Advanced Deep Reinforcement Learning for Heat Pump Control in Residential Buildings

Gökhan Demirel[1], Ömer Ekin[1], Jianlei Liu[1], Luigi Spatafora[1], Kevin Förderer[1], Veit Hagenmeyer[1]
[1]Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology, Germany
{goekhan.demirel, oemer.ekin, jianlei.liu, luigi.spatafora, kevin.foerderer, veit.hagenmeyer}@kit.edu

***Abstract*—Residential heating is a significant c ontributor to carbon emissions. Replacing conventional on/off and heating curve controls with smart strategies is essential for decarbonization. This paper presents eight state-of-the-art control strategies for residential air-source heat pumps in the open-source environment LLECBuildingGym, which emulates the heat pump house at the Living Lab Energy Campus (LLEC). We compare three rule-based controllers (fuzzy, PI, and PID), a model-predictive controller (MPC), and four advanced deep reinforcement learning (RL) algorithms (A2C, DDPG, PPO, and SAC) in a 1R1C thermal building model with continuous heating and cooling control. The model captures nonlinear thermal dynamics using Euler discretization, models sensor uncertainties as reflected Wiener processes and integrates dynamic electricity tariffs. We define s ingle-objective (temperature) a nd m ulti-objective tasks that minimize thermal discomfort and energy costs. An extensive ablation study identifies t he b est p erforming R L algorithm configuration t hat r educes c ost b y 6 % c ompared t o rule-based controllers, outperforms MPC by 1% and underperforms MPC with perfect prediction by less than 4%.**

***Index Terms*—Heat Pump, Reinforcement Learning, Model Predictive Control, Fuzzy control, Energy Management**

I. Introduction

Decarbonizing residential heating and cooling introduces significant c hallenges t o e nergy s ystems d ue t o t he rapid adoption of electric heat pumps (HP) [1]. With residential heating accounting for 18% of the UK's greenhouse gas emissions, there is an urgent need to develop and implement efficient h eating t echnologies [2]. A r ecent a nalysis o f the German heating sector also recommends a hybrid solution using HPs for well-insulated single-family homes and district heating in densely populated urban areas [3]. In the United States, buildings consume 75% of electricity and 40% of total energy, underscoring their key role in the energy transition [4].

Current HP control strategies depend on static heating curves that directly convert outdoor temperatures into heating actions [5]. This simplified a pproach o verlooks d ynamic factors such as weather forecasts, user behavior, building aging, renovation effects, and fluctuating e lectricity p rices, leading to suboptimal performance. Effective control mechanisms aim to combine comfort for occupants, energy efficiency, and dynamic tariffs. Dealing with these multi-objectives requires intelligent control systems adapting to external signals and internal system states. To address these points, this work investigates the following research questions:

1) What influence do different observation space configurations and reward settings have on the learning behavior and performance of reinforcement learning (RL) agents?
2) What is the performance of different control methods from proportional-integral (PI) and proportional-integral-derivative (PID) to fuzzy control, model predictive control (MPC), and data-driven RL algorithms in maintaining the indoor temperature while minimizing the energy costs in HP operation?

To answer these questions and facilitate future research, we propose a comprehensive control portfolio including rule-based, model-based, and data-driven approaches, whereby a perfect-foresight MPC serves as the optimal ground-truth solution. The key contributions of this work are:

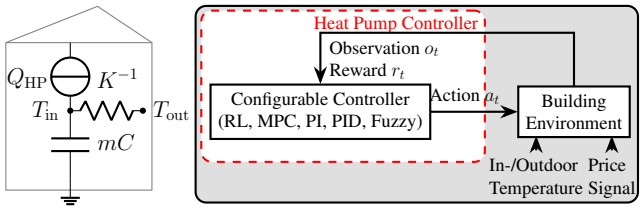
- We formulate heat pump control as a Partially Observable Markov Decision Process (POMDP) within our open-source *LLECBuildingGym*[1] environment that simulates nonlinear 1R1C building dynamics, outdoor disturbances, and dynamic EPEX spot market tariffs.
- We conduct a systematic RL ablation study to determine the most effective observation and reward configurations for residential heat pump systems. In addition, we provide a comprehensive comparison of state-of-the-art control strategies. For reinforcement learning, we use StableBaselines3 [6] to implement four advanced deep RL algorithms. We also evaluate a MPC approach implemented in Pyomo [7] using the Interior Point OPTimizer (`IPOPT`) solver [8], alongside fuzzy, PID, and PI controllers.

The remainder of this paper is structured as follows: Section II reviews state-of-the-art HP control. Section III introduces the problem formulation, while Section IV details the environment and controller implementation. Section V presents the experimental setup (V-A) and the controller evaluation conducted as part of the ablation study (V-B). Finally, Section VI concludes this paper and outlines directions for future research.

II. Related Work

Numerous approaches have been proposed in building automation to control thermal comfort and economic costs, which inherently demands multidisciplinary research [9], [10]. Existing control strategies can be categorized into rule-based methods (e.g., PI, PID, or fuzzy control), model-based methods (e.g., MPC), and learning-based methods (e.g., RL). MPC

[1]Code available at: https://github.com/KIT-IAI/LLECBuildingGym.



(a) (b)

Fig. 1: Lumped thermal model of a residential building used for HP control in the `LLEC-HeatPumpHouse-1R1C` environment. (a) 1R1C thermal–electrical equivalent circuit. (b) Modular controller–environment interaction for RL (single- and multi-objective), MPC, PI, PID, and Fuzzy controllers.

requires an accurate mathematical description of the building's thermal dynamics to perform optimally [11]. Similarly, rule-based controllers depend on user-defined rules often designed for specific building types and operating conditions [12]. Consequently, both categories have difficulty scaling and adapting to different building types. In contrast, RL is a data-driven, often model-free method that learns building dynamics through interaction with the environment, removing the need for detailed physical models. In [13], a Q-learning-based heating, ventilation, and air conditioning (HVAC) controller achieves 10% energy savings while maintaining thermal comfort and outperforming a day-night rule-based control strategy. The authors of [14] compared a Deep Q-Network (DQN) HVAC controller with an on/off rule-based baseline that uses fixed temperature thresholds. It achieved 11% energy savings. Moreover, [15] used a Deep Deterministic Policy Gradient (DDPG) for continuous temperature and humidity control, outperforming DQN and Q-learning in efficiency but without comparing it to a rule-based controller. Similarly, [16] showed that Proximal Policy Optimisation (PPO) can approach MPC-level performance for HP control with load shifting, and [17] demonstrated that a customized DQN controller yielded about 6% cost savings relative to a threshold-based controller.

Many existing RL approaches focus on simple rule-based controllers, leaving a gap in comparing feedback control (PID, PI, fuzzy) and model-based MPC methods with advanced RL control for HPs. This paper addresses that gap by systematically comparing these controllers for residential buildings under identical conditions and provides an open-source benchmarking environment for more comparability.

III. Problem Formulation

Throughout this present paper, subscripts denote discrete timesteps or temporal horizons (e.g., prediction and control horizons). Predicted values are written as $\hat{a}$, measured values as $\tilde{a}$; scalars are represented by a, and vectors by $\mathbf{a}$.

Residential HP control is modeled as a sequential decision-making problem under uncertainty and modeled as a POMDP. The POMDP is defined by the 7-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, \mathcal{T}_0, R, \gamma)$, where both the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are continuous. The transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ describes the probability transitioning from state $s_t \in \mathcal{S}$ to $s_{t+1} \in \mathcal{S}$ for the action $a_t \in [-1, 1]$; positive actions for heating and negative values for cooling. The initial-state distribution is denoted by $\mathcal{T}_0$, and $\gamma \in [0, 1)$ is the discount factor. The latent state s_t comprises physical and contextual variables relevant for HP control. At each timestep t, the controller receives a noisy observation $o_t = s_t + \varepsilon_t \in \mathcal{O}$, and a scalar reward $R : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is returned by the environment. The noise ε_t is distributed as $\varepsilon_t \sim \mathcal{N}(0, 0.01)$. Given a noisy partial observation o_t of the latent state s_t, the controller selects an action $a_t \in \mathcal{A} = [-1, 1]$. The agent aims to learn the optimal control policy π^*:

$$\pi^*(a \mid \mathbf{o}) = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{H} \gamma^t R_t\Big]. \tag{1}$$

A trajectory τ within an episode is defined as the sequence of states, observations, actions, and rewards up to the terminal timestep H. The goal is to maximize the expected return over one episode, given by the cumulative discounted reward $\sum_{t=0}^{H} \gamma^t R_t$. In a stochastic environment, the trajectory distribution $\mathcal{T}(\tau \mid \pi)$ represents the probability of observing trajectory τ under the policy π, starting from an initial state s_0 with probability $\mathcal{T}(s_0)$:

$$\mathcal{T}(\tau \mid \pi) = \mathcal{T}(s_0) \prod_{t=0}^{H} \mathcal{T}(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid o_t). \tag{2}$$

In practice, the policy π is typically represented by a parameterized function approximator, such as a neural network with weights θ, resulting in a policy $\pi_\theta(a \mid \mathbf{o})$.

IV. LLECBuildingGym Environment

This section introduces the LLECBuildingGym environment, a unified testbed for evaluating HP control strategies in residential buildings. It covers the underlying HP house model (IV-A), the implemented controller portfolio (IV-B), and the underlying reward design to decision-making (IV-C).

A. Heat Pump House Environment Model

Figure 1a illustrates the single-zone building model, which is based on a first-order resistance–capacitance (1R1C) thermal model. In the thermal–electrical analogy, T_{in} and T_{out} represents the indoor and the outdoor temperature correspond to node voltages, whereas heat flows are analogous to electrical currents. The thermal heat capacity in the building mass and air is lumped into a capacitor mC; the overall transmittance of walls and windows is represented by a resistance K^{-1}; and a controllable current source injects (or extracts) heat controlled by the HP. Kirchhoff's current law is applied to the indoor temperature node, assuming a positive sign convention for heat inflow. Discretizing the resulting energy balance using a fixed timestep Δt with the forward Euler method yields the state transition equation of the environment:

$$T_{\text{in},t+\Delta t} = T_{\text{in},t} - \frac{\Delta t}{mC}\left[K\left(T_{\text{in},t} - T_{\text{out},t}\right) - a_t Q_{\text{HP}}^{\max}\right], \tag{3}$$

where $Q_{\text{HP}}^{\max}$ denotes the maximum thermal power of the HP.

TABLE I: Observation space overview.

#	Variable	Description	Length
1	$\tilde{T}_{\text{in}} - T_{\text{set}}$	Noisy indoor temp. deviation from setpoint.	(1)
2	time_of_day	Normalized timestamp $[0, 1]$	(1)
3	a_{t-1}	Previous action$[-1, 1]$	(1)
4	p_t	Current normalized energy price $[0, 1]$	(1)
5	$\hat{\boldsymbol{p}}_{t:t+n_p}$	Prediction normalized energy prices $[0, 1]$	(n_p)

B. Controller Portfolio

This section presents the implemented controllers and their integration. At each timestep, each controller maps the observation o_t to an action $a_t \in [-1, 1]$ based on its control, as shown in Fig. 1b. The observation space includes temperature deviation, time of day, previous action, current energy price, and forecasted prices (see Table I).

- **Reinforcement Learning (RL):** In the training phase, a model-free agent learns a policy $\pi_\theta(a_t \mid o_t)$, parameterized by neural network weights θ, by interacting with the environment in order to maximize the expected cumulative reward defined in Eq. (1). The agent receives partial observations $o_t \in \mathcal{O}$ (see Table I) and selects actions $a_t \in [-1, 1]$ to control the HP, receiving a scalar reward $R_t \in \mathbb{R}$ and a new partial observation at each timestep. In the test phase, the learned policy is executed in a feed-forward manner to determine actions without exploration noise.
- **Model Predictive Control (MPC):** The MPC controller solves a constrained optimization problem over a finite prediction horizon n_p. At each timestep t, a sequence $\{a_t, \ldots, a_{t+n_p-1}\}$ is computed by minimizing a cost objective based on predicted states and prices:

$$\min_{\{a_t\}} \sum_{t=1}^{n_p} w_{\text{temp}} \cdot |T_{\text{in},t} - T_{\text{set}}| + w_{\text{econ}} \cdot \hat{p}_t \cdot |a_t Q_{\text{HP}}^{\max}|. \quad (4)$$

  Only the first control action a_t is applied, and the problem is re-solved at the next timestep. After applying the solution $\boldsymbol{a}_1$ at the initial timestep, the time horizon is shifted forward and the process repeats.
- **PI Controller:** Implements a control law that combines a proportional and an integral term:

$$a_t = K_p e_t + K_i \sum_{i=0}^{t} e_i \, \Delta t, \quad (5)$$

  where $e_t = T_{\text{set},t} - T_{\text{in},t}$ is the temperature error between the setpoint and the current indoor temperature. The integral term accumulates the historical error to eliminate steady-state deviations. Since the controller operates in discrete time, the integral $\int_0^t e(\tau)\, d\tau$ is approximated by the sum $\sum_{i=0}^{t} e_i \cdot \Delta t$ using the rectangle rule.
- **PID Controller:** Extends the PI controller by incorporating a derivative term to improve the response to sudden changes:

$$a_t = \underbrace{K_p e_t}_{\text{P-term}} + \underbrace{K_i \sum_{i=0}^{t} e_i \, \Delta t}_{\text{I-term}} + \underbrace{K_d \frac{e_t - e_{t-1}}{\Delta t}}_{\text{D-term}}. \quad (6)$$

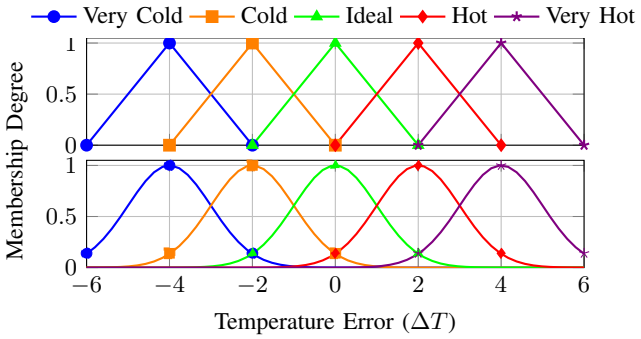


Fig. 2: Membership functions for temperature error using triangular (top) and Gaussian (bottom) shapes.

  The derivative term approximates the rate of change of the error using a finite difference and helps reduce oscillations by reacting to rapid temperature fluctuations.
- **Fuzzy Controller:** Applies rule-based inference using fuzzy logic, providing an interpretable and robust alternative to conventional feedback controllers. The temperature error $e_t = T_{\text{set},t} - T_{\text{in},t}$ is mapped to linguistic categories such as "Very Cold", "Cold", "Ideal", "Hot", and "Very Hot" which are represented by triangular or Gaussian membership functions, as illustrated in Fig. 2. The control action is computed using centroid defuzzification:

$$a_t = -\frac{\sum_i \mu_i(e_t) \cdot \nu_i}{\sum_i \mu_i(e_t)}, \quad (7)$$

  where $\mu_i(e_t)$ is the membership degree of the error e_t, and ν_i is the numerical control value associated with fuzzy rule i. These linguistic categories are mapped to predefined control intensities within the normalized action space $[-1, 1]$. The negative sign reflects the environment's convention, where positive fuzzy outputs indicate cooling actions.

C. Reward Design

After executing the action a_t, the controller receives a scalar reward according to the selected reward mode. Three modes are considered:

1) Temperature-Based Reward (Comfort-Oriented):

$$R_1 = \exp\left(-|T_{\text{in},t} - T_{\text{set},t}|\right). \quad (8)$$

2) Economic Reward (Cost-Oriented):

$$R_2 = -\frac{p_t \cdot |a_t|}{p_{\max}}, \quad (9)$$

3) Combined Reward (Multi-Objective Setting):

$$R_t = w_{\text{temp}} R_1 + w_{\text{econ}} R_2, \quad (10)$$

   where w_{temp} and w_{econ} are weighting factors, p_t denotes the electricity price under a dynamic tariff, and T_{set} is the user-defined temperature setpoint.

These reward formulations form the basis for learning and are key elements to the ablation study, where reward modes and observation components are systematically removed to evaluate their influence on RL performance.

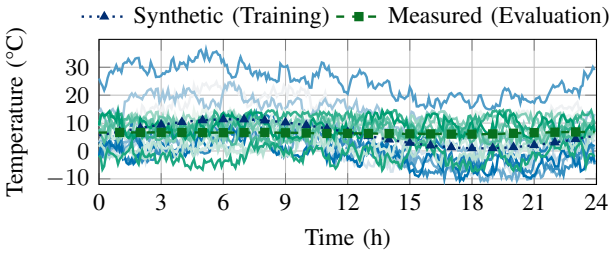


Fig. 3: Temperature data with noise-free synthetic (blue triangles) and measured (green squares) references, and corresponding Wiener-noisy training (blue) and evaluation (green).

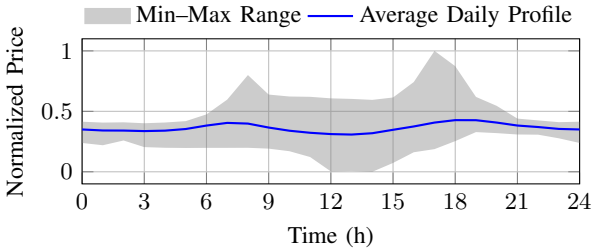


Fig. 4: Normalized daily electricity price profile based on EPEX Spot market prices for Austria (January–April 2025).

V. Evaluation

A. Experimental Setup

Experiments are conducted in a simulated single-zone building with HP control, modeled by a first-order energy balance and discretized with explicit Euler integration (see Section IV).

Figure 3 shows the training and evaluation datasets. The training set contains synthetic, class-based sinusoidal temperature profiles with reflected Wiener noise (±5 °C), shown as blue lines and mimicking daily outdoor temperature patterns. In contrast, the evaluation set contains 5-minute resampled measurements from the Living Lab Energy Campus (LLEC) [18] with identically parameterized Wiener noise added using fixed random seeds, shown as green lines. An ablation study systematically investigates the influence of reward design and observation sets on RL performance. Variants T01–T04 use the temperature reward from Eq. (8); variants C01–C04 add the economic penalty from Eq. (9), introducing a cost-awareness component and additional observations #4 and #5. Table II lists the ablation variants T01–C04.

Our evaluation uses Austria's day-ahead electricity prices from the EPEX Spot platform [19], resampled to 5-minute intervals. The values are then normalized between 0 and 1, as shown in Figure 4. To reflect price uncertainty [20], each price forecast $\hat{\lambda}_t$ includes Gaussian noise:

$$\hat{\lambda}_t = \lambda_{\text{orig},t} + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2), \sigma = 0.01. \tag{11}$$

Training is conducted over 1×10^6 timesteps using 5-minute intervals ($\approx$ 3472 day-long episodes of 288 timesteps). Four parallel environments are initialized with different random seeds during training to encourage diverse exploration.

TABLE II: Ablation study results over evaluation dataset for different observation space and reward configurations.

Variant	#1	#2	#3	#4	#5	Reward Mode
T01	✓	✗	✗	✗	✗	Temperature Eq. (8)
T02	✓	✓	✗	✗	✗	Temperature Eq. (8)
T03	✓	✗	✓	✗	✗	Temperature Eq. (8)
T04	✓	✓	✓	✗	✗	Temperature Eq. (8)
C01	✓	✗	✗	✓	✓	Combined Eq. (10)
C02	✓	✓	✗	✓	✓	Combined Eq. (10)
C03	✓	✗	✓	✓	✓	Combined Eq. (10)
C04	✓	✓	✓	✓	✓	Combined Eq. (10)

#1–#5 refer to the observation components listed in Table I.

Performance metrics are cumulative rewards Eqs. (8–10) averaged over the evaluation set with 99% confidence interval (CI) bounds. We evaluate the performance of state-of-the-art RL algorithms through an ablation study, including Advantage Actor-Critic (A2C) [21], DDPG [22], PPO [23], and Soft Actor-Critic (SAC) [24]. All agents share a learning rate of 3×10^{-4} and a batch size of 64. Off-policy methods use a replay buffer 1×10^5 transitions and start learning after 2880 timesteps (ten episodes); on-policy methods update once per one episode. All experiments used a 4g/20GB NVIDIA A100 MIG instance with CUDA 12.4 and cuDNN 9.0.5.

B. Ablation Study Results

Table III presents the cumulative reward results for all controller types across temperature (T01–T04) and combined (C01–C04) settings in the `LLEC-HeatPumpHouse-1R1C` environment. In T01, where only temperature deviation is observed, the PI and PID controllers outperform the fuzzy controller by 11.6%. MPC further improves performance by 5.9% over PI/PID, while the best (perfect-foresight) MPC achieves the highest score with an additional increase of 5.1%. PPO and SAC trail MPC by 1.2% and 1.7%, respectively, but outperform the fuzzy controller by over 16%. A2C and DDPG yield lower performance than the PI/PID. In T03, where temperature deviation and the previous action are observed, PPO and SAC outperform MPC by up to 1% and achieve the best RL performance. MPC maintains consistent performance in the combined variant, achieving cumulative rewards of 246.68 (MPC) and 260.33 (Best MPC). Performance in C01–C04 decreases slightly due to added economic penalties. In C02, PPO performs best among RL algorithms, followed by DDPG, while A2C drops by 41.5% compared to its T02 score. In C03, SAC yields the highest reward, though all RL algorithms remain within 2.4%, indicating similar adaptation without time-of-day observation. In C04, PPO remains robust, whereas DDPG drops by 22.5% relative to C02, highlighting sensitivity to observation features such as previous actions.

Although MPC performs best with accurate forecasts and precise models, real-world applications rarely provide such ideal conditions. In contrast, RL agents learn adaptive strategies to handle uncertainty and conflicting objectives better. Figure 5 shows the indoor temperature trajectories under dynamic setpoints for the best-performing controllers across T01 to T04.

TABLE III: Ablation study results showing cumulative reward (mean ± 99 % CI) for each controller (T01–C04) in the `LLEC-HeatPumpHouse-1R1C` environment.

Variants:	**Temperature (T01–T04)**			**Combined (C01–C04)**		
Controller	Mean	Min*	Max*	Mean	Min*	Max*
T01/C01						
FUZZY	225.69	217.27	234.11	205.51	194.64	216.38
MPC	266.90	266.01	267.78	246.68	243.58	249.79
PI	251.98	250.46	253.51	230.86	226.17	235.55
PID	251.98	250.45	253.51	230.86	226.17	235.55
BEST MPC	280.51	280.40	280.62	260.33	257.19	263.48
A2C	247.67	243.10	252.25	234.44	227.28	241.59
DDPG	250.36	245.27	255.45	230.64	224.13	237.16
PPO	263.62	260.76	266.47	236.44	224.17	248.71
SAC	262.32	258.75	265.89	232.16	221.51	242.81
T02/C02						
A2C	251.75	245.72	257.78	142.01	119.22	164.79
DDPG	238.09	230.91	245.26	240.50	238.20	242.80
PPO	260.81	256.04	265.58	242.86	236.92	248.79
SAC	259.44	255.24	263.64	230.87	216.37	245.38
T03/C03						
A2C	206.42	196.46	216.38	227.90	220.11	235.69
DDPG	228.83	215.39	242.26	234.57	230.84	238.30
PPO	269.74	267.83	271.65	234.02	225.53	242.52
SAC	269.85	268.78	270.91	238.85	228.66	249.05
T04/C04						
A2C	245.55	234.67	256.44	229.34	219.30	239.37
DDPG	250.81	247.15	254.46	186.41	148.08	224.75
PPO	242.42	240.51	244.32	241.19	235.55	246.83
SAC	268.71	267.24	270.17	221.93	213.54	230.32

*99% CI bounds.

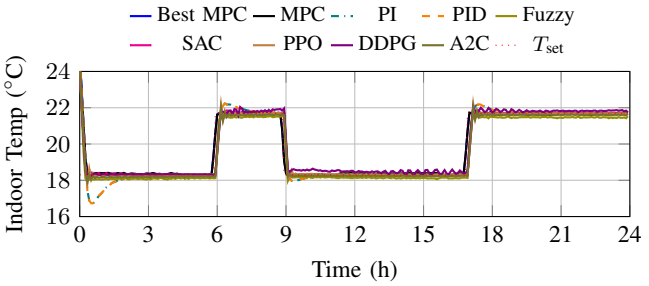


Fig. 5: Indoor temperature trajectories on the evaluation set for the best controllers across T01–T04 with dynamic setpoints.

VI. Conclusion

In this paper, we introduced *LLECBuildingGym*, an open-source simulation environment (available at https://github.com/KIT-IAI/LLECBuildingGym) for evaluating PI, PID, fuzzy, MPC, and RL-based control strategies for heat pump operation in buildings. The modular framework enables reproducible controller comparisons and systematic variations.

An ablation study addressing the first research question showed that combining temporal and economic observation variables with suitable reward formulations significantly improved RL agents' learning efficiency and control performance. Regarding the second research question, our evaluation showed that the best RL agent outperforms the most effective rule-based controller by 6%, exceeds the performance of a well-tuned MPC by 1%, and falls short of a perfect-foresight MPC by only 4%. Future research will focus on integrating subsystems such as stratified thermal storage, multiple heat pump types, PV generation, batteries, district heating, and electric vehicle charging.

VII. Acknowledgements

This work was supported in part by the Energy System Design (ESD) Project; in part by the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI and the HAICORE@KIT partition.

References

[1] Eurofound, "Decarbonisation of residential heating and cooling: The heat pump challenge," Eurofound research paper, Luxembourg, 2024.
[2] National Audit Office (NAO), "Decarbonising home heating," HC 581, Session 2023–24, March 2024.
[3] R. Dickel, "Decarbonizing Germany's heating sector," Oxford Institute for Energy Studies, OIES Paper ET29, Feb 2024.
[4] U.S. Department of Energy. (2024) Data and Analysis for Buildings Sector Innovation. [Online]. Available: https://www.energy.gov/eere
[5] D. Rolando *et al.*, "Heat pump system control: the potential improvement based on perfect prediction of weather forecast and user occupancy," in *12th IEA Heat Pump Conf.*, Rotterdam, Netherlands, 2017, pp. 1–9.
[6] A. Raffin *et al.*, "Stable-baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021.
[7] M. L. Bynum *et al.*, *Pyomo – Optimization Modeling in Python*, 3rd ed., ser. Springer Optim. Appl. Springer Cham, 2021, vol. 67.
[8] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Math. Program.*, vol. 106, no. 1, pp. 25–57, Mar 2006.
[9] K. Dalamagkidis *et al.*, "Reinforcement learning for energy conservation and comfort in buildings," *Build. Environ.*, vol. 42, pp. 2686–2698, 2007.
[10] A. Dounis and C. Caraiscos, "Advanced control systems engineering for energy and comfort management in a building environment—A review," *Renew. Sustain. Energy Rev.*, vol. 13, no. 6, pp. 1246–1261, 2009.
[11] M. Frahm *et al.*, "Occupant-oriented demand response with multi-zone thermal building control," *Applied Energy*, vol. 347, p. 121454, 2023.
[12] F. Calvino *et al.*, "The control of indoor thermal comfort conditions: introducing a fuzzy adaptive controller," *Energy Build.*, vol. 36, no. 2, pp. 97–102, 2004.
[13] E. Barrett *et al.*, "Autonomous HVAC Control, A Reinforcement Learning Approach," in *Mach. Learn. Knowl. Discov. Databases*. Cham: Springer, 2015, pp. 3–19.
[14] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," in *2017 54th ACM/EDAC/IEEE DAC*, 2017, pp. 1–6.
[15] G. Gao, J. Li, and Y. Wen, "DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings Via Reinforcement Learning," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8472–8484, 2020.
[16] T. Rohrer *et al.*, "Deep Reinforcement Learning for Heat Pump Control," in *Intelligent Computing*. Cham: Springer Nature Switzerland, 2023, pp. 459–471.
[17] Z. Jiang *et al.*, "Building HVAC control with reinforcement learning for reduction of energy cost and demand charge," *Energy and Buildings*, vol. 239, p. 110833, 2021.
[18] E. Tajalli-Ardekani *et al.*, "Experimental Data of Two Buildings Supplied by Heat Pump and District Heating under Multiple Heating Scenarios during Winter 2023-2024 ," Feb. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.14810476
[19] aWATTar GmbH, "Hourly Tariff – EPEX Spot Market-Based Electricity Prices," https://www.awattar.at/tariffs/hourly, 2025.
[20] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *Int. J. Forecast.*, vol. 30, no. 4, pp. 1030–1081, 2014.
[21] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," 2016.
[22] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2019.
[23] J. Schulman *et al.*, "Proximal Policy Optimization Algorithms," 2017.
[24] T. Haarnoja *et al.*, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," 2018.