KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-026-00908-0

PROJECT REPORTS

Deep Reinforcement Learning for Price-Aware Building Heating Control

Qiong Huang[1] · Adrian Till Assmuth[1] · Felix Langner[1] · Benjamin Schäfer[1] · Veit Hagenmeyer[1]

Received: 1 August 2025 / Accepted: 25 March 2026
© The Author(s) 2026

Abstract
Heating systems account for a significant share of residential energy consumption, and rising energy prices call for intelligent, cost-aware control strategies. Traditional methods, such as rule-based or model predictive control (MPC), often require detailed system modeling or lack adaptability to dynamic price signals. This work explores the use of deep reinforcement learning (DRL) to control heat pumps in a way that balances occupant comfort with energy-cost minimization. We evaluate deep Q-network (DQN) and proximal policy optimization (PPO) methods across discrete and continuous action spaces. The agents are trained in simulation using real weather and electricity price data, with a model representing the thermal dynamics of the building. Short-term electricity price forecasts are included to enable anticipatory heating strategies. Reward functions combine price penalties with piecewise-linear or quadratic comfort penalties. Among the DRL variants, a DQN agent with discrete actions and a piecewise-linear comfort reward achieves the best overall trade-off between comfort and cost. MPC still performs best in absolute cost terms because it uses an exact model, while the DQN policy approaches MPC performance and retains the model-free, adaptive advantages of RL. The findings highlight the potential of DRL for adaptive and price-aware heating control without the need for detailed physical modeling.

Keywords Reinforcement learning · Heat pump control · Smart buildings

1 Introduction

Residential heating systems represent a major component of household energy use and are important driver of peak electricity demand during the colder seasons [9]. With the increasing integration of intermittent renewable energy sources such as solar and wind, aligning energy demand with variable supply is becoming a crucial strategy for achieving grid stability and decarbonization goals [8]. Due to their inherent thermal inertia, buildings are well-suited for such demand side management (DSM): they can preheat during periods of low-cost, abundant electricity and reduce heating when prices or grid stress increase.

Traditional control approaches for building heating, including rule-based controllers and model predictive control (MPC), have notable limitations. Rule-based methods are simple but rigid, lacking adaptability to changing external conditions or user preferences. MPC, while more sophisticated, relies on accurate building models and involves solving optimization problems at every control step, which is computationally intensive and difficult to scale [1, 19].

In recent years, machine learning techniques have gained traction for building energy management, particularly reinforcement learning (RL) methods which learn control policies through interaction with the environment. RL has emerged as a promising model-free alternative in optimizing complex tasks by balancing multiple objectives, such as occupant comfort and energy cost [20]. Deep reinforcement

GitHub Repository: https://github.com/Flywienix/rl-heatpump.git.

✉ Qiong Huang
qiong.huang@kit.edu

Adrian Till Assmuth
adrian.assmuth9@kit.edu

Felix Langner
felix.langner@kit.edu

Benjamin Schäfer
benjamin.schaefer@kit.edu

Veit Hagenmeyer
veit.hagenmeyer@kit.edu

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Published online: 17 April 2026

learning (DRL) methods extend this capability by using neural networks to approximate value functions or policies, enabling the handling of high-dimensional state spaces, which is typical in real-world applications [12].

Our use case considers a residential heat-pump controller operating at fixed sampling intervals under time-varying weather and electricity prices. The agent receives thermal states and price-forecast features, and selects a bounded heating action that must satisfy comfort requirements while reducing cost. We make modeling assumptions and discuss how these assumptions affect interpretation and transferability. We investigate the use of DRL for cost-aware heat pump control under dynamic electricity pricing. Specifically, we compare two DRL algorithms, deep Q-network (DQN) [11] and proximal policy optimization (PPO) [17], in both discrete and continuous action spaces. Agents are trained in a simulated environment using real-world weather and price data, and their performance is evaluated in terms of both thermal comfort and energy cost. We design a reward function that combines a comfort zone penalty with cost incentives, enabling agents to learn anticipatory heating strategies based on price forecasts.

The remainder of the paper is organized as follows: Sect. 2 summarizes related work; Sect. 3 introduces the methodology; Sect. 4 presents the experimental setup and results; Sect. 5 discusses key findings, followed by the conclusion and outlook in Sect. 6.

2 Related Work

Early work already demonstrated that model-free RL can control residential heat pumps without detailed white-box models, including thermostat set-back strategies and residential demand-response formulations [14, 15, 18]. These studies established feasibility, but typically used simpler control objectives and limited action formulations compared with recent DRL approaches.

More recent DRL literature reports stronger control performance for HVAC and heat-pump applications, including improved comfort tracking and energy reduction in simulation studies [2, 12, 13, 20]. Several studies also target economically oriented operation under variable prices and flexible assets, showing that DRL can learn anticipatory behavior and load shifting when reward design and observation spaces are chosen carefully [6, 9, 16].

In parallel, transfer-learning based controllers are increasingly explored to reduce training burden and improve scalability across buildings, climates, and operational contexts [4, 7, 10]. This direction is particularly relevant for practical deployment, where training from scratch for each building can be costly.

For cost-aware operation under explicit constraints, model predictive control remains a strong baseline because it can jointly optimize comfort and cost when an accurate model is available [1, 19]. However, deployment at scale can be limited by model-identification effort and repeated online optimization.

Against this background, our study focuses on a controlled comparison of DQN and PPO for price-aware heat pump control, with explicit benchmarking against MPC, modeling assumptions, and validation on unseen weather and price profiles.

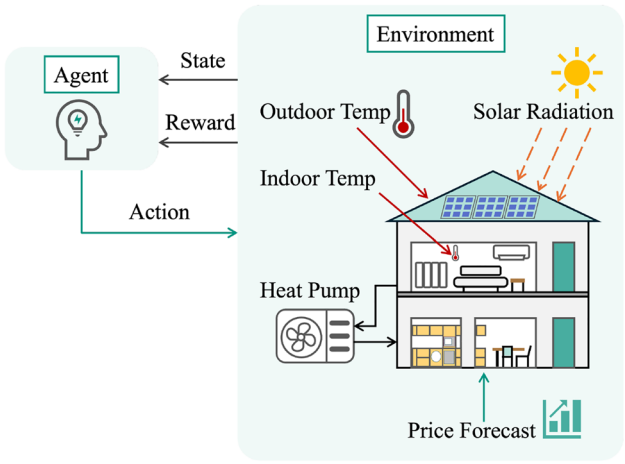

Fig. 1 System model used for RL-based heating control, including the DRL controller (Action), weather and price inputs (with forecasts), heat-pump actuation, the building thermal state-space model, and the observation/reward feedback loop

3 Method

The RL agents are trained in a simulation environment that models the thermal dynamics of a building with a heat pump. An overview of the building simulation environment is shown in Fig. 1, where the environment provides feedback in the form of state observations and rewards based on temperature deviations and energy costs.

Simulation Environment The environment is based on a linear state-space model derived from Vallianos et al. [19], which captures the indoor air and thermal mass temperatures, as well as the effects of ambient temperature and solar radiation. The agents learn to control the power output of the heat pump to maintain a comfortable indoor temperature while minimizing energy costs. Data are sampled at 15-minute intervals using a Typical Meteorological Year for a Csb (warm-summer Mediterranean, e.g., Seattle) climate from European Commission [5]. The state vector includes five temperature states and, for price-aware agents, five forecasted electricity prices (current, +6 h, +12 h, +18 h, +24 h),

where the electricity price data are taken from Bundesnetzagentur [3].

Action and Agent Design The control input is a normalized heating modulation $\alpha_t \in [0, 1]$. For DQN, we use the discrete set $\{0, 0.25, 0.5, 0.75, 1.0\}$, and PPO outputs continuous values in [0, 1]. Thus, "25%" means 25% of the nominal heating actuation in the model, not a direct supply-temperature setpoint. The action enters the building dynamics through the input vector $u_t = [\alpha_t, T_{amb,t}, I_{solar,t}]^\top$ and is translated by the identified matrix B_i into state changes. In this first study, the heat pump is modeled as an ideal modulating actuator with bound constraints only ($0 \leq \alpha_t \leq 1$). We do not model start-up/shut-down transients, minimum on/off times, ramp-rate limits, defrost cycles, domestic hot-water draw, or hydronic network details (e.g., radiator-level flow balancing). In practical deployment, α_t should be converted by a lower-level controller into feasible equipment setpoints (e.g., supply temperature and valve positions). Further details can be found in Appendix A.

Reward Function As a simple baseline, one could penalize deviation from a fixed temperature target using a quadratic loss:

$$r_t = -(T_{in,t} - T_{target,t})^2, \tag{1}$$

where $T_{in,t}$ is the indoor air temperature and $T_{target,t}$ is the target temperature (21° C in the fixed-target setting). Defining the temperature error as $e_t = T_{in,t} - T_{target,t}$, the comfort term can be written with two variants:

- Piecewise-quadratic:

$$f_{quad}(e_t) = \begin{cases} (e_t + T)^2 & e_t < -T, \\ 0 & -T \leq x \leq T, \\ (e_t - T)^2 & T < e_t \end{cases}$$

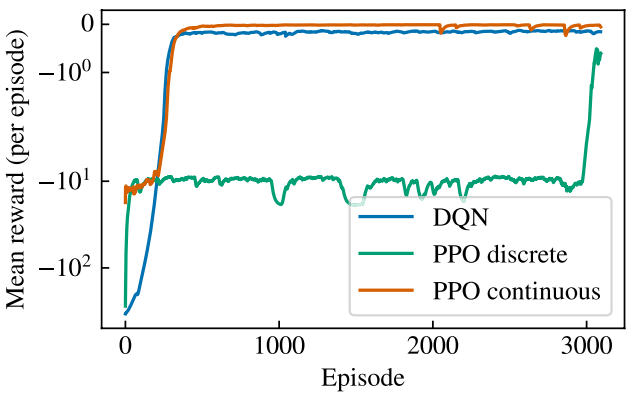

Fig. 2 Smoothed mean training reward over episodes for a DQN agent and two PPO agents using fixed temperature as target

- Piecewise-linear:

$$f_{linear}(e_t) = \begin{cases} \sigma/2 - e_t - T - \sigma & e_t < -T - \sigma, \\ (e_t + T)^2/(2 \cdot \sigma) & -T - \sigma \leq e_t < -T, \\ 0 & -T \leq x \leq T, \\ (e_t - T)^2/(2 \cdot \sigma) & T < e_t \leq T + \sigma, \\ \sigma/2 + e_t - T - \sigma & T + \sigma < e_t, \end{cases}$$

where the threshold $T = 1$ defines the comfort interval as [20, 22]° C, and σ denotes the transition width. The piecewise-quadratic function is shown in Fig. 7 compared to the piecewise-linear function.

The full reward is defined as a weighted sum of the temperature and price terms:

$$r_t = -w_t \cdot f_{\{quad,linear\}}(e_t) - w_p \cdot p_t \cdot \alpha_t, \tag{2}$$

where w_t and w_p are the weights for the temperature and price terms, p_t denotes electricity price, and α_t is the normalized heating action. The piecewise-linear function is designed to provide a sharper transition between comfort and discomfort, while the piecewise-quadratic function emphasizes larger penalties for significant deviations from the target temperature. Further information on the training is provided in Appendix B.

4 Experiments and Results

Fixed Temperature Control We compare three agents: a DQN agent with discrete actions, a PPO agent with discrete actions, and a PPO agent with continuous actions. The goal is to maintain a fixed indoor temperature of 21° C while minimizing temperature deviations and maximizing rewards based on the reward function defined in Eq. (1). The agents are trained for up to 3,000 episodes, with each episode spanning 16,166 time steps (equivalent to 168 days of real-time data). The DQN agent uses a replay buffer and target network for stability, while the PPO agents use a policy gradient approach with entropy regularization to encourage exploration. Figure 2 shows the smoothed mean training reward over episodes for the DQN agent and the two PPO agents using a fixed temperature as the target. The DQN agent and the continuous PPO agent improve much faster than the discrete PPO agent. The continuous PPO agent achieves the best performance with the highest reward, while the discrete-action PPO agent shows slower convergence and lower overall performance. These results suggest that continuous control enables more precise temperature regulation, while the DQN agent, with its simpler structure, still outperforms PPO in the discrete setting.

Figure 3 shows the indoor air temperatures achieved by the different methods over a whole episode compared to the

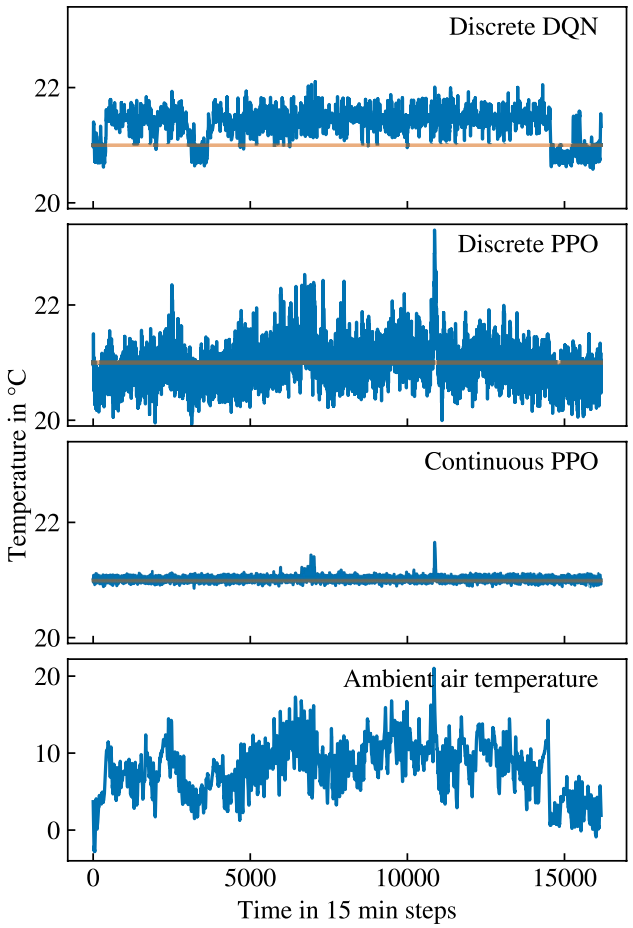

Fig. 3 Indoor air temperatures achieved by different methods over a whole episode compared to ambient air temperature. Orange line indicates the setpoint temperature 21° C

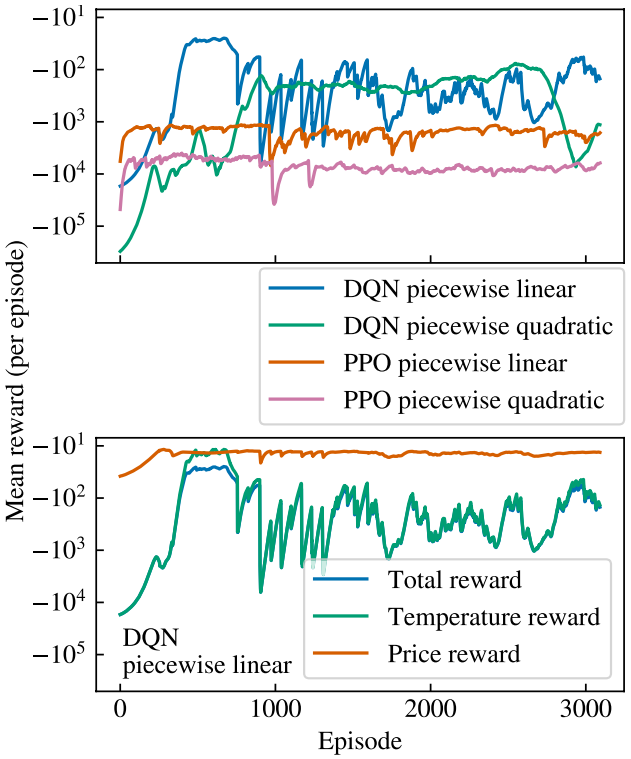

Fig. 4 Smoothed mean training reward over episodes across two DQN agents and two PPO agents trained with weighted sum reward settings

ambient air temperature. The DQN agent and the continuous PPO agent maintain the target temperature more closely than the discrete PPO agent, which exhibits larger fluctuations. The DQN agent's discrete action space allows it to make more significant adjustments, while the continuous PPO agent can fine-tune its actions for better temperature regulation.

Temperature-band and Price-Aware Control We extend the experiments with price-aware control using short-term electricity price forecasts. We compare four agents: two DQN (discrete action space) and two PPO (continuous action space) agents, each with one of the piecewise-quadratic and piecewise-linear reward function variants. The reward function combines temperature penalties and energy costs, as defined in Eq. (2). The weight of the temperature term w_t is set to a high value of $w_t = 1,000$, ensuring that comfort is prioritized when violated. The weight of the price term is set to $w_p = 1$. The temperature term of the reward function is designed to overrule the price term whenever the indoor air temperature exceeds the comfortable interval.

Figure 4 shows the smoothed mean training reward curves for the four agents. The upper plot shows the total reward, while the bottom plot decomposes the reward into temperature and price components for the DQN agent with piecewise-linear comfort reward, which is the best-performing combination. The DQN agent with piecewise-linear comfort reward achieves the highest average reward, indicating better performance in balancing comfort and cost. The PPO agents show slower convergence and lower overall performance compared to the DQN agents, particularly the PPO agent with piecewise-quadratic comfort reward, which struggles to maintain comfort while minimizing costs. In the decomposition reward terms, the total reward (blue) closely follows the higher weighted temperature term (green) as long as it is higher than the price term (orange). Initially, the temperature term dominates due to its high weight. The best model is achieved in episode 369. As training progresses and the agent learns to stay within the comfort range, the temperature term approaches 0 and the agent is able to keep the temperature within the specified bounds. For this, the price term components become the focus and take over to minimize the cost of heating.

Action analysis Fig. 5 compares the behavior of the piecewise-linear DQN variant over three days. The upper plot shows the resulting indoor temperature (blue) next to the temperature achieved by MPC (orange) with the

comfort zone in orange area defined by the reward function. The bottom plot shows the learned heating action (pink) compared to the action taken by MPC (blue) control and the electricity price (green) at each timestep. We observe that the DQN agent learns to anticipate price changes and adjust its heating strategy accordingly. In the first half of the period, when prices are low, the agent preheats the building and maintains a higher indoor temperature (around 21° C) to prepare for potential price spikes. It switches off the heat pump when prices spike, allowing the indoor temperature to fall within the comfort zone. In the second half, when a price drop is forecasted, it allows the indoor temperature fall to the lower comfort limit (around 20° C), and reheats during the zero-price period, demonstrating learned cost minimization behavior. However, it stops heating when indoor temperatures reach 22° C at the upper boundaries of the comfort zone.

The MPC more precisely exploits the allowed temperature range to minimize cost, maintaining the temperature at the boundaries of the comfort zone, while the DRL agent tends to stay closer to the midpoint (21° C). This indicates that MPC is able to exploit situations of very high or very low prices better than the RL agent resulting in the performance gap. The temperatures achieved by the RL agent are generally closer to 21° C because the agent is not as good as the MPC to fully exploit the temperature constraints.

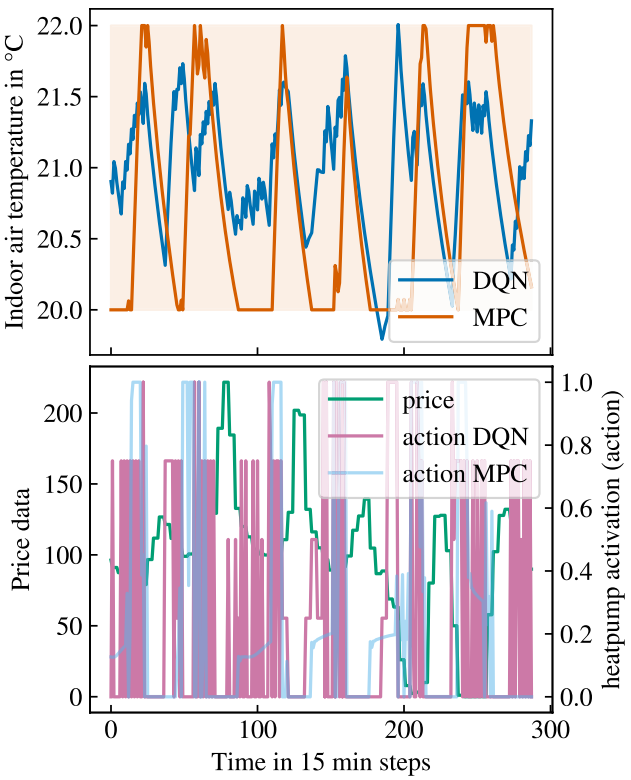

Fig. 5 Upper: Indoor air temperature; Bottom: learned action and energy price over the course of three days for a DQN agent trained with the piecewise linear function in comparison to MPC

Generalization on Unseen Data To assess generalization, we evaluate the best-performing agent on a separate validation dataset that was not used for training. This validation set contains different weather and electricity price trajectories; the input profiles are shown in Appendix C in Figs. 10 and 11. Figure 6 shows the resulting control behavior on this unseen dataset. The learned policy preserves the key strategy observed on the training scenario: preheating before high-price periods, reduced heating during price peaks, and operation within the comfort band of 20° C to 22° C. This indicates that the learned control policy transfers to changed exogenous conditions rather than only fitting one specific time series.

Comfort and Cost Analysis For comfort comparison, we use the deviation from 21° C relative to MPC. Table 1 reports this deviation metric together with total heating cost on the held-out test dataset. For RL methods, the reported values are mean ± standard deviation across 3 independent training runs (different random seeds). MPC and the P-controller are deterministic baselines evaluated once. The total cost is calculated as: $\text{total cost} = \sum_{t=0}^{n} p_t \cdot \alpha_t$,

where p_t is the electricity price at time step t and α_t is the action taken by the agent at time step t. The total cost is normalized to the MPC agent (in bold), which is set to 100%. For the case considering both temperature comfort and price-aware control, the DQN agent with piecewise-linear method (in bold) achieves the lowest total cost, closely approaching the MPC agent. The DQN agent with piecewise-quadratic comfort reward shows higher costs, indicating that the quadratic penalty may lead to over-penalization of temperature deviations, resulting in less efficient heating strategies. The PPO agents, particularly the piecewise-quadratic variant, exhibit higher costs due to slower convergence and less effective price-aware control strategies. The P-controller serves as a baseline, showing significantly higher costs due to its inability to adapt to dynamic price signals and reliance on fixed rules.

5 Discussion

In fixed temperature control experiments, both DQN and PPO are evaluated using discrete action spaces, with PPO additionally being tested in a continuous setting. The continuous PPO agent achieves the best performance when solely maintaining a target temperature. However, after extending the objective to also account for electricity cost using a price forecast, DQN with a discrete action space outperformed PPO. For this reason, the DQN agent using a piecewise-linear reward function is selected for subsequent

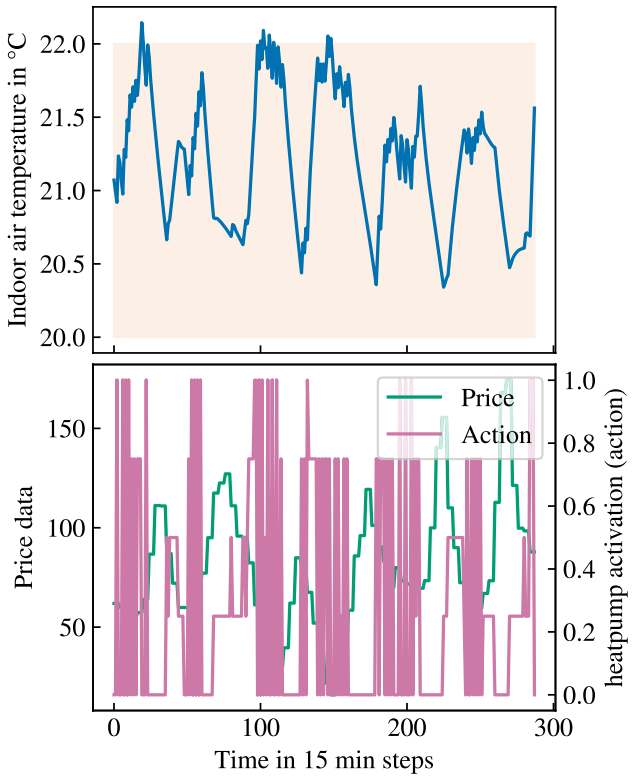

Fig. 6 Validation behavior of the a DQN agent on unseen weather and price data. The controller keeps indoor temperature within the comfort band and anticipates upcoming price changes

Table 1 Performance comparison of the different methods across 3 independent training runs

Method	Deviation from 21° C relative to MPC	Heating cost relative to MPC
P-Controller	2.7%	228%
MPC	**100%**	**100%**
PPO without price forecast	93% ± 509%	158% ± 276%
DQN piecewise linear	**129%**± 1780%	**143%**± 642%
DQN piecewise quadratic	149%± 287%	178%± 261%
PPO piecewise linear	522%± 479%	305%± 285%
PPO piecewise quadratic	534%± 568%	327%± 319%

evaluations, including benchmarking against MPC. Accordingly, "best performance" in this work refers to the best-performing DRL configuration (comfort–cost trade-off) rather than outperforming MPC in absolute cost.

Compared to MPC, the RL agent achieves comparable but not equal performance. Specifically, the total heating cost for a full episode is approximately 29% higher than that of MPC. Although MPC remains the gold standard in terms of optimality, assuming accurate building models, its reliance on detailed system identification and repeated real-time optimization makes it computationally expensive and less scalable [1]. In our setting, this means MPC must solve a constrained optimization problem at every control step over a multi-step prediction horizon, whereas DRL inference requires only one neural-network forward pass per step. The DRL computational burden is shifted mainly to offline training. In contrast, once training is completed, online DRL control remains lightweight, effectively independent of the optimization horizon length, and adaptable. Here, "adaptability" means that a fixed trained policy can react online to changing observed conditions such as weather, thermal state, and price forecasts without re-solving an optimization problem. If long-term operating conditions drift substantially, adaptation would require periodic offline retraining or fine-tuning of newly collected data rather than continuous online learning.

The design of the reward function has been shown to be critical to training success. Both DQN and PPO perform better with the piecewise-linear variant of the comfort penalty compared to the piecewise-quadratic version. Although the quadratic penalty increases more rapidly with distance from the target temperature, the linear variant introduces a sharper transition at the edges of the comfort zone (e.g., 20–22° C). This means deviations are penalized more strongly, which results in better adherence to the comfort interval and higher cumulative rewards. The sharper transition appears to provide a clearer learning signal, facilitating faster convergence and more consistent control behavior.

Despite the theoretical advantages of continuous action spaces, such as finer control, discrete actions often yield more robust training outcomes. In this study, discrete DQN agents demonstrate greater stability and reproducibility, likely due to simpler policy structures and reduced variance in action selection during exploration. A discrete action space with five levels is used in all experiments. While this discretization limits the agent's resolution compared to continuous control, it is effective both in comfort and price-aware scenarios. Increasing the number of discrete levels could approximate continuous control more closely, but at the cost of increased policy complexity and longer training time. Conversely, reducing the number of actions could simplify control policy without significantly compromising performance, especially in real-world settings where precise actuation may not be critical. Future work could explore this trade-off more systematically.

6 Conclusion and Outlook

This work explores the use of DRL for adaptive, cost-aware control of residential heat pumps. By comparing DQN and PPO across discrete and continuous action spaces, we demonstrated that a DQN agent with a discrete action space and a piecewise-linear comfort reward achieves the best trade-off between thermal comfort and energy cost. While

DRL does not yet surpass the performance of MPC, it offers several practical advantages, including model-free training and real-time inference. In this context, adaptability refers to responsive decision-making under varying inputs with a fixed trained policy; updating the policy itself would require additional fine-tuning or retraining when the operating regime changes substantially.

The results highlight the importance of reward function design, particularly the choice between piecewise-linear and piecewise-quadratic comfort penalties. The piecewise-linear function provided a clearer learning signal, enabling faster convergence and more consistent control behavior. The discrete action space also proves to be more robust than continuous control in this context. DRL thus presents a scalable and flexible alternative to traditional control strategies in smart building applications. This further lays the groundwork for model-free approaches of optimized DSM, given an adequate price or control signal for the RL agent to optimize against.

Future work will explore systematic tuning of reward weights, extension to multi-zone building models, and further improvements through advanced DRL algorithms such as Soft Actor-Critic or hybrid model-based approaches. We will further implement transfer learning to verify whether the learned knowledge could be transferred to other buildings and reduce the training time. Real-world validation and deployment remain essential next steps to bridge the gap between simulation and practical implementation.

Building Model Dynamics

To evaluate the proposed controller under realistic and diverse thermal conditions, we utilize a library of linear state-space building models from Vallianos et al. [19]. The models were identified from large-scale smart thermostat measurements including approximately 60,000 residential buildings. For each building i, thermal behavior is represented by the discrete-time tuple (A_i, B_i, C_i). The state update and observation equations are

$$\begin{aligned} x_{t+1} &= A_i x_t + B_i u_t, \\ o_t &= C_i x_t, \end{aligned} \tag{3}$$

where $x_t \in \mathbb{R}^5$ is the latent thermal state and o_t is the measured output.

State Vector The state vector captures the dominant thermal inertia of the building and consists of five temperatures:

1. Indoor air temperature (T_{in})
2. Interior thermal mass temperature (T_{int_mass})
3. Envelope thermal mass temperature (T_{env_mass})
4. Heater component temperature (T_{heater})
5. Sensor casing temperature (T_{sensor})*Input Vector* The input $u_t \in \mathbb{R}^3$ combines control action and disturbances:

$$u_t = [\alpha_t, T_{amb,t}, I_{solar,t}]^\top, \tag{4}$$

with heating modulation $\alpha_t \in [0, 1]$, ambient outdoor temperature $T_{amb,t}$, and solar irradiance $I_{solar,t}$. Here, α_t is a normalized heat-input command (0%–100%) that scales the nominal actuation level represented in the identified model matrices.

For clarity, this abstraction assumes an ideal modulating heat source with instantaneous tracking of α_t. Equipment-level constraints (e.g., compressor cycling limits, ramping dynamics, and emitter/hydronic control) are not explicitly represented and are deferred to future work.

Output Mapping The controller observes indoor air temperature, so the output matrix is chosen as $C_i = [1, 0, 0, 0, 0]$, which yields $o_t = T_{in,t}$.

Additional Training Details

Table 2 summarizes the main training hyperparameters used for the DQN and PPO agents.

The piecewise-linear and piecewise-quadratic reward functions are defined in the main text, and their shapes are shown in Fig. 7. The piecewise-linear function provides a sharper transition at the comfort boundaries, while the piecewise-quadratic function penalizes larger deviations more heavily. The choice of reward function significantly

Table 2 Main hyperparameters used for training the DRL agents

Hyperparameter	DQN	PPO
Policy / network type	MLP policy	MLP policy
Hidden-layer architecture	[64, 64]	[64, 64]
Activation function	ReLU	ReLU
Learning rate	1×10^{-4}	3×10^{-4}
Batch size	32	64
Discount factor γ	0.99	0.99
Replay buffer size	10^6	n/a
Gradient steps per update	1	n/a
Target network update interval	10,000 steps	n/a
Exploration schedule	linear decay, ϵ: 1.0 $\rightarrow$ 0.05 over 10% of training	n/a
Rollout length n_{steps}	n/a	2048
Epochs per update	n/a	10
GAE parameter λ	n/a	0.95
Clipping range	n/a	0.2
Entropy coefficient	n/a	0.0
Value-function coefficient	n/a	0.5
Max gradient norm	10	0.5

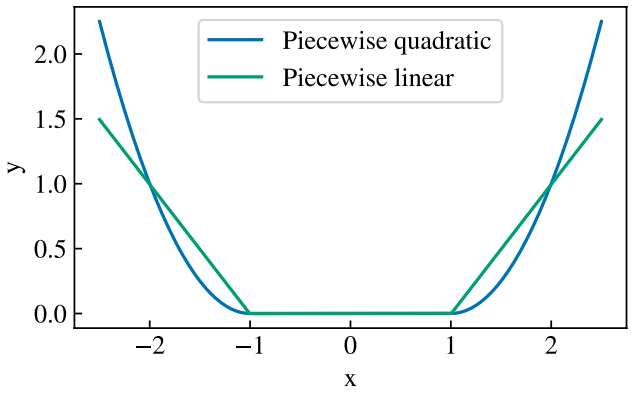

Fig. 7 The piecewise-quadratic f_{quad} temperature term function compared to the piecewise-linear f_{linear} temperature term function

impacts the learning process, with the piecewise-linear variant leading to better performance in our experiments due to its clearer learning signal around the comfort zone.

Validation data

Figures 8 and 9 show the input profiles of the training dataset used in Sect. 4.

Figures 10 and 11 show the exogenous input profiles of the unseen validation dataset used in the generalization experiment described in Sect. 4. These trajectories were not seen during training and differ from the training weather and price data.

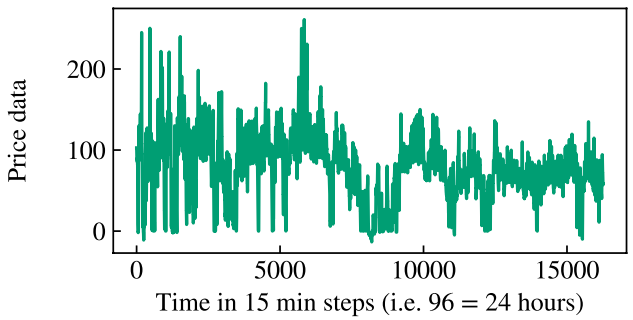

Fig. 9 Electricity price profile (in €/MWh) of the training data from Csb climate zone

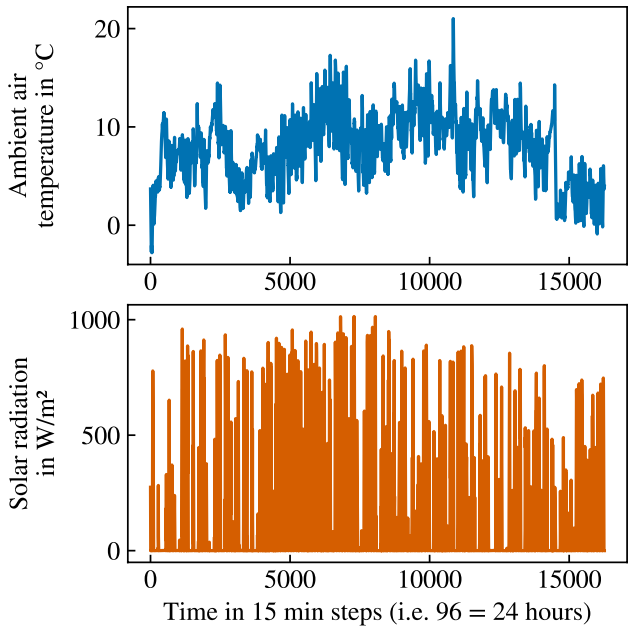

Fig. 8 Weather profile of the training data from Csb climate zone

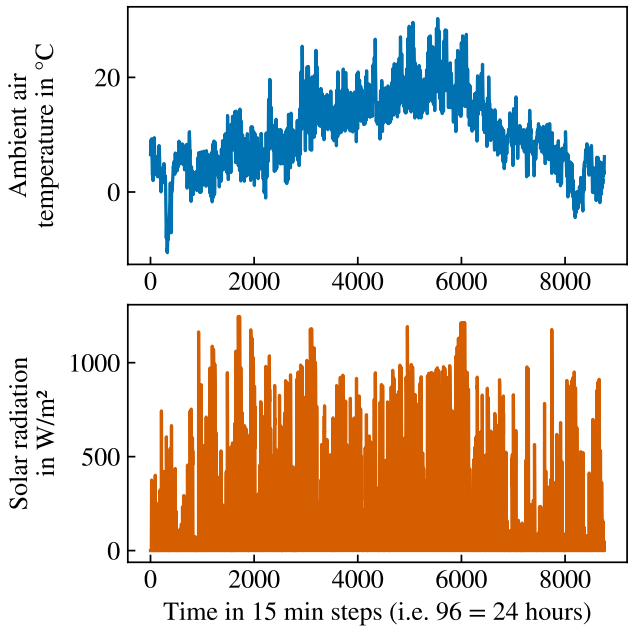

Fig. 10 Weather profile of the unseen validation dataset

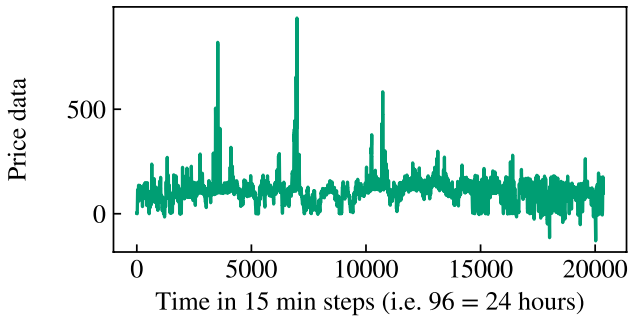

Fig. 11 Electricity price profile (in €/MWh) of the unseen validation dataset

Acknowledgements We gratefully acknowledge funding from the Helmholtz Association under grant No. VH-NG-1727, the Networking Fund through Helmholtz AI, and within the framework of the Program-Oriented Funding POF IV in the program Energy Systems Design (ESD, project number 37.12.01)

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The linear state-space environment model is derived from a publicly available dataset in the Zenodo repository: https://doi.org/10.5281/zenodo.8347091. The Typical Meteorological Year climate data for Csb (warm-summer Mediterranean) climate are available from the Photovoltaic Geographical Information System (PVGIS) at https://re.jrc.ec.europa.eu/pvg_tools/de/tools.html. Electricity price data are available from SMARD at https://www.smard.de/home/downloadcenter/download-marktdaten/. All datasets used in this study are publicly accessible; no proprietary or restricted data were employed.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Arroyo J, Spiessens F, Helsen L (2020) Identification of multi-zone grey-box building models for use in model predictive control. J Build Perform Simul 13(4):472–486
2. Brandi S, Piscitelli MS, Martellacci M, Capozzoli A (2020) Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energy Build 224:110225. https://doi.org/10.1016/j.enbuild.2020.110225
3. Bundesnetzagentur (2025) Smard. https://www.smard.de/home/downloadcenter/download-marktdaten/
4. Coraci D, Brandi S, Hong T, Capozzoli A (2023) Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. Appl Energy 333:120598. https://doi.org/10.1016/j.apenergy.2022.120598
5. European Commission JRC (2025) Photovolatic geographical information system. https://re.jrc.ec.europa.eu/pvg_tools/de/tools.html
6. Han G, Joo HJ, Lim HW, An YS, Lee WJ, Lee KH (2023) Data-driven heat pump operation strategy using rainbow deep reinforcement learning for significant reduction of electricity cost. Energy 270:126913. https://doi.org/10.1016/j.energy.2023.126913
7. Kadamala K, Chambers D, Barrett E (2024) Enhancing hvac control systems through transfer learning with deep reinforcement learning agents. Smart Energy 13:100131. https://doi.org/10.1016/j.segy.2024.100131
8. Kohlhepp P, Harb H, Wolisz H, Waczowicz S, Müller D, Hagenmeyer V (2019) Large-scale grid integration of residential thermal energy storages as demand-side flexibility resource: a review of international field studies. Renew Sustain Energy Rev 101:527–547
9. Langer L, Volling T (2022) A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems. Appl Energy 327:120020
10. Lissa P, Schukat M, Keane M, Barrett E (2021) Transfer learning applied to drl-based heat pump control to leverage microgrid energy efficiency. Smart Energy 3:100044. https://doi.org/10.1016/j.segy.2021.100044
11. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
12. Nagy Z, Henze G, Dey S, Arroyo J, Helsen L, Zhang X, Chen B, Amasyali K, Kurte K, Zamzam A et al (2023) Ten questions concerning reinforcement learning for building energy management. Build Environ 241:110435
13. Rohrer T, Frison L, Kaupenjohann L, Scharf K, Hergenröther E (2023) Deep reinforcement learning for heat pump control. In: science and information conference, Springer, pp 459–471
14. Ruelens F, Iacovella S, Claessens BJ, Belmans R (2015) Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. Energies 8(8):8300–8318. https://doi.org/10.3390/en8088300
15. Ruelens F, Claessens BJ, Vandael S, De Schutter B, Babuška R, Belmans R (2017) Residential demand response of thermostatically controlled loads using batch reinforcement learning. IEEE Trans Smart Grid 8(5):2149–2159. https://doi.org/10.1109/TSG.2016.2517211
16. Schmitz S, Brucke K, Kasturi P, Ansari E, Klement P (2024) Forecast-based and data-driven reinforcement learning for residential heat pump operation. Appl Energy 371:123688. https://doi.org/10.1016/j.apenergy.2024.123688
17. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347
18. Urieli D, Stone P (2013) A learning agent for heat-pump thermostat control. In: proceedings of the 2013 international conference on autonomous agents and multi-agent systems, pp 1093–1100
19. Vallianos C, Candanedo J, Athienitis A (2024) Thermal modeling for control applications of 60,000 homes in north america using smart thermostat data. Energy Build 303:113811
20. Wei T, Wang Y, Zhu Q (2017) Deep reinforcement learning for building hvac control. In: proceedings of the 54th annual design automation conference 2017, pp 1–6