data, the update cycle runs K PPO epochs: contiguous sequences are sampled to preserve temporal order for BackPropagation Through Time (BPTT) through the LSTM actor, GAE advantages are computed, and both the clipped policy objective and the value-function loss are optimized via Adam with gradient clipping.

As illustrated in Fig. 2, training proceeds on-policy: the current policy interacts with the environment to generate trajectories, and the collected

Table 2

Fixed loads parameters (adopted from [21]). Duration and start time are expressed in 15-minute time-step indices.

Fixed Load	Power Rating	Duration	Start time
Refrigerator	1.66 kW	96	[0, 95]
Microwave	1.2 kW	1	[4, 7] [46, 51] [58, 62]
Oven	2.4 kW	2 6	[2, 5] [54, 60]
Kettle	2.0 kW	1	[2, 5] [46, 51] [58, 62]
TV	0.28 kW	20	[48, 54]
Computer	0.2 kW	12	[58, 62]
Lighting	0.2 kW	20	[48, 54]
Vacuum	0.6 kW	2	[44, 49]
Toaster	0.8 kW	1	[2, 7]
Iron	2.4 kW	1	[3, 6]
Security cameras	0.2 kW	96	[0, 95]
Water pump	1.2 kW	3	[69, 73]

Table 3

Deferrable loads parameters (adopted from [21]). Duration and t_{ini}^d are expressed in 15-minute time-step indices.

Shiftable Load	Power Rating	Duration	t_{ini}^d
Dishwasher	1.32 kW	4	[64, 67]
Washing Machine	1.4 kW	4	[44, 48]
Clothes dryer	3.8 kW	4	[72, 75]

Table 4

EV and ESS parameters (adopted from [7,21]).

Parameter	EV	ESS
E_{max}	16 kWh	10 kWh
E_{min}	1.6 kWh	1 kWh
E_{tr}^{EV}	70%	—
$E_{initial}$	[4 – 9] kWh	[2 – 6] kWh
$p_{c,i}$	3.3 kW	3 kW
$p_{d,i}$	3.3 kW	3 kW
$\eta_{c,i}, \eta_{d,i}$	0.95	0.95
$t_{departure}$	[11 – 14]	—
$t_{arrival}$	[45 – 50]	—

Table 5

HVAC parameters (adopted from [7,30]).

Parameter	Value	Parameter	Value
$T_{max}^{HVAC}(F)$	75.2	$T_{min}^{HVAC}(F)$	66.2
$T_{initial}^{HVAC}(F)$	[69 – 73]	$p_{HVAC}^{HVAC}(kW)$	1.75
C^{HVAC}	0.33	R^{HVAC}	13.5
η^{HVAC}	2.5	—	—

Table 6

Reward function Weights.

Parameter	Value
EV Weight (ω_{EV})	0.50
HVAC Comfort Weight (ω_{HVAC})	0.25
EWB Comfort Weight (ω_{EWB})	0.125
Cost Weight (ω_{cost})	0.10
Deferrable waiting weight (ω_d)	0.015

Table 7

EWB parameters (adopted from [26,31,32]).

Parameter	Value	Parameter	Value
$T_i^{tap}(F)$	59	SA (m^2)	2.5
V (L)	150	TR (hour. m^2 .F/kJ)	0.75
C_p (kJ/(°F. kg))	4.2	ρ (kg/L)	1
p_{max} (kW)	4.5	F_i (L/hour)	3
$T_{max}^{EWB}(F)$	131	$T_{min}^{EWB}(F)$	122
$T_{initial}^{EWB}(F)$	[122 – 131]	T_{amb}	T_i^{in}

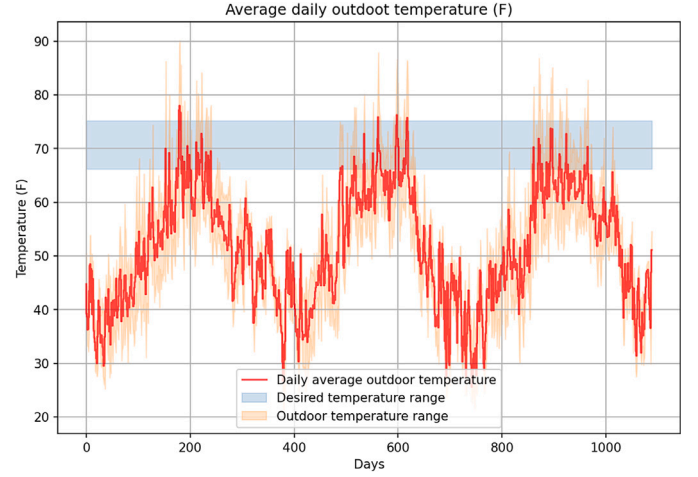


Fig. 3. Daily outdoor temperature characteristics: Average (red line) and range (Orange shaded area) across all days in the dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

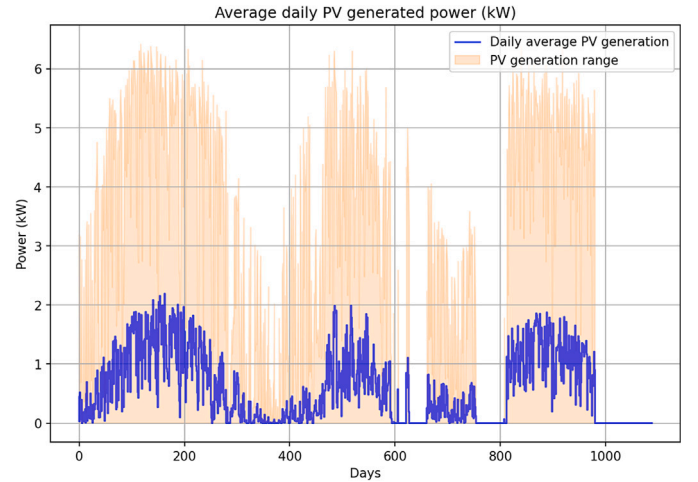


Fig. 4. Daily PV generation characteristics: Average (blue line) and range (Orange shaded area) across all days in the dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Utrecht, Netherlands (2015–2017). Temperature and PV generation data were obtained from Hugging Face [39], and electricity price data were sourced from Ember Energy [40]. All signals were aligned and resampled to 15-minute resolution ($\Delta t = 15$ min). Days with missing values were removed, resulting in 1089 complete days.

This dataset was chronologically split into a training set comprising the first 75% of the days (816 days) and a testing set containing the remaining 25% (273 days). The testing set spans various seasons across the covered years, providing diverse scenarios for evaluating the generalization performance of the trained agent.

Figs. 3–5 visualize the characteristics of the dataset. Fig. 3 shows the average daily outdoor temperature and its typical daily range. Significant seasonal variation is evident, with temperatures frequently falling outside typical indoor comfort zones, necessitating active HVAC operation.

Fig. 4 illustrates the average daily PV generation profile and its variability. The substantial fluctuations, influenced by weather and seasonality, pose a challenge for the HEMS in optimizing ESS charging/discharging and grid interaction strategies.

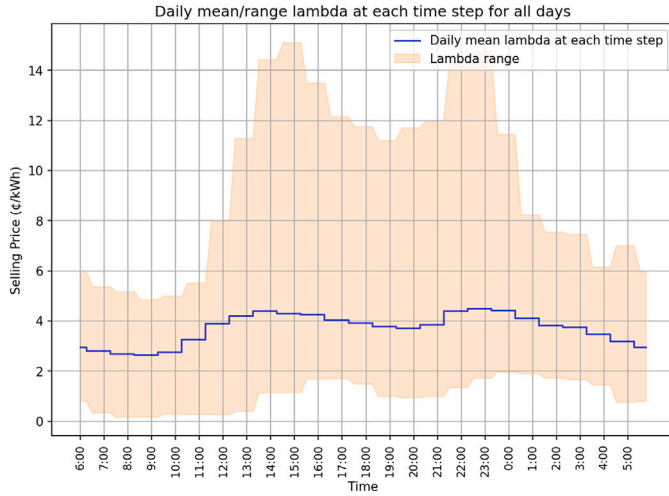


Fig. 5. Electricity purchase price characteristics: Average (blue line) and range (Orange shaded area) at each 15-minute interval across all days in the dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

A key aspect of this work is the use of realistic, day-varying electricity price trajectories (Fig. 5). The figure reports the mean and range of the purchase price at each 15-minute index across all days, highlighting substantial intra-day and inter-day volatility.

ProHome is evaluated under strict forecast-free constraints. At time step t , the agent observes only the current state s_t , including the current purchase price λ_t^+ , and has no access to future prices. Although the simulator uses a full 24-hour historical price trajectory to advance the environment, the policy receives the price one step at a time; recurrence (LSTM) is therefore needed to exploit temporal regularities without explicit forecasts.

6.1.3. RPPO architecture and training protocol

The ProHome agent employs an actor-critic architecture based on RPPO.

- **Actor Network:** Consists of an LSTM layer (128 hidden units) followed by three fully connected MLP layers (128 neurons each). The LSTM layer processes the input state sequence, capturing temporal features, which are then fed into the MLP layers. Tanh activation functions are used in the hidden layers. The output layer generates the parameters for the action distributions: 4 means and 4 standard deviations for the Gaussian distributions (controlling HVAC, ESS, EV, and EWH actions) and 3 probabilities for the Bernoulli distributions (controlling the 3 deferrable appliances).
- **Critic Network:** Comprises four MLP hidden layers (128 neurons each) with Tanh activations, followed by an output layer producing a single scalar value representing the estimated state value $V_\phi(s_t)$. This network does not include recurrent layers.

Standard practices are followed for training: Input observations (states) are normalized based on statistics gathered from initial environment rollouts using a random policy. The Adam optimizer is used for training both networks. A cosine annealing learning rate schedule [41] adjusts the learning rate over the course of training: from an initial 1×10^{-3} down to a final 1×10^{-5} for ProHome, from 1×10^{-3} down to 2×10^{-5} for standard PPO, and from 3×10^{-4} down to 1×10^{-5} for SAC. Key PPO hyperparameters include a clipping parameter $\epsilon = 0.2$ and a discount factor $\gamma = 0.9$. Gradients are clipped at a maximum norm of 1.0. No entropy bonus is used. No fixed random seed is set; each training run uses a fresh environment seed. Experiences (sequences of

transitions) are stored in an on-policy rollout buffer that is cleared between update cycles, and prioritized sampling (PER [38]) is used only as a batch-sampling strategy within the current rollout to emphasize high-error transitions during the PPO optimization epochs. The critic network is optimized using an L2 loss (mean squared error) against the calculated target values.

The implementation utilizes PyTorch 2.4.0 and was executed on a Windows 11 system with a 12th Gen Intel Core i9-12900 processor, 32 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU.

6.1.4. Benchmark algorithms

The proposed RPPO-based ProHome agent is benchmarked against:

1. Standard PPO (feedforward): isolates the effect of recurrence.
2. SAC: a strong off-policy actor-critic baseline for continuous control [42].
3. Perfect Information Optimum (PIO): an offline MILP benchmark with full future information.
4. Random policy: an uncontrolled lower bound.

The PIO baseline assumes perfect foresight of all future system dynamics. Under this perfect-foresight assumption, the energy management problem can be formulated as a deterministic Mixed-Integer Linear Programming (MILP) problem, following the PIO approach adopted from [21]. The optimal solution, representing the minimum achievable cost, is computed offline using the PySCIPOpt solver [43]. This serves as a theoretical lower bound on cost.

The standard PPO benchmark utilizes identical network architectures for its actor and critic: each consists of four fully connected MLP hidden layers with 128 neurons per layer (using Tanh activation), matching the MLP component depth of the RPPO agent but lacking the LSTM layer. All other hyperparameters and training settings (optimizer, learning rate schedule, rollout buffer, PER sampling, ϵ , γ , loss functions) are kept consistent with the RPPO implementation to ensure a fair comparison focused on the impact of recurrence.

The SAC benchmark also employs identical actor and critic networks, each with four MLP hidden layers (128 neurons, Tanh activation). As SAC inherently operates in continuous action spaces, its actor outputs a single continuous action vector. To accommodate the discrete actions required for the deferrable appliances, the last three action dimensions are discretized by a hard threshold after squashing: we map the last three normalized outputs to $[0, 1]$ via $(a + 1)/2$ and then apply rounding to obtain $\{0, 1\}$ start/wait commands.

This post-hoc discretization is a heuristic: SAC does not natively model discrete decisions, and thresholding/rounding can reduce the quality of gradient information around the decision boundary compared to a policy that explicitly parameterizes Bernoulli distributions. We acknowledge this introduces a fairness limitation, as SAC cannot natively optimize binary scheduling decisions as directly as the proposed composite PPO/RPPO architecture.

Table 8 summarizes the key architectural and training parameter details of the implemented DRL algorithms.

A formal parameter-level sensitivity analysis (systematically varying α , reward weights, and random seeds) was not performed and is identified as future work. The existing experimental design does, however, provide evidence of robustness under the tested conditions: the cross-algorithm comparison (RPPO, PPO, SAC, PIO, random policy) shows consistent relative performance across different control paradigms; the chronological 75/25 train/test split across 1089 days spanning multiple seasons shows stable behavior across weather, price, and load variability; and the cumulative cost analysis under the base-case export-price multiplier ($\alpha = 0.5$) shows the DRL agents generating higher revenue (€199–€207) than PIO (€177), indicating that they exploit export opportunities PIO ignores. These comparisons demonstrate robustness across conditions rather than a controlled sensitivity sweep.

Table 8
HEMS algorithms details.

	ProHome	PPO	SAC
Actor layers	1 LSTM + 3 hidden MLP	4 hidden MLP	4 hidden MLP
Critic layers	4 hidden MLP	4 hidden MLP	4 hidden MLP
Neurons per layer	128	128	128
Optimizer	Adam	Adam	Adam
Initial LR	1×10^{-3}	1×10^{-3}	3×10^{-4}
Final LR	1×10^{-5}	2×10^{-5}	1×10^{-5}
LR scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
Critic loss type	L2	L2	L2
Discount γ	0.9	0.9	0.99
PPO clip ϵ	0.2	0.2	—
GAE λ	0.95	0.95	—
PPO epochs K	10	10	—

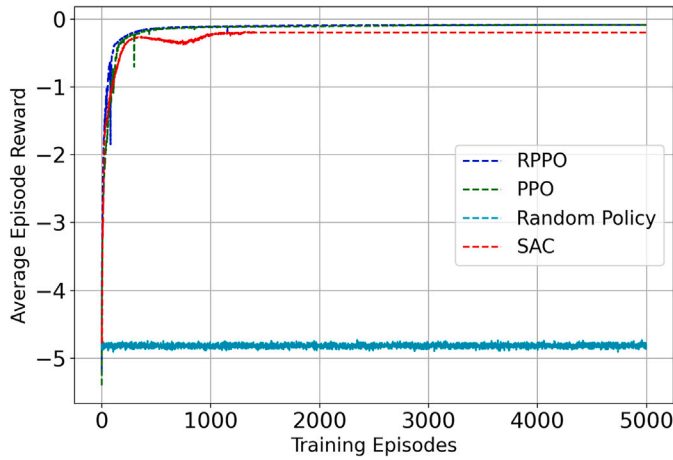


Fig. 6. Training curves showing average episode reward for RPPO, PPO, and SAC agents.

6.2. Performance comparisons

This subsection compares the performance of the trained ProHome agent (using RPPO) against the benchmark algorithms (standard PPO, SAC, PIO, and a random policy) on the unseen test dataset.

6.2.1. Training convergence

Fig. 6 displays the learning curves (average episode reward vs. training episodes) for the DRL algorithms. Both RPPO and standard PPO demonstrate efficient learning, converging to a similar average reward level of approximately -0.084 after about 1000 episodes. SAC exhibits slower convergence, requiring around 1410 episodes to reach a plateau at a significantly lower average reward of -0.2 . This suggests that, for this specific HEMS problem, the PPO-based algorithms converge faster and potentially find better policies than SAC. The final learned policy from each trained agent is used for evaluation on the test set.

6.2.2. Quantitative evaluation on test set

Table 9 summarizes the key performance metrics averaged over the 273-day test set.

As expected, the PIO benchmark achieves the best theoretical performance due to its perfect foresight. It incurs the lowest cost and maintains perfect user satisfaction (zero negative reward for HVAC and EWH comfort violations, guaranteed by hard constraints in the MILP) and ensures the EV reaches exactly the target SoE (E_{tr}^{EV}) at departure.

Comparing the DRL agents:

- ProHome (RPPO) achieves the highest total reward (least negative) among the DRL agents (-1983) and, critically, the lowest average daily electricity cost (68.28€). It demonstrates excellent user

Table 9
Comparison among the HEMS algorithms (cost reported in $\text{€}/\text{day}$).

	ProHome	PPO	SAC	PIO	Rand Pol
Reward sum	-1983	-2024	-4833	-1080	$-125,243$
Mean daily cost (€)	68.28	69.03	112.79	39.2	364
Mean r_t^{HVAC}	0	0	-0.16	0	-12.1
Mean r_t^{EWH}	-0.01	0	-0.08	0	-10.8
Mean $E_{departure}^{EV}$ (%)	67.77	65.67	54.2	70	39.4

comfort performance, with near-zero average penalties for HVAC and EWH temperature deviations, indicating successful maintenance of comfort bounds. The average EV departure SoE (67.8%) is very close to the target (70%), representing a 2.1 percentage point improvement over the standard PPO and signifying high user satisfaction regarding travel needs.

- Standard PPO performs comparably to RPPO in terms of total reward (-2024) and comfort metrics. Its average daily cost (69.03€) is slightly higher than RPPO's. It shows marginally better EWH satisfaction (average reward closer to zero) by a negligible amount (0.01). The average EV departure SoE is slightly lower (65.7%) than RPPO's.
- SAC lags significantly behind both PPO variants in total reward (-4833) and incurs the highest average daily cost (112.79€). While it manages HVAC/EWH comfort reasonably well (low average penalties), its EV departure SoE is considerably lower (54.2%), indicating poorer handling of the EV charging constraint or objective.
- The Random Policy performs poorly across all metrics, establishing a baseline of uncontrolled operation.

These results highlight the benefit of recurrence under forecast-free observation. Compared to standard PPO, RPPO reduces the mean daily cost by about 1% (68.28 vs. $69.03 \text{€}/\text{day}$, a difference of $0.75 \text{€}/\text{day}$) while maintaining near-zero comfort penalties. This improvement is consistent with the LSTM capturing recent price/SoE/temperature trends and enabling more anticipatory decisions (e.g., charging storage before higher-price intervals) without using explicit forecasts. The small cost difference is mirrored in other aggregate metrics: cumulative revenue ($\text{€}199$ vs. $\text{€}207$) and cumulative purchase cost ($\text{€}386$ vs. $\text{€}395$) are closely matched, suggesting that the LSTM produces only marginal refinements to temporal load-shifting rather than a qualitatively different device-level strategy. Confirming this at per-timestep resolution (e.g., overlaying RPPO and PPO temperature and power trajectories on a common test day) is identified as future work, as it requires rerunning the evaluation pipeline with per-timestep state logging.

6.2.3. Cumulative cost and revenue analysis

Fig. 7 provides insights into the economic performance by plotting the cumulative electricity purchase costs and sales revenues over the 273-day test period for each algorithm.

Analysis of the cumulative results reveals:

- Revenue: The DRL algorithms generate similar cumulative revenues from selling surplus electricity (PPO: $\text{€}207$, RPPO and SAC: $\text{€}199$). The PIO generates the lowest revenue ($\text{€}177$), indicating it prioritizes self-consumption or avoids selling at low prices.
- Cost: Significant differences appear in cumulative purchase costs. PIO achieves the lowest cost ($\text{€}285$) due to perfect optimization. Among DRL agents, ProHome (RPPO) incurs the lowest cost ($\text{€}386$), followed closely by PPO ($\text{€}395$), while SAC incurs the highest cost ($\text{€}507$).
- Net Cost: Considering the net financial outcome (Cost - Revenue), ProHome (RPPO) achieves the best result among DRL agents ($\text{€}386 - \text{€}199 = \text{€}187$), slightly better than PPO ($\text{€}395 - \text{€}207 = \text{€}188$), and substantially better than SAC ($\text{€}507 - \text{€}199 = \text{€}308$). The PIO achieves the lowest net cost ($\text{€}285 - \text{€}177 = \text{€}108$).

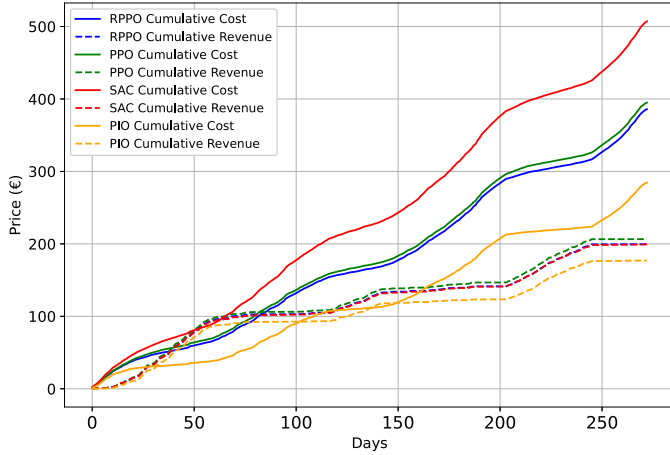


Fig. 7. Cumulative electricity purchase cost and sales revenue over the 273-day test period (€).

The PIO's strategy confirms that minimizing overall cost often involves maximizing self-consumption of PV energy rather than maximizing sales revenue, especially when the feed-in tariff (λ^-) is significantly lower than the purchase price (λ^+). ProHome's superior cost performance compared to other DRL methods suggests its recurrent structure allows it to better learn and exploit temporal patterns in prices and PV generation for more effective load shifting and storage management, leading to reduced grid imports during expensive periods. This self-consumption versus export behavior is the primary source of the 29.08 cent/day gap to the PIO lower bound: the DRL agents export more aggressively when feed-in prices are favorable, generating higher revenue but also incurring higher purchase cost, whereas PIO preferentially stores or self-consumes PV energy.

Although RPPO achieves the lowest daily cost among the tested forecast-free DRL agents, the PIO benchmark remains substantially better (39.2 vs. 68.28 cent/day, a gap of 29.08 cent/day). However, this gap should not be interpreted as a shortcoming of the RL formulation nor as a gap that forecast-based MILP methods could easily close in practice. Three structural barriers limit the deployability of forecast + MIP approaches: (i) Non-deferrable (fixed) loads are inherently stochastic: their exact activation times and power draws cannot be predicted in advance, yet PIO assumes perfect knowledge of these profiles; (ii) Forecast errors compound over the 96-step planning horizon (24 hours at 15-minute resolution): even marginal per-step prediction errors for outdoor temperature, PV generation, and EV arrival SoE accumulate and degrade the MILP solution quality; and (iii) Day-ahead MILP schedules are rigid and cannot adapt to mid-day deviations from forecasted conditions, whereas an RL policy reacts online to the observed state at each time step. PIO should therefore be interpreted strictly as a perfect-information lower bound rather than a deployable benchmark under realistic operating constraints.

6.3. HVAC operational analysis

The ProHome HEMS manages the HVAC system to maintain the indoor temperature T^{in} within the user-defined comfort range $[T^{\text{HVAC}}_{\text{min}}, T^{\text{HVAC}}_{\text{max}}]$ by activating heating or cooling modes as necessary. The following case studies and aggregate results illustrate its performance.

6.3.1. Mixed heating and cooling scenario

Fig. 8 presents a representative day where the outdoor temperature T^{out} exhibits significant variation: starting cold, rising above the upper comfort bound during the day, and dropping below the lower bound at night. The ProHome agent successfully maintains T^{in} strictly within the specified comfort bounds throughout the entire day. Notably, the

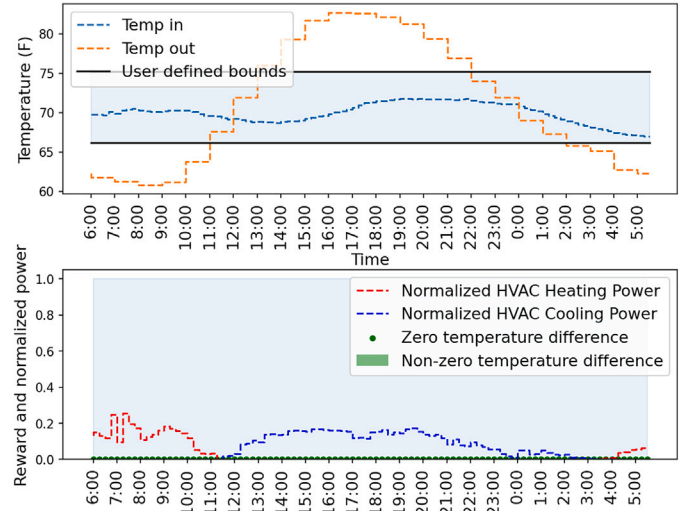


Fig. 8. Example day: HVAC operation (bottom panel, showing normalized heating/cooling power) maintaining indoor temperature (top panel, blue line) within comfort bounds (top panel, dashed red lines) despite varying outdoor temperature (top panel, Orange line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

HVAC achieves this with relatively low power consumption, generally operating between 0% and 25% of its maximum capacity. The agent demonstrates anticipatory control; for example, it transitions from heating to cooling proactively as the outdoor temperature rises, anticipating the need to counteract solar gains or further temperature increases, thereby enhancing both comfort and efficiency.

6.3.2. Predominant heating scenario

Fig. 9 depicts a day characterized by consistently cold outdoor temperatures, remaining significantly below the lower comfort bound ($T^{\text{HVAC}}_{\text{min}}$). Despite this persistent thermal load, the ProHome HEMS effectively operates the HVAC in heating mode to keep the indoor temperature T^{in} within the desired range. The heating power consumption fluctuates, primarily between 25% and 50% of maximum capacity, responding to the temperature dynamics. The indoor temperature often tracks near the lower comfort bound, indicating the agent's strategy balances minimizing energy consumption (by avoiding overheating) with strictly maintaining user comfort.

6.3.3. Aggregate comfort performance

User satisfaction with HVAC control across the entire 273-day test set is evaluated using the daily average HVAC reward (r^{HVAC} from Eq. (20), averaged over the day). Fig. 10 plots this daily average reward. A value of zero signifies perfect comfort maintenance for that day. The results show that the vast majority of days achieve an average reward of zero. On days where minor deviations occur, the average penalty is extremely small (lowest value approx. -0.006), corresponding to a negligible average temperature deviation from the bounds. Furthermore, the maximum instantaneous temperature deviation observed across all time steps in the test set was only 0.27°F . This aggregate performance underscores the ProHome agent's capability to consistently ensure thermal comfort while optimizing energy use. At the aggregate level, all three agents (RPPO, PPO, and PIO) maintain thermal comfort at comparable levels: RPPO and PPO both achieve near-zero HVAC and EWH daily penalties, and PIO achieves zero violations by construction through hard MILP constraints. This indicates that the three agents follow broadly similar temperature trajectories that respect the comfort band, and the cost differences between them cannot be attributed to comfort trade-offs.

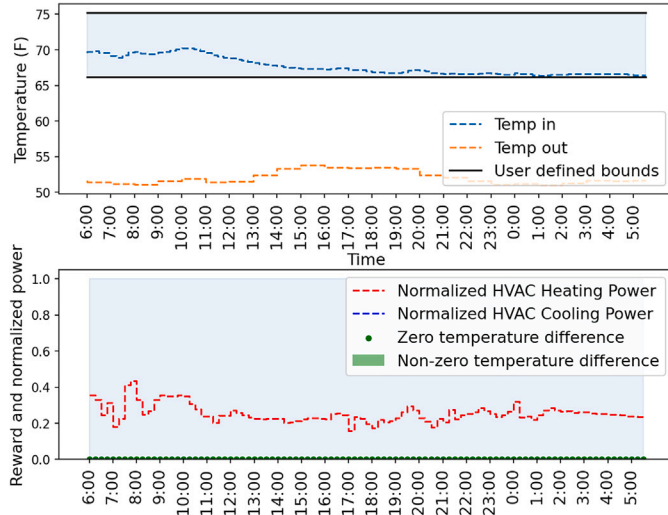


Fig. 9. Example day: HVAC primarily in heating mode maintains indoor comfort during cold outdoor conditions.

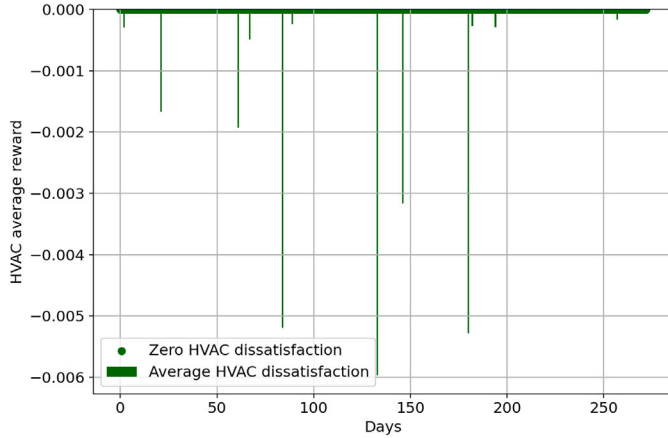


Fig. 10. Distribution of daily average HVAC rewards over the test set, indicating high consistency in maintaining comfort.

6.4. EWH operational analysis

The ProHome HEMS controls the EWH's heating element (p_t^{EWH}) to maintain the water temperature T^{EWH} within its specified range $[T_{\min}^{\text{EWH}}, T_{\max}^{\text{EWH}}]$, balancing energy costs against hot water availability.

6.4.1. Example day operation

Fig. 11 shows the EWH operation on a representative test day. The agent generally maintains the water temperature within the desired bounds. A brief dip below the lower limit occurs late in the day (around 23:00), potentially due to significant hot water usage coinciding with a decision to defer heating based on high electricity prices or other system constraints. The heating element power is modulated throughout the day, often operating below its maximum capacity (peak usage around 40% shown) and allowing the temperature to drift toward the lower bound ($T_{\min}^{\text{EWH}}$) when possible to conserve energy. This illustrates the HEMS strategically minimizing heating energy while largely respecting the comfort constraint.

6.4.2. Aggregate comfort performance

Similar to the HVAC analysis, Fig. 12 presents the distribution of the daily average EWH reward (r_t^{EWH} from Eq. (21), averaged over the

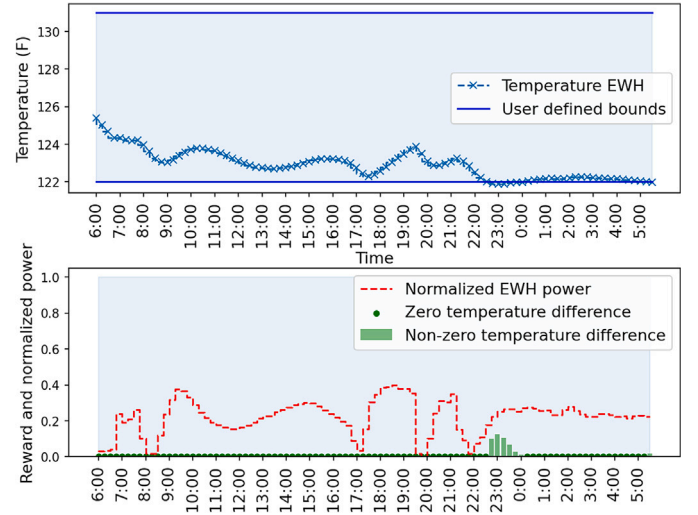


Fig. 11. Example day: EWH operation (bottom panel, showing normalized heating power) maintaining water temperature (top panel, blue line) largely within comfort bounds (top panel, dashed red lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

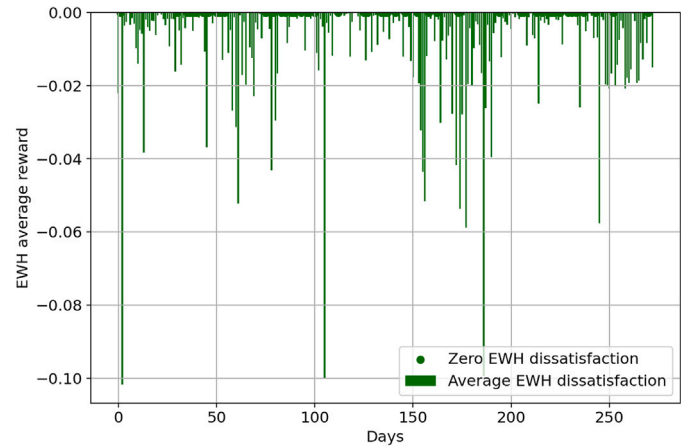


Fig. 12. Distribution of daily average EWH rewards over the test set.

day) across the test set. While some days exhibit small negative average rewards (minimum approx. -0.1), indicating occasional or minor temperature deviations, the overall average reward across all test days is only -0.008 . This corresponds to a negligible average deviation from the target temperature range. The maximum instantaneous temperature drop below $T_{\min}^{\text{EWH}}$ observed across all time steps was 0.95°F . These results demonstrate that the ProHome agent effectively manages the EWH, achieving significant energy savings by allowing the temperature to float within the acceptable range while ensuring user satisfaction remains high, with only minor and infrequent comfort boundary violations.

6.5. Fixed load analysis

Fixed loads, such as refrigerators or background electronics, operate based on their internal controls or immediate user needs and are not actively scheduled by ProHome. Their aggregated power consumption, however, constitutes a significant and often unpredictable component of the total household load that the HEMS must account for. Fig. 13 displays the combined power profile of all fixed loads for a sample day. The profile shows a base consumption level (around 0.4 kW, largely

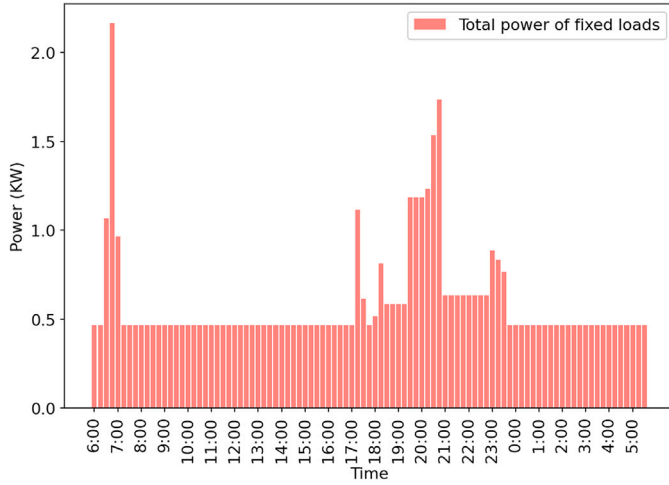


Fig. 13. Example profile of aggregated power consumption from all fixed loads over one day.

from continuous loads like refrigeration) superimposed with intermittent peaks from other fixed devices. This inherent variability and lack of direct control present a challenge for the HEMS, which must adapt its management of controllable assets in response to these fluctuating background loads.

6.6. Deferrable appliance scheduling analysis

ProHome schedules the operation of deferrable appliances (washing machine, dishwasher, clothes dryer) by deciding their start time (t_{start}^d) after user initialization (t_{ini}^d). This decision involves a trade-off: starting immediately minimizes user waiting time (improves r_t^d) but might occur during high electricity prices, whereas delaying shifts operation toward lower-price periods at the cost of a waiting penalty. Since future prices are unknown, the HEMS must act based on the current state and learned temporal patterns.

Fig. 14 illustrates this scheduling process for a specific test day.

- The washing machine and dishwasher are initialized by the user at 17:00 and 22:00, respectively (indicated by vertical dashed lines). ProHome starts both appliances immediately ($t_{start}^d = t_{ini}^d$). This strategy avoids any waiting penalty and also preempts a rise in electricity prices occurring later in the evening (shown by the price curve overlay), proving cost-effective in hindsight for these specific times.
- The clothes dryer is initialized at 23:45, coinciding with a peak in electricity prices. ProHome opts to delay the operation, starting the dryer cycle only at 01:15 when the price has significantly decreased. This decision prioritizes cost savings over immediate operation.

This example demonstrates ProHome's ability to balance the conflicting objectives of user convenience and cost minimization based on the perceived state and learned value of delaying versus immediate execution. The relative importance of these objectives can be tuned by adjusting the weight ω_d associated with the waiting time reward (Eq. 24) during agent training or potentially as a user preference setting.

6.7. Electricity trading and energy flow analysis

This subsection analyzes the economic performance resulting from ProHome's management of energy flows, specifically focusing on electricity trading with the grid. The net daily financial outcome is calculated as the total revenue from selling electricity minus the total cost of purchasing electricity over 24 hours, using the time-varying prices λ_t^-

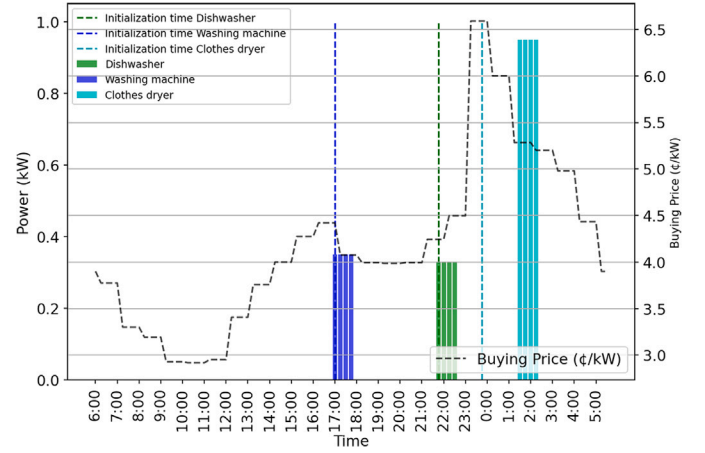


Fig. 14. Example scheduling of deferrable loads (Dishwasher, Washing Machine, Clothes Dryer). Vertical lines indicate user initialization times; solid blocks show actual operation periods. The background displays the electricity price curve.

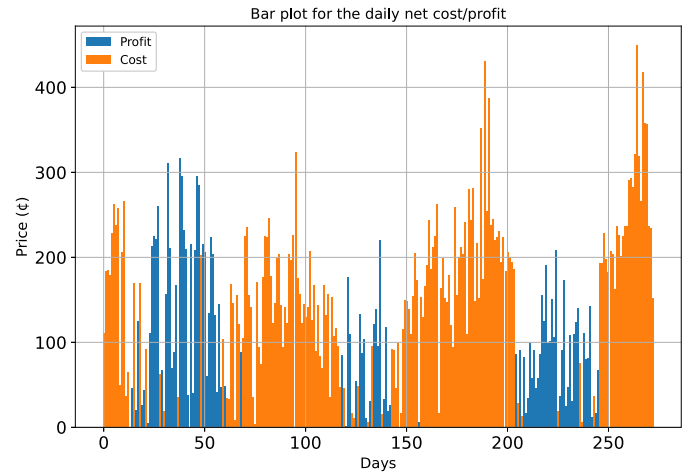


Fig. 15. Net financial outcome (Profit/Cost) for each day in the test set.

and λ_t^+ (where $\lambda_t^- = 0.5\lambda_t^+$). A positive outcome represents a net profit, while a negative outcome signifies a net cost.

Fig. 15 plots this net daily outcome for each day in the test set. As anticipated, there is significant day-to-day variability, strongly correlated with PV generation levels. Days with high solar output tend toward net profit, whereas days with low generation typically result in a net cost. The overall average across the test set corresponds to the average daily cost reported previously (approx. 68 €). To provide deeper insights into the agent's decision-making under different conditions, three representative case study days are examined: one resulting in net profit, one in net cost, and one near break-even.

6.7.1. Case 1: Net profit day (€+2.1)

On this day, characterized by substantial PV generation, ProHome achieves a net profit. Fig. 16 shows the instantaneous cost/profit profile alongside the electricity price, while Fig. 17 details the energy flows of major components (EV/ESS SoE, PV generation, total load).

- Morning Charging: The EV starts with 6 kWh and is charged using low-cost grid power early in the morning (before 09:00) to reach its target SoE (approx. 11.2 kWh) before departure. The ESS also utilizes cheap grid power initially and then PV generation (starting around 07:00) to charge.

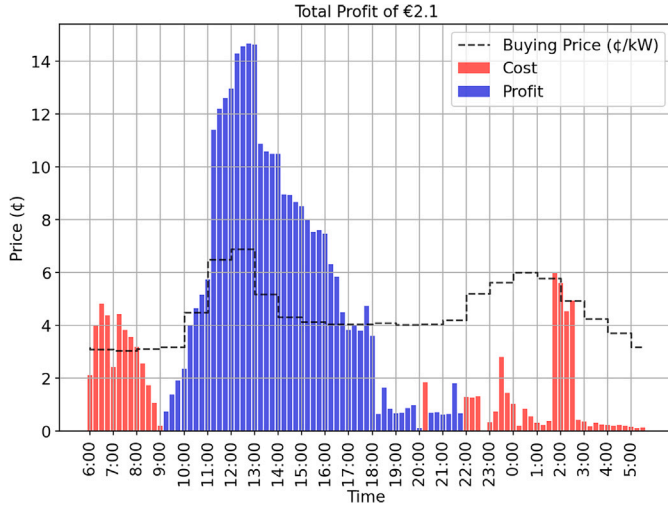


Fig. 16. Case 1 (Net Profit): Instantaneous cost (positive bars) / profit (negative bars) profile and electricity purchase price.

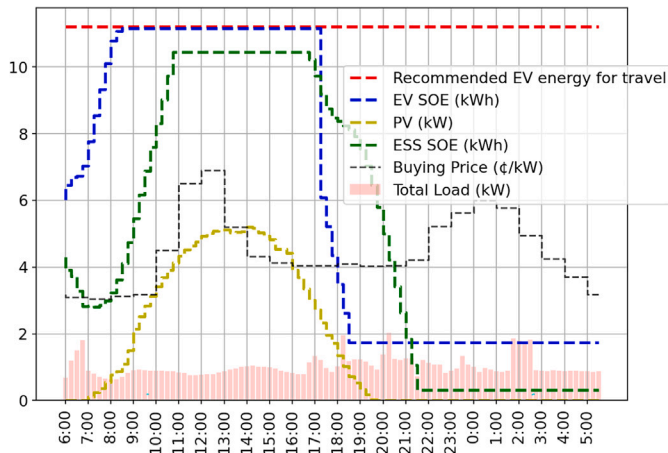


Fig. 17. Case 1 (Net Profit): Energy flow dynamics showing EV/ESS SoE, PV generation, total load, and electricity price.

- **PV Utilization and Export:** As PV generation ramps up, grid purchases cease (by 09:00). Surplus PV energy is used to meet loads and charge the ESS, with excess sold back to the grid, particularly during higher price periods later in the day, maximizing revenue.
- **Evening Discharge:** After PV generation ends and the EV returns (17:15), both the ESS and the returned EV discharge to meet evening loads, minimizing grid purchases during typically higher price periods. Minimal grid import is needed late at night (e.g., for the clothes dryer).

This case demonstrates ProHome's effective coordination of charging, PV self-consumption, and strategic grid export to capitalize on favorable conditions.

6.7.2. Other trading scenarios: net cost and break-even days

While Case 1 illustrates a highly profitable day due to abundant solar generation, ProHome's control logic remains robust under less favorable conditions. On days with minimal PV output (Net Cost Day, €-2.4), the agent relies heavily on the grid but strategically charges the ESS during low-price periods to mitigate expenses during anticipated afternoon peaks. Conversely, on days with moderate PV generation (Break-Even Day, €-0.002), the system balances grid imports with profitable surplus exports. Across all scenarios, despite lacking explicit price forecasts, the

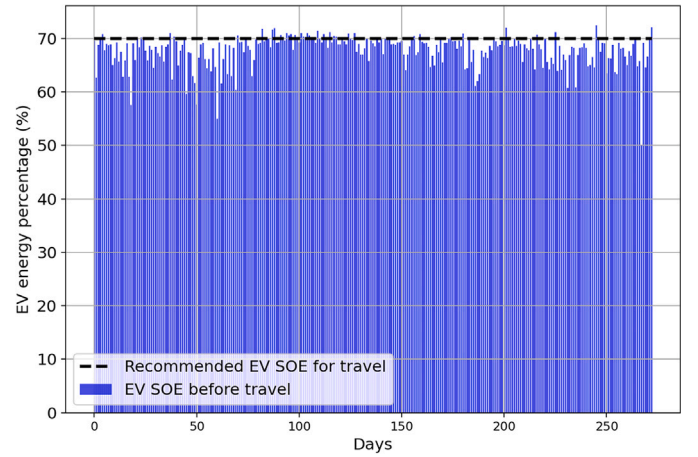


Fig. 18. EV state of energy (SoE) at departure time for each day in the test set, compared to the 70% target level (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

RPPO agent consistently demonstrates a learned strategy of purchasing energy during low-price valleys and avoiding grid purchases (or exporting) during peak periods.

6.8. EV departure state of energy satisfaction

User satisfaction related to the EV is primarily linked to having sufficient charge for the intended travel at the time of departure. Fig. 18 plots the EV's SoE ($E_{\text{departure}}^{\text{EV}}$), expressed as a percentage of its maximum capacity, at the scheduled departure time for each day in the test set. The target SoE for travel ($E_{\text{tr}}^{\text{EV}}$) was set at 70% in these experiments.

The results show that ProHome consistently charges the EV to a level very close to the 70% target on most days. The average departure SoE across the entire test set is 67.8%. This indicates that the HEMS effectively prioritizes the EV charging requirement, ensuring high user satisfaction and vehicle readiness, while simultaneously optimizing charging times based on costs and other system demands.

6.9. Limitations and deployment considerations

The following limitations apply to the current study:

1. **Sub-optimal performance relative to PIO.** As discussed in Section 6, RPPO incurs 68.28 cents/day compared with 39.2 cents/day for the perfect-information optimum, a gap of 29.08 cents/day. This gap indicates significant room for improvement. Potential directions include removing the grid-export option to enforce self-consumption (the PIO consistently prioritizes storing PV energy over selling), exploring alternative architectures such as transformers that may capture longer-range temporal dependencies, and integrating short-horizon forecasts to narrow the information gap.
2. **Dataset and geographic specificity.** All experiments use data from a single location (Utrecht, Netherlands, 2015–2017) with a specific tariff structure and climate profile. Real-world deployment requires validation across diverse climates, home configurations, and electricity pricing regimes. A foundation-model approach (pretraining on large-scale multi-region datasets and fine-tuning per home) would be necessary for broad applicability.
3. **Computational overhead of recurrence.** The LSTM layer in RPPO adds sequential processing and additional parameters compared to the feedforward PPO baseline, increasing inference latency and memory requirements. Deployment on resource-constrained edge devices may require model compression or distillation.

4. Simulation-only evaluation. All training and testing are performed in a calibrated simulator driven by historical data; no real-home deployment was conducted. Performance may differ under real measurement noise, hardware actuation latency, and unanticipated user behavior. However, the ProHome environment formulation is deliberately comprehensive (covering fixed, deferrable, and modulatable loads with configurable appliance sets) and can represent a wide range of typical households.
5. Reward-design and tariff dependency. Results depend on the specific reward weights (Table 6) and the base-case export-price multiplier ($\alpha = 0.5$). Different tariff structures, user comfort preferences, or home configurations would require retraining and re-calibration.

7. Conclusion

This paper presents ProHome, a comprehensive Home Energy Management System (HEMS) based on Recurrent Proximal Policy Optimization (RPPO) for real-time scheduling and control under dynamic pricing. ProHome jointly manages a rich appliance set (fixed and deferrable loads, HVAC, EWH, ESS, and EV) using a unified composite action representation (continuous setpoints and binary start decisions) and operates without access to future electricity prices.

Results on 1089 days of historical data (Utrecht, 2015–2017) show that recurrence provides an incremental performance benefit under forecast-free observations: RPPO reduces the mean daily electricity cost compared to the non-recurrent PPO baseline while maintaining near-zero thermal comfort violations and high EV departure SoE satisfaction.

Future work will focus on (i) extending ProHome to community/multi-home settings while mitigating multi-agent non-stationarity and coordination overhead, (ii) integrating forecasting or uncertainty-aware price features while preserving non-anticipative deployment constraints, and (iii) exploring alternative architectures (e.g., transformers) and pretraining on larger, multi-region datasets to improve generalization and narrow the gap to the PIO lower bound. The large gap to the PIO lower bound (29.08 cent/day) reflects the economic value of future information; however, as discussed in Section 6.9, closing this gap through forecasting alone is structurally limited by stochastic loads, compounding prediction errors, and schedule rigidity. Hybrid approaches combining recurrent RL with short-horizon robust forecasts and model-predictive re-optimization represent a promising direction.

CRedit authorship contribution statement

Mahmoud Sallam: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kuljeet Kaur:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

This work is supported by the Department of Science and Technology, the India-Canada Centre for Innovative Multidisciplinary Partnerships to Accelerate Community Transformation and Sustainability (IC-IMPACTS), and the Centre of Excellence (Research on Remoulding Existing Buildings into Smart Carbon Neutral Buildings: A Green Cyber-Physical Approach (Green-CPS), NO. (DST/IC/IC-IMPACTS/2022/P-8 (G)), and also the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2021-04013.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mahmoud Sallam reports that financial support was provided by IC-IMPACTS. If there are other authors, they declare that they have no

known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Perfect-information MILP (PIO) benchmark

The Perfect-Information Optimum (PIO) benchmark is intended as a deterministic offline optimization that provides a lower bound on cost under a perfect-foresight assumption: the optimizer has access to the full-day exogenous trajectories before making decisions. The benchmark follows the PIO/MILP approach adopted from [21].

A.1. Horizon and objective

The problem is solved day-by-day over a 24-hour horizon with $T = 96$ time steps (15-minute resolution). The objective is to minimize the total electricity cost defined in (17):

$$\min \sum_{t=0}^{T-1} C_t. \quad (\text{A.1})$$

A.2. Key constraints (consistent with the DRL environment)

PIO decisions are constrained to follow the same component models and limits used in the simulation environment:

- Energy balance and grid exchange: Net household power is computed as in (16), with the corresponding import/export cost computed using (17).
- Thermal comfort: HVAC dynamics and comfort bounds follow (2)–(3e); EWH dynamics and temperature bounds follow (4) and (6).
- Storage and EV: ESS and EV SoE dynamics and bounds follow (7)–(8e) and (9)–(10g), including EV availability and the departure energy requirement.
- Deferrable appliances: Tasks are scheduled as non-interruptible runs according to (11)–(12).

A.3. Implementation note

The solver implementation used to compute the PIO trajectories is not included in the current submission package; therefore, this appendix documents the intended optimization model specification on which the reported PIO results are based.

Data availability

The data used in this study are publicly available from the cited sources (see Section 6.1).

References

- [1] H. Shareef, M.S. Ahmed, A. Mohamed, E. Al Hassan, Review on home energy management system considering demand responses, smart technologies, and intelligent controllers, *IEEE Access* 6 (2018) 24498–24509.
- [2] S. Mokeke, L.Z. Thamae, The impact of intermittent renewable energy generators on Lesotho national electricity grid, *Electr. Power Syst. Res.* 196 (2021) 107196.
- [3] M. Erol-Kantarci, H.T. Moufah, Wireless sensor networks for cost-efficient residential energy management in the smart grid, *IEEE Trans. Smart Grid* 2 (2) (2011) 314–325.
- [4] F.A. Qayyum, M. Naeem, A.S. Khwaja, A. Anpalagan, L. Guan, B. Venkatesh, Appliance scheduling optimization in smart home networks, *IEEE Access* 3 (2015) 2176–2190.
- [5] M. Khan, J. Seo, D. Kim, Real-time scheduling of operational time for smart home appliances based on reinforcement learning, *IEEE Access* 8 (2020) 116520–116534.
- [6] P. Lissa, C. Deane, M. Schukat, F. Seri, M. Keane, E. Barrett, Deep reinforcement learning for home energy management system control, *Energy AI* 3 (2021) 100043.
- [7] L. Yan, X. Chen, Y. Chen, J. Wen, A hierarchical deep reinforcement learning-based community energy trading scheme for a neighborhood of smart households, *IEEE Trans. Smart Grid* 13 (6) (2022) 4747–4758.
- [8] J. Lee, W. Wang, D. Niyato, Demand-side scheduling based on multi-agent deep actor-critic learning for smart grids, in: 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), IEEE, 2020, pp. 1–6.

- [9] J. Lu, P. Mannion, K. Mason, A multi-objective multi-agent deep reinforcement learning approach to residential appliance scheduling, *IET Smart Grid* 5 (4) (2022) 260–280.
- [10] Z. Wang, F. Xiao, Y. Ran, Y. Li, Y. Xu, Scalable energy management approach of residential hybrid energy system using multi-agent deep reinforcement learning, *Appl. Energy* 367 (2024) 123414.
- [11] F. Charbonnier, T. Morstyn, M.D. McCulloch, Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility, *Appl. Energy* 314 (2022) 118825.
- [12] D. Latoń, J. Grela, A. Ozadowicz, Applications of deep reinforcement learning for home energy management systems: a review, *Energies* 17 (24) (2024) 6420.
- [13] Q. Fu, Z. Li, Z. Ding, J. Chen, J. Luo, Y. Wang, Y. Lu, ED-DQN: an event-driven deep reinforcement learning control method for multi-zone residential buildings, *Build. Environ.* 242 (2023) 110546.
- [14] C. Huang, H. Zhang, L. Wang, X. Luo, Y. Song, Mixed deep reinforcement learning considering discrete-continuous hybrid action space for smart home energy management, *J. Mod. Power Syst. Clean Energy* 10 (3) (2022) 743–754.
- [15] M. Sallam, K. Kaur, G. Kaddoum, Advanced home energy management using proximal policy optimization with a comprehensive appliance set, in: *ICC 2025 - IEEE International Conference on Communications*, 2025, pp. 5779–5784.
- [16] A. Molderink, V. Bakker, M.G. Bosman, J.L. Hurink, G.J. Smit, Management and control of domestic smart grid technology, *IEEE Trans. Smart Grid* 1 (2) (2010) 109–119.
- [17] K.M. Tsui, S.-C. Chan, Demand response optimization for smart home scheduling under real-time pricing, *IEEE Trans. Smart Grid* 3 (4) (2012) 1812–1821.
- [18] A. Agnetis, G. De Pascale, P. Detti, A. Vicino, Load scheduling for household energy consumption optimization, *IEEE Trans. Smart Grid* 4 (4) (2013) 2364–2373.
- [19] M.A.A. Pedrasa, T.D. Spooner, I.F. MacGill, Coordinated scheduling of residential distributed energy resources to optimize smart home energy services, *IEEE Trans. Smart Grid* 1 (2) (2010) 134–143.
- [20] R.K. Yadav, P. Hrisheeksha, V.S. Bhadoria, Grey wolf optimization based demand side management in solar PV integrated smart grid environment, *IEEE Access* 11 (2023) 11827–11839.
- [21] J. Aldahmashi, X. Ma, Real-time energy management in smart homes through deep reinforcement learning, *IEEE Access* 12 (2024) 43155–43172.
- [22] T. Logenthiran, D. Srinivasan, T.Z. Shun, Demand side management in smart grid using heuristic optimization, *IEEE Trans. Smart Grid* 3 (3) (2012) 1244–1252.
- [23] J. Li, Z. Jiang, Z. Chen, J. Liu, L. Cheng, CuEMS: deep reinforcement learning for community control of energy management systems in microgrids, *Energy Build.* 304 (2024) 113865.
- [24] H. Shengren, P.P. Vergara, E.M.S. Duque, P. Palensky, Optimal energy system scheduling using a constraint-aware reinforcement learning algorithm, *Int. J. Electr. Power Energy Syst.* 152 (2023) 109230.
- [25] L. Langer, T. Volling, A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems, *Appl. Energy* 327 (2022) 120020.
- [26] S. Li, D. Cao, Q. Huang, Z. Zhang, Z. Chen, F. Blaabjerg, W. Hu, A deep reinforcement learning-based approach for the residential appliances scheduling, *Energy Rep.* 8 (2022) 1034–1042.
- [27] S. Lee, D.-H. Choi, Energy management of smart home with home appliances, energy storage system and electric vehicle: a hierarchical deep reinforcement learning approach, *Sensors* 20 (7) (2020) 2157.
- [28] J. Ruddick, G. Ceusters, G. Van Kriekinge, E. Genov, C. De Cauwer, T. Coosemans, M. Messagie, Real-world validation of safe reinforcement learning, model predictive control and decision tree-based home energy management systems, *Energy AI* 18 (2024) 100448.
- [29] H. Li, Z. Wan, H. He, A deep reinforcement learning based approach for home energy management system, in: *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, IEEE, 2020, pp. 1–5.
- [30] N. Lu, An evaluation of the HVAC load potential for providing load balancing service, *IEEE Trans. Smart Grid* 3 (3) (2012) 1263–1270.
- [31] M.H. Nehrir, R. Jia, D.A. Pierre, D.J. Hammerstrom, Power management of aggregate electric water heater loads by voltage control, in: *2007 IEEE Power Engineering Society General Meeting*, IEEE, 2007, pp. 1–6.
- [32] H. Li, Z. Wan, H. He, Real-time residential demand response, *IEEE Trans. Smart Grid* 11 (5) (2020) 4144–4154.
- [33] R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*, 2 ed, The MIT Press, 2018.
- [34] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, 2014.
- [35] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
- [36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347*, 2017.
- [37] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, *arXiv preprint arXiv:1506.02438*, 2015.
- [38] T. Schaul, Prioritized experience replay, *arXiv preprint arXiv:1511.05952*, 2015.
- [39] Hugging Face, EDS-lab/pv-generation (dataset), 2024, <https://huggingface.co/datasets/EDS-lab/pv-generation> (Accessed 3 October 2024).
- [40] Ember, Ember energy, 2024, <https://ember-energy.org/> (Accessed 3 October 2024).
- [41] PyTorch, CosineAnnealingLR (PyTorch documentation), 2024, https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.CosineAnnealingLR.html (Accessed 3 October 2024).
- [42] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 1861–1870.
- [43] S. Maher, M. Miltenberger, J.P. Pedroso, D. Rehfeldt, R. Schwarz, F. Serrano, PySCIPOpt: mathematical programming in Python with the SCIP optimization suite, in: *Mathematical Software – ICMS 2016*, Springer International Publishing, 2016, pp. 301–307.