Improving energy efficiency in smart building using deep reinforcement learning control strategy⋆

Oleksandr Vyshnevskyy[1,*,†], Liubov Zhuravchak[1,†] and Vitaliy Yakovyna[2,†]

[1] *Lviv Polytechnic National University, Bandery street 12, Lviv, Ukraine*
[2] *University of Warmia and Mazury in Olsztyn, Oczapowskiego street 2, Olsztyn, Poland*

Abstract

This study explores using control strategies to reduce overall building energy load. According to the literature review, most researchers used complex model-based methods or less effective Q-learning. The paper introduces a new technique for Heating, Ventilation, and Air Conditioning system control in mid-sized office buildings, leveraging an intelligent reinforcement learning controller built upon the model-free Proximal Policy Optimization algorithm. Our methodology integrates EnergyPlus building simulations, enabling an accurate and dynamic representation of system behavior across various control scenarios, with a specific focus on supply air temperature regulation. To refine our control strategy, we developed a Gymnasium co-simulation environment, which served as a robust platform for implementing and optimizing reinforcement learning algorithms. The performance of our developed deep reinforcement learning controller was evaluated against a traditional controller that employed an outdoor air temperature reset approach. The results revealed a substantial 27.8 % improvement in energy savings achieved while maintaining comfortable indoor temperature, humidity, and CO2 concentration levels. The reinforcement learning methodology demonstrates the potential to make sophisticated control strategies more accessible and easier to deploy in real-world building management systems. This study emphasizes the synergy between AI and building control systems, facilitating the adoption of intelligent energy management solutions in practical scenarios. In contrast to traditional model-based optimization techniques, deep reinforcement learning operates without the need for precise mathematical models of the physical system. Instead of relying on complex equations, control decisions are derived directly from the observed relationships between the actions taken and their subsequent effects on the system's state.

Keywords

Reinforcement learning, energy efficiency, load, building, control, agent, optimization

1. Introduction

Buildings are a major contributor to global energy consumption and CO_2 emissions, accounting for about 30 % compared to other consumers. This significantly affects climate change and forces to build new green (energy-efficient) buildings and work on renovation plans for existing buildings to meet energy efficiency targets. In old buildings, it is essential to use control strategies or policies that can reduce overall energy consumption [1].

Building management systems often depend on predefined rule-based thresholds and simple heuristics, limiting their ability to optimize energy consumption effectively. Smart regulation techniques, such as model predictive control (MPC) and reinforcement learning (RL), present compelling options and have demonstrated significant capability to outperform conventional approaches. MPC requires explicit physics process modeling in the building system, which can be a very complex problem. Building systems can also be represented using Semantic Web Technologies that can standardize different existing equipment vendors and sequence control strategies [2].

ICyberPhyS'25: 2nd International Workshop on Intelligent & CyberPhysical Systems, July 04, 2025, Khmelnytskyi, Ukraine
[1*] Corresponding author.
[†] These authors contributed equally.
oleksandr.k.vyshnevskyi@lpnu.ua (O. Vyshnevskyy); liubov.m.zhuravchak@lpnu.ua (L. Zhuravchak); yakovyna@matman.uwm.edu.pl (V. Yakovyna);
0009-0005-4857-9669 (O. Vyshnevskyy); 0000-0002-1444-5882 (L. Zhuravchak); 0000-0003-0133-8591 (V. Yakovyna);

In RL, the agent (the demand response controller) discovers the best strategy of action by learning from its experiences through trial and error instead of depending on pre-programmed instructions. This model-free characteristic makes RL highly adaptable for tackling control and optimization challenges, even when dealing with systems that are not fully understood or where information is limited.

Deep RL (DRL) represents a more sophisticated evolution of this approach, consolidating deep learning (DL) with RL principles. By learning from data, DRL can uncover complex patterns and automatically optimize control policies. Numerous studies have demonstrated that both MPC and RL can lead to significant energy savings in buildings without compromising occupant comfort. These findings highlight the potential of these advanced control techniques for creating more sustainable and efficient building systems.

This research goal is to investigate how a controller (agent) with advanced Heat, ventilation, and air conditioning (HVAC) control policy based on a novel model-free RL algorithm can improve energy efficiency compared to traditional methods.

Tasks include analyzing the state of art for control strategies, finding methods commonly used, determining issues in current approaches, choosing the appropriate algorithm, integrating it into the co-simulation software framework, training and tuning of optimal control policy, and comparing the energy savings and user comfort with baseline method.

This research introduces a novel approach by developing an RL agent using an advanced Proximal Policy Optimization (PPO) algorithm for supply air temperature control, aiming to enhance multiple optimization goals. We tackle the challenges associated with MPC strategies by employing a data-driven RL approach that doesn't require explicit system physics process modeling, simplifies and speedup development and utilization in real case scenarios, where there can be a diversity of different building structures and creating models for individual buildings is not reasonable as it is time, resource and cost consuming.

2. Literature review

Researchers extensively explored predictive control methods for HVAC systems, demonstrating their capability to decrease building energy consumption and enable more adaptable system operation. A recent study [3] showcased a simple, affordable, and scalable method to implement MPC in commercial buildings with older equipment, replacing traditional weather-based controls. The system cleverly uses the ventilation exhaust temperature to estimate the indoor temperature. Testing showed promising results, including a 33 % reduction in energy costs during a one-month spring period.

As a model-free, real-time learning strategy, RL enables agents to engage directly with their surroundings, capturing uncertainties as they happen. In [4], the authors presented a comprehensive review of RL applications in the energy sector, categorizing existing research and analyzing their strengths and limitations. The review highlights that while many studies demonstrate the potential of RL for improving energy system performance by 20 %, there is still room for improvement. The authors observed a lack of utilization of advanced DRL techniques and a reliance on simpler methods like Q-learning (QL). While data-driven approaches like RL offer a promising alternative, the adoption of RL in complex energy systems is not straightforward.

In [5], the authors introduced a novel approach for managing home energy systems that combines MPC with RL. The system focuses on optimizing energy use in a house. The primary goal is to leverage the house's thermal properties and battery storage to shift energy consumption away from peak pricing periods and capitalize on selling excess solar energy back to the grid. To address challenges due to model inaccuracies, uncertainties in weather forecasting, and unpredictable user behavior, the researchers developed a solution that uses a parameterized MPC framework to approximate the optimal energy management strategy. This framework is continually refined using a compatible delayed deterministic actor-critic algorithm, a type of RL method. Simulations

demonstrated that their approach effectively balances user comfort with economic considerations, even when faced with imperfect models and unpredictable real-world factors.

Research [6] explores the potential of RL as a more intelligent and energy-efficient alternative to the commonly used simple strategies based on predefined if-then rules. The authors presented a multi-agent RL framework that optimizes HVAC operation to minimize energy use while simultaneously incorporating occupant feedback on thermal comfort. The study investigates techniques like parameter sharing among multiple agents and various pre-training strategies. The results demonstrate that this framework can save controller training time and achieve notable energy savings, around 6 % for an entire building and up to 8 % for a single room, compared to traditional rule-based control. Importantly, these energy reductions are achieved without compromising occupant comfort, as feedback remains comparable to or even better than the baseline system.

Study [7] introduced a new method for optimizing the energy performance and thermal comfort of multizone buildings by using a combination of Artificial Intelligence (AI) and rule-based control for the central Air Handling Unit (AHU). The researchers first trained a DRL agent to make energy-efficient decisions regarding the AHU's supply water temperature, chiller valve, and economizer damper. To make this complex AI logic applicable in real-world building automation systems, they developed a method for extracting clear, actionable rules from the DRL agent's decision-making process. Results showed that the rule-based controller derived from the AI agent achieved almost the same energy efficiency while being much simpler to implement. This approach demonstrates the potential of combining AI and rule-based systems to make advanced energy management more accessible for real-world applications.

Occupants' unpredictable behavior, especially hot water usage patterns, presents a significant challenge in optimizing building energy systems. Traditional control systems often rely on conservative, energy-intensive strategies to ensure user comfort. Study [8] employed a model-free RL approach, allowing the system to be transferable to different buildings without requiring detailed building-specific models. The RL agent is trained offline using a stochastic hot water use model to simulate realistic occupant behavior and accelerate the learning process. The framework's effectiveness was validated using real-world data collected from a residential house, demonstrating significant energy savings (about 20 %) compared to conventional rule-based control systems. These energy savings were achieved without compromising occupant comfort or hot water availability. The authors highlight the potential of RL-based control systems to optimize building energy management by effectively learning and adapting to the dynamic and often unpredictable behavior of occupants.

Research [9] addresses the potential of demand response in residential buildings for achieving significant energy savings. Recognizing the need for fully automated energy management systems to unlock this potential, the authors proposed an approach using RL. They formulated the energy management system scheduling task as an RL problem, arguing that it can be effectively addressed by dividing appliances into clusters and optimizing each cluster's schedule independently. Compared to existing methods, this approach offers several advantages: it eliminates the need for explicitly modeling user satisfaction, enables the energy management system to proactively initiate tasks, allows users to make more flexible requests, and reduces computational complexity. The study demonstrated the application of QL to illustrate the effectiveness of their proposed framework.

In [10], the authors applied a DRL approach to optimize energy management within smart buildings. They authors established a comprehensive framework for the building energy management system, modeling various components such as energy storage, photovoltaic generation, electric vehicle charging, and household appliances. The authors developed a QL model incorporating relevant operational restrictions. Simulations validated the effectiveness of the proposed approach. The results have shown that the DRL-based energy management strategy can effectively meet the energy demands of the intelligent building while ensuring efficient energy allocation. The proposed method outperforms other control algorithms, maintaining errors within an acceptable range of 10 %.

Traditional building management systems struggle to handle the increasing complexity and diversity of data from various sources, such as sensors and occupant schedules. In [11], the authors addressed the challenge of managing indoor air quality in buildings using a knowledge graph-enhanced DRL approach. They proposed using semantic web knowledge graphs to represent diverse building information related to indoor air quality in a structured and interconnected manner. This knowledge graphs-based approach captures the relationships between various factors influencing indoor air quality, providing a more comprehensive and context-aware representation of the building's state. The authors outlined that this approach simplifies building management, reducing the complexity faced by facility managers in maintaining optimal indoor air quality.

In [12], the authors applied an RL-powered energy management system designed for smart buildings operating within a smart grid, where buildings can intelligently exchange energy with the grid, incorporating local renewable sources like solar panels, energy storage systems, and electric vehicle charging stations with vehicle-to-grid capabilities. The proposed system described the energy management task as Markov decision process (MDP), defining the possible states, actions, transition probabilities, and rewards. A QL algorithm was employed to enable the system to learn optimal energy dispatch decisions over time, adapting to uncertain factors such as building load demands, charging requirements, and solar energy generation. Simulations using real-world data demonstrated the algorithm's effectiveness in reducing energy costs compared to existing methods and random decision-making. The system effectively minimizes expenses. The authors outlined that this RL-based approach can be applied to various smart grid environments, including microgrids and industrial settings.

Research [13] introduced a multi-agent control system for optimizing building comfort and energy efficiency. Their framework consists of three independent agents, each responsible for managing a specific aspect of indoor environmental quality: air quality and visual and thermal comfort. A stochastic model, leveraging probabilistic and evolutionary algorithms, estimates occupant presence based on CO_2 levels. System parameters are represented using fuzzy logic to account for uncertainty. The control agents utilize Fuzzy QL to handle the continuous nature of the building's state and control actions. Simulations demonstrate the system's effectiveness, showcasing accurate occupancy estimation and significant energy savings of up to 56 % while maintaining decent levels of occupant comfort.

The deep Q-learning (DQL) approach utilizes artificial neural networks instead of the traditional discrete Q-table in conventional QL, allowing it to investigate more extensive action and state spaces. In [14], the authors proposed a system based on a neural network-based QL algorithm, enabling it to learn and adapt to household energy consumption patterns intelligently, aiming to reduce peak energy demand and promote energy conservation in residential buildings. By employing an advanced Neural Fitted QL technique, the system makes agile and efficient decisions regarding energy usage, striking a balance between minimizing costs and maintaining comfort. Simulations using a classic Canadian home demonstrated the system's ability to significantly decrease energy consumption during peak hours, contributing to a smaller carbon footprint for residential dwellings. The authors outlined that widespread adoption of such an approach in residential areas could substantially reduce peak demand, leading to optimized resource allocation for utility companies and potentially lower energy tariffs for consumers.

In [15], the use of DRL was investigated to optimize energy consumption scheduling in residential buildings. The authors explored DQL and Deep Policy Gradient (DPG) algorithms to determine their effectiveness in managing household energy usage patterns. The study examined two specific goals: decreasing energy costs and smoothing the overall energy consumption profile. The authors explored the impact of dynamic electricity pricing signals to actuate shifting energy use to off-peak hours. Through simulations using the Pecan Street dataset, the study demonstrated that a single DRL agent can effectively tackle multiple energy management objectives. The results showed that DPG outperforms DQL for real-time energy scheduling. However, using RL for energy management in actual buildings presents some challenges. The exploration phase for agents in a real building setting can result in higher operational costs and longer training durations. To overcome this, authors have

applied the Transfer Learning (TL) approach in [16], which has shown promise as a way to make DRL more practical for managing building energy use. This is important because DRL often requires a lot of data and can be difficult to implement in real-world settings. They achieved enhanced temperature regulation and energy savings and, therefore, demonstrated the online TL as a viable way to make DRL controllers more scalable and practical for real-world building applications.

Thus, this research will contribute to improving the development of robust energy-efficient control solutions. According to the literature review most researchers used less effective Q-Learning or complex model-based methods. Our research employs an advanced Reinforcement Learning control approach with a cutting-edge PPO algorithm to enhance building energy efficiency, maintaining multiple occupant comfort objectives.

3. Methodology

In our research, we applied an advanced model-free approach with Reinforcement Learning [17]. RL is a type of machine learning in which an agent (controller) learns the best way to solve a problem through a process of experimentation [18]. The controller learns by experiencing consequences, either positive or negative, based on its choices within the environment. The agent's learning process involves finding the best action to take in each possible state to maximize its overall reward [19]. The agent's goal is to develop a strategy that leads to the highest total reward over time. This involves a trade-off: the agent must balance exploring new and potentially better strategies with exploiting the knowledge it has already gained [20].

3.1. Proximal policy optimization

We applied the Proximal Policy Optimization (PPO) algorithm, a flexible and widely used RL technique for solving control problems. Recently, the PPO algorithm has gained prominence and has been the standard RL algorithm at OpenAI since 2018. PPO algorithm has been utilized in various domains, including energy system scheduling. It shows superior fit to hyperparameters and enhanced training efficiency when compared to DDPG, along with delivering better objective function values. Researchers indicated that the PPO algorithm effectively explores continuous action spaces with a more stable update mechanism, resulting in lower operating costs than DDPG [21].

The PPO algorithm enhances the stability of RL agent training by preventing excessively large updates to the policy. It achieves this by employing a ratio that reflects the disparity between the current and previous policies and then clipping this ratio within a predetermined range. This clipping mechanism ensures that policy updates remain moderate, promoting greater stability during training. Essentially, PPO aims to improve training stability by restricting the extent to which the policy is altered in each training episode, thereby avoiding abrupt and potentially detrimental shifts.

3.2. Markov decision process

A number of researchers described optimal building control Reinforcement Learning tasks as Markov decision process (MDP), defining possible states, actions, transition probabilities, and rewards [22,23]. Accordingly, our RL problem can be structured as MDP defined by:

- $\{S\}$ – different possible situations the agent might encounter (state space)
- $\{A\}$ – actions the agent can take (action space)
- $P_a(s_t, s_{t+1})$ – the probability of transitioning between states s_t *and* s_{t+1} based on action a
- $R_a(s_t, s_{t+1})$ – reward received after transition from state s_t *to* s_{t+1} with action a

At each time step t, the agent, following a policy π, reacts with the environment by taking an action a_t based on the observed state s_t according to (1)

$$a_t = \pi(s_t). \tag{1}$$

The environment responds to this action with a reward R_{a_t} value and then acquires a new state s_{t+1} and provides the agent with a scalar reward $R_{a_{t+1}}$. The agent's goal is to discover the best strategy, denoted as π, which leads to the highest total reward over a series of actions, as illustrated in (2)

$$max\ \mathrm{E}\left[\sum_{t=0}^{\infty}\gamma_t R_{a_t}(s_t, s_{t+1})\right], \tag{2}$$

where γ – discount factor, determines the relative importance of rewards, $0 \leq \gamma_t \leq 1$.

3.2.1. State space

For each step, the agent receives environment observations, which represent the state of the system at this time point according to (3)

$$s_t = \{T_{out}, T_{in}, H_{in}, C_{in}, T_{htg\_spt}, T_{clg\_spt}, E_{heating\_load}, E_{cooling\_load}\}, \tag{3}$$

where T_{out}, T_{in} – temperature of outdoor air and room air, H_{in} – room air relative humidity, C_{in} – room air CO_2 concentration, T_{htg_spt}, T_{clg_spt} – setpoint temperature for heating and cooling, $E_{heating\_load}$ – natural gas heating load, $E_{cooling\_load}$ – electricity cooling load.

The DRL agent's decision-making process considered eight different variables, measured in the case study building during simulation with a time step of 15 minutes. These variables, along with their allowable ranges and units, are outlined in Table 1.

Table 1
State space variables

Variable	Min	Max	Unit	Description
Tout	-40	40	°C	Outdoor air temperature
Tin	0	40	°C	Room air temperature
Hin	0	100	%	Room air relative humidity
Cin	0	100000	ppm	Room air CO2 concentration
Thtg_spt	0	30	°C	Setpoint temperature for heating
Tclg_spt	0	30	°C	Setpoint temperature for cooling
Eheating_load	0	28	kWh	Natural gas heating load
Ecooling_load	0	28	kWh	Electricity cooling load

3.2.2. Action space

The agent makes actions at each time step *t* (every 15 minutes), representing the decisions for control of the building's HVAC system with heating and cooling devices. Within this study, the agent determines the optimal supply air temperature as a control variable that serves as the agent's mechanism for optimizing the energy efficiency, so actions at time step *t* can be represented as (4). The action space is discrete, with 100 possible actions. In our study, these discrete actions are rescaled to a continuous range of 15.0 to 30.0, representing supply air temperature setpoint ranges.

$$a_t = \{T_{sat_t}\}, \tag{4}$$

where T_{sat_t} – supply air temperature at time step t.

3.2.3. State transition

When the agent makes a control decision, it affects the state of the environment, potentially leading to a new situation. The transitioning to a particular new state depends not only on the action taken

but also on various unpredictable factors within the environment. Modeling these transitions can be very complex due to the inherent uncertainty in how the environment responds [24]. However, DRL offers a way to sidestep this challenge by learning directly from experience. Instead of trying to explicitly model the probability of each transition, DRL uses neural networks to capture the effects of uncertainty based on the observed data.

3.2.4. Reward function

The main function of the DRL controller is to optimize the supply air temperature (T_{sat}) by regulating the actuator variable every 15 minutes. The supply air temperature setpoint is within a range from 15 °C to 30°C. The optimization process aims to minimize zone temperature discomfort, as well as maintain humidity within acceptable limits while minimizing energy consumption and CO_2 level. The reward function R_t used for the DRL agent is provided in (5). By structuring the reward function in this way, the MDP seeks to identify energy management schedules that maximize the cumulative reward over time, leading to minimized operational costs.

$$R_t = -(E_{penalty} + T_{penalty} + H_{penalty} + CO2_{penalty}), \quad (5)$$

where:

$E_{penalty} = (E_{heating\_load} + E_{cooling\_load})$,

$$T_{penalty} = \begin{cases} \min\big((T_{in} - T_{clg\_spt}), (T_{htg\_spt} - T_{in})\big), \ if\ (T_{in} < T_{clg\_spt}\ or\ T_{in} > T_{htg\_spt}) \\ 0, \ else \end{cases},$$

$$H_{penalty} = \begin{cases} \min((H_{in} - 40), (60 - H_{in})), \ if\ (H_{in} < 40\ or\ H_{in} > 60) \\ 0, \ else \end{cases},$$

$CO2_{penalty} = C_{in}$.

The energy load reflects the overall energy used by building HVAC system, it combines the energy load of a natural gas meter (measures heating) and an electricity meter (measures cooling). Zone temperature discomfort is minimized by penalizing the deviations of zone air temperature (T_{in}) from heating setpoint temperature (T_{htg_spt}) and cooling setpoint temperature (T_{clg_spt}). When room temperature falls below the heating setpoint temperature, the penalty increases proportionally to how much it's too cold. However, when the temperature rises above the cooling setpoint temperature, the penalty increases proportionally, reflecting the discomfort of being too warm, particularly during summer periods when cooling is needed. For optimal humidity level, the penalty is applied whenever the humidity (H_{in}) goes outside its defined comfortable range of 40 % to 60 %. This penalty increases accordingly as the humidity moves further away from this allowed range.

4. Development of RL agent

This section describes the development of the RL controller that makes decisions based on its pre-trained optimal policy and transmits control signals to the simulated HVAC system. We developed a DRL controller in Python 3.11 language, using Farama Foundation Gymnasium framework 0.28.1 (fork of Open AI Gym). Gymnasium is specifically designed to develop and test RL algorithms using Gymnasium environments, a configurable space where agents iteratively interact with and learn from the environment via actions and rewards mechanisms under specified conditions and refine its control strategy.

We developed a custom Gymnasium environment for EnergyPlus 24.2, a comprehensive building energy simulation engine that models energy consumption in buildings for various purposes, including heating, cooling, and ventilation. It has specified paths to selected medium office building model "idf" file and weather "epw" described in detail in Section 5. We defined observation space according to (3), specifying appropriate EnergyPlus variables and meter names as Gym Box with possible value ranges according to minimum and maximum limits specified in Table 1. We implemented the action space defined in (4) as a discrete Gym space with 100 possible actions. Then,

these discrete actions are rescaled to a continuous range of 15.0 to 30.0, representing supply air temperature setpoint ranges. The schema of the constructed co-simulation is described in Figure 1.

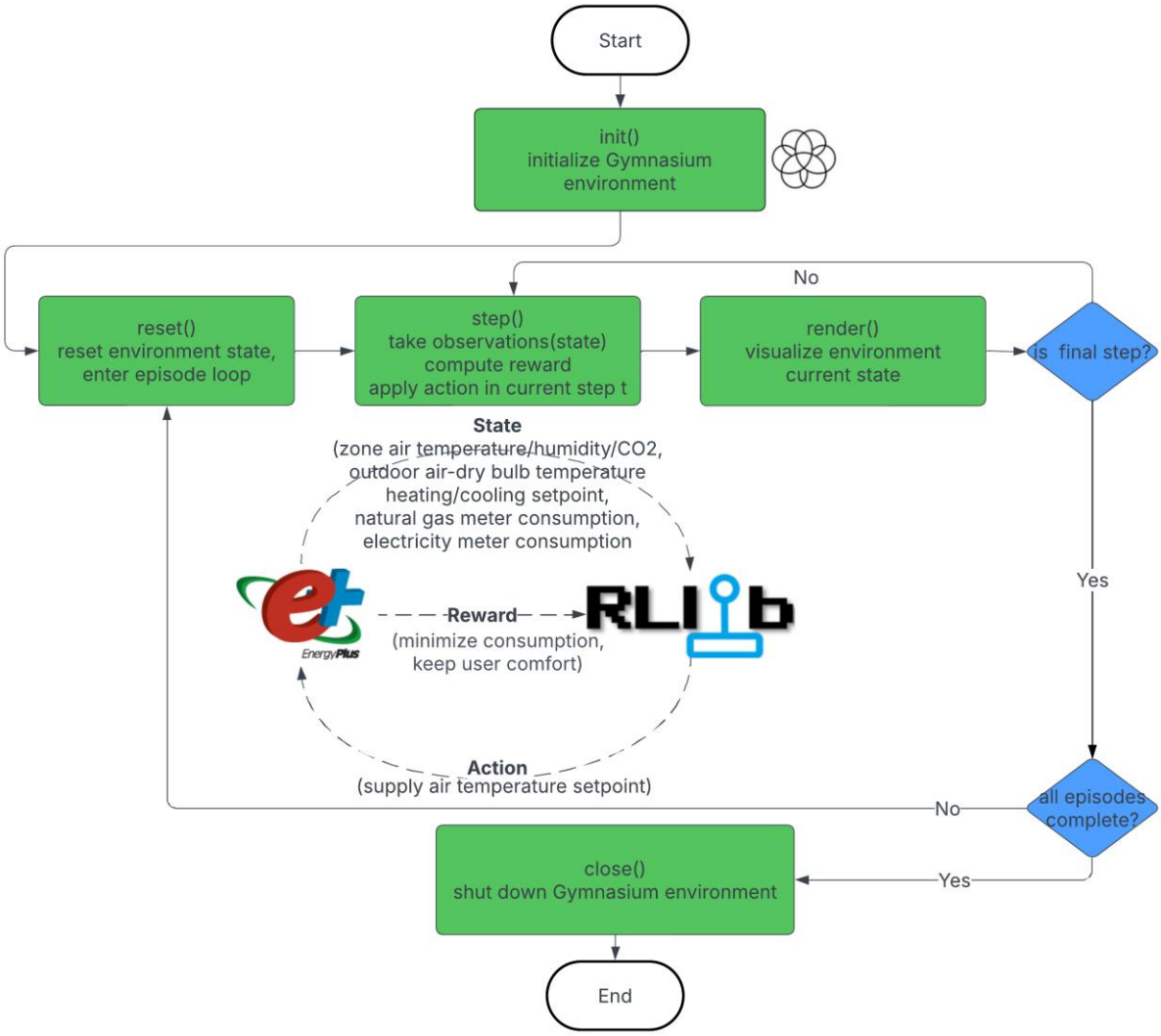

Figure 1: Constructed RL co-simulation with EnergyPlus and Gymnasium.

The reward function was constructed according to (5) to quantify the effectiveness of actions taken by the agent in achieving desired objectives, such as minimizing energy consumption while maintaining comfort levels, it is crucial in taking control of the learning process.

We used the EnergyPlus Python API programmatic access to communicate with the EnergyPlus simulation process, which provides integration capabilities with the Gymnasium framework, allowing the Gymnasium Environment to send commands to and receive data from EnergyPlus simulations. This integration defines the bidirectional communication channel, where the evolving state of the simulation is continuously fed back into the learning algorithm, and actions proposed by the algorithm are applied in EnergyPlus. Using Ray RL Library 2.20.2 (RLib), the algorithm is then configured to operate within this custom environment. RLib is an open-source high-level library that provides sophisticated, production-ready, scalable RL algorithms, including the PPO algorithm and many others.

For configuring the PPO algorithm, we employed the following set of parameters.: the discount factor gamma=0.95, learning rate lr=0.003, Adam optimizer was applied, KL divergence start coefficient kl_coeff=0.3, train_batch_size=2880, sgd_minibatch_size=360, value function loss coefficient vf_loss_coeff=0.01, clip parameter clip_param=0.2, use critic as baseline use_critic=true, use the Generalized Advantage Estimator (GAE) use_gae=true, Long Short-Term Memory (LSTM) layer was enabled to improve performance with temporal dependencies. The Torch framework was

used for training. A number of training and validation cycles were conducted to refine the model, involving iterative simulation runs where the agent's policies continually evolve based on feedback from the environment. The training loop continues until a specified number of timesteps=2880000 (1000 episodes) have been completed, after which the best-performing policy is saved.

Our developed co-simulation application has a console interface. The sample output with the current iteration, episode, and reward is provided in Figure 2. EnergyPlus simulation settings include options for generating CSV output files containing simulation results. The file "eplusout.csv" is generated at the end of the simulation so we can review the state of output variables, setpoints, and meters during all steps of the simulation. This data was used to create charts and visualize the simulation process. Ray library logs results to the Tensor Board allowing the analysis of the training metrics during all iterations.

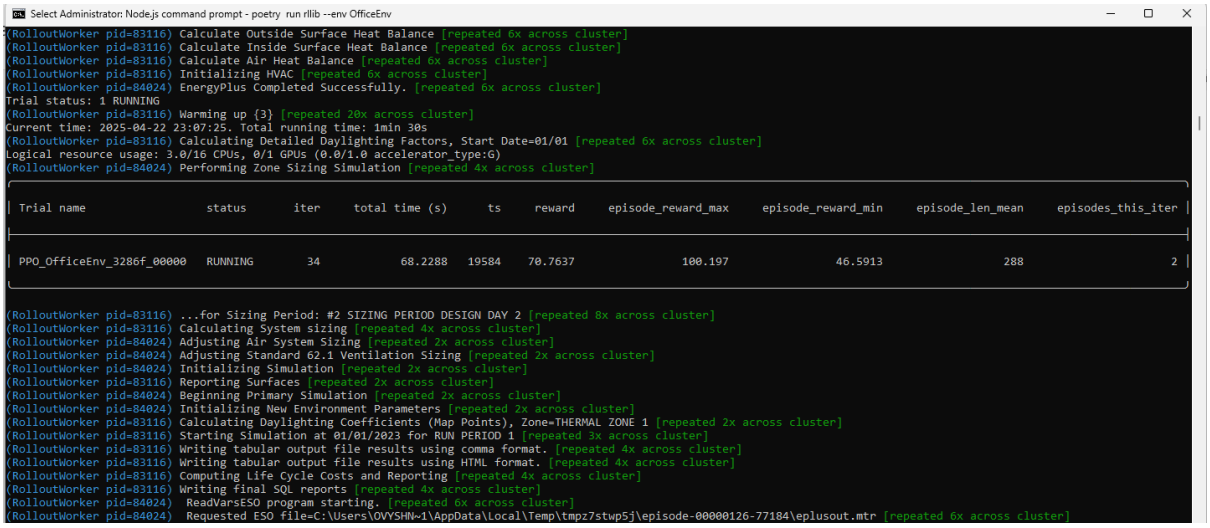

Figure 2: Developed co-simulation application console output.

5. Results

Our experiments were implemented on an Asus Vivobook Pro laptop supplied with AMD Ryzen 7 6800H 8 core processor (3.20 GHz), 32 GB RAM, and AMD Radeon 680M 12 core graphics processor with Windows 11 Pro operating system.

The study utilized a simulated building environment using EnergyPlus 24.2.0. We used a medium office building model from OpenStudio Application 1.8.0 example models illustrated in Figure 3. We used the weather file "UKR_LV_Lviv.Intl.AP.333930_TMYx.2009-2023.epw" for simulation downloaded from the Repository of Building Simulation Climate Data. The building has one story with 3m height, four spaces with 10m width and 10m length each, it is constructed with 100mm brick exterior walls, 100mm lightweight concrete roof, two windows, and one metal door.

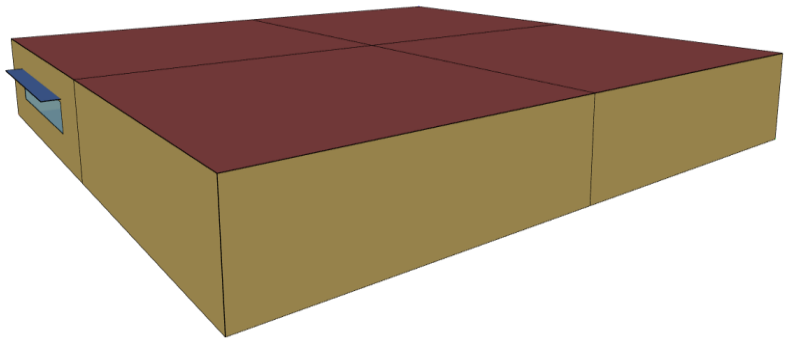

Figure 3: Medium office building model used for simulation.

The building model has an air loop HVAC system controlling a single thermal zone represented in Figure 4.

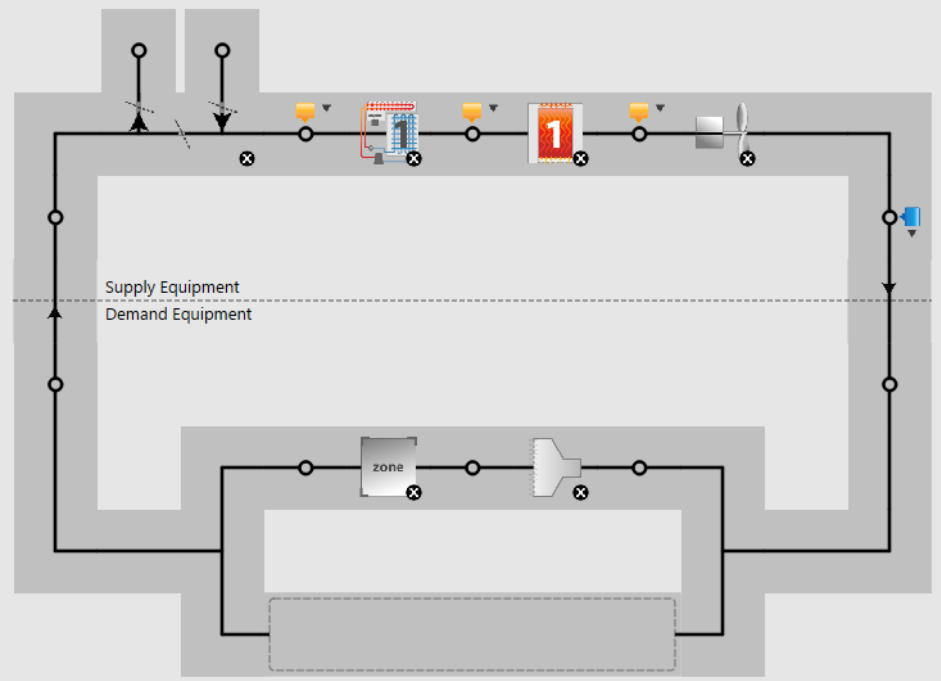

Figure 4: Air Loop HVAC system.

The system incorporates an electric-powered cooling mechanism. For heating, the air circulation system utilizes a gas-operated heating element. A simple fan model circulates air through the system. The outdoor air system is responsible for managing the outdoor air intake and ventilation for the main air loop. It controls the mixing of fresh outdoor air with return air, which has already circulated through the building. This mixed air stream then gets conditioned (cooled or heated) before being supplied back to the zones. The system uses constant airflow to the zone without the ability to adjust airflow or reheat. The model has people, lights, and electric equipment load definitions for each space. Schedule for lightning, number of people, electric equipment, and cooling/heating setpoints covers different hours of the day, working days, and holidays.

We performed a series of experiments with the training of our DRL agent hyperparameters tuning within the developed co-simulation environment, starting with a short simulation period duration (1 day), 96000 time steps, and 1000 episodes, this allowed the PPO algorithm more frequent policy updates, making it easier for the agent to learn the daily patterns, reduce the variance in experiences and therefore adjusted reward function can provide a more consistent learning signal. We followed a curriculum learning approach with gradually increased simulation duration as the agent performance improved. This speeds up the development and testing cycle but requires extending to longer simulation periods as a refined approach to capture all the complexities of HVAC control. EnergyPlus simulation takes a one-month period, from April 1 to April 30 2023. with timestep=4 (15 minutes), other model and simulation parameters are provided in Table 2.

Following this training phase, the effectiveness of the learned control strategy was evaluated by deploying it and controlling our building model over a month period. The DRL controller performance was assessed by analyzing the total accumulated rewards. Overall, the results demonstrated the controller's capability to effectively optimize multiple objectives defined in (5) concurrently.

After the training phase, the DRL controller performance was compared with a baseline supply air temperature control strategy provided by the EnergyPlus “OutdoorAirReset” Setpoint Manager. This manager provides a good balance between the warmest and coldest strategies based on the outdoor air temperature reset method. The setpoint manager was configured to dynamically adjust

the supply air temperature based on outdoor conditions, with a setpoint of 12.8 °C when the outdoor temperature is 15.6 °C or below, gradually increasing to a setpoint of 18.3 °C when the outdoor temperature reaches 26.7 °C or above, providing efficient cooling while maintaining comfort across a range of weather conditions.

Table 2
Building model simulation parameters

Parameter	Value	Unit
People per space floor area	0.05	people/m^2
Carbon dioxide generation rate	0.000038	$L/s \cdot W$
Watts per space floor area (lightning)	10	W/m^2
Watts per space floor area (electric equipment)	5	W/m^2
Watts per space floor area (printer)	200	W
Setpoint for heating (from 22 PM to 6 AM)	15.6	°C
Setpoint for heating (from 06 AM to 22 PM)	21	°C
Setpoint for cooling (from 22 PM to 6 AM)	26.7	°C
Setpoint for cooling (from 06 AM to 22 PM)	24	°C
Algorithm for heat balance	Conduction transfer function	
Algorithm for room air heat balance	Third order backward difference	
Inside surface convection algorithm	TARP	
Outside surface convection algorithm	DOE-2	
Number of time steps per hour	4	

The results of the supply air temperature control policy by RL agent and baseline for the first week of April are represented in Figure 5, for comparative analysis, we included data on temperature thresholds for heating and cooling systems, as well as a graph depicting external air temperature variations.

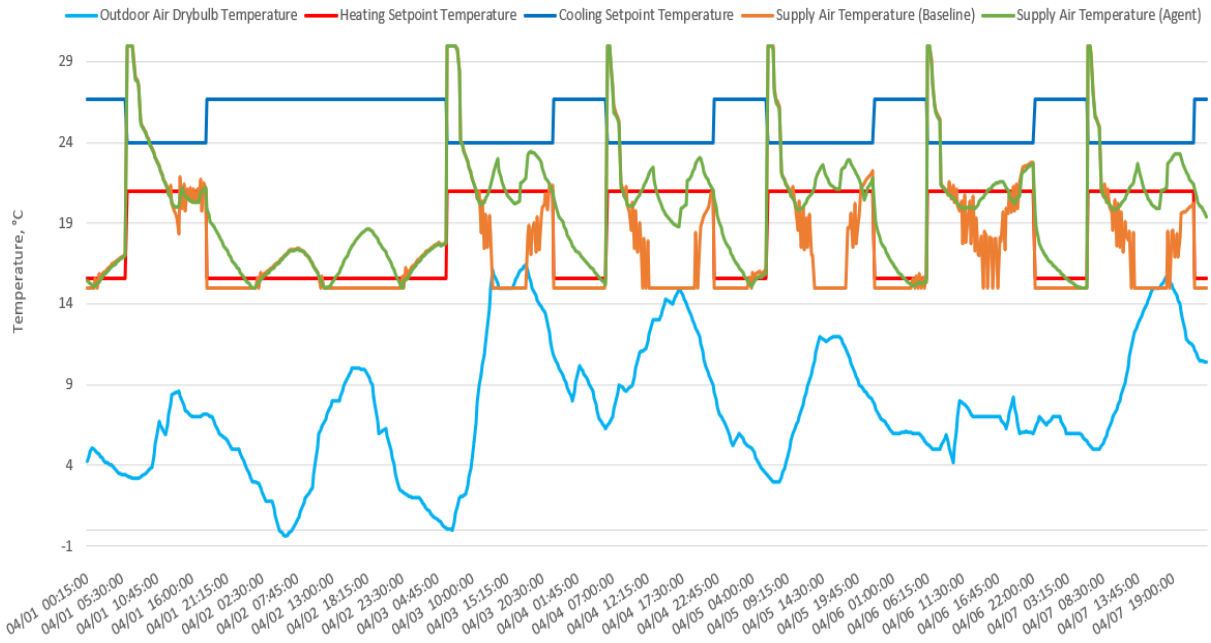

Figure 5: Supply air temperature control policy for 1 week period by RL agent and baseline.

It can be observed that on April 1, at the start of the day working hours both controllers tried to increase the supply air temperature to raise room temperature inside comfort ranges. Then, during the day, the supply air temperature goes down as outdoor temperature increases. On April 3, during working hours, the RL agent tries to increase the supply air temperature to satisfy room temperature

inside heating and cooling thermostat setpoints. On the other hand, the baseline controller tries to decrease the supply air temperature to the minimum value, providing the room temperature at the minimum comfort level. A comparison of room comfort conditions, including room temperature, humidity, and air CO_2 concentration, is represented in Figure 6.

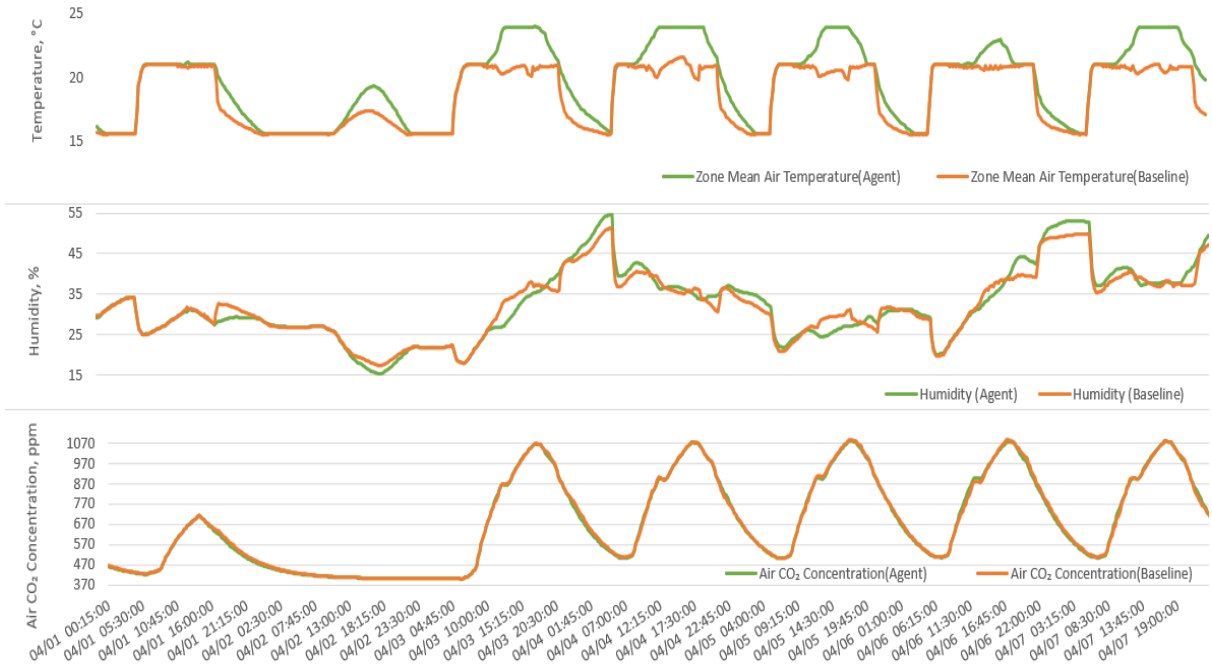

Figure 6: Zone comfort comparison.

The cumulative sum of the natural gas load established by our RL agent and the conventional control system utilizing a setpoint manager is illustrated in Figure 7. As we can see, the RL algorithm demonstrates superior efficiency, with 962.5 kWh of energy load per month. This represents a significant 27.8 % energy reduction compared to the standard controller, which utilizes 1248.8 kWh over the same period while maintaining an appropriate room comfort level. Our research shows that proposed advanced control strategies, specifically based on the DRL approach, can make buildings use energy much more efficiently.

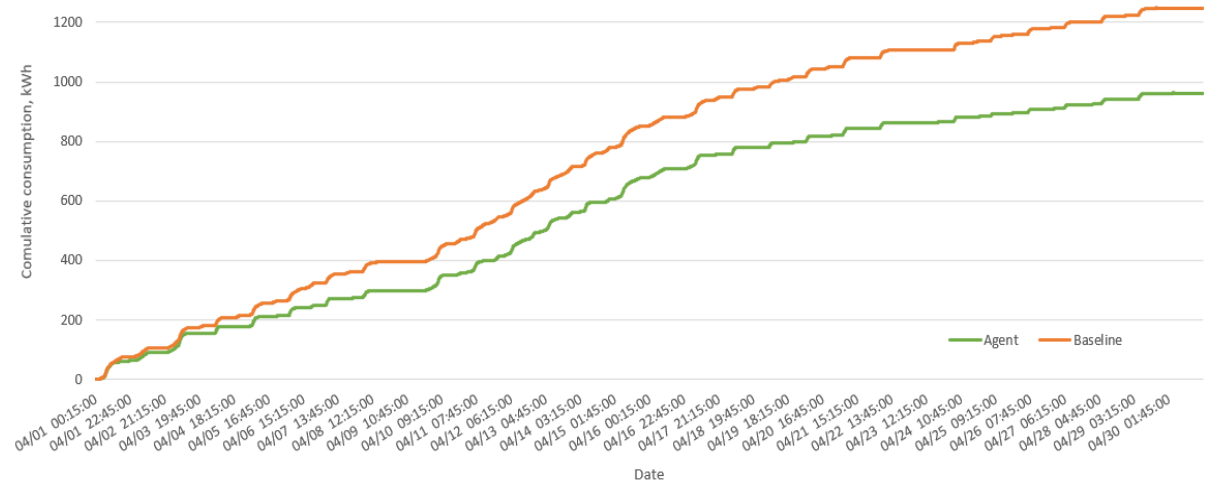

Figure 7: Cumulative natural gas load by RL agent and baseline.

Conclusion

This research proposes an innovative approach to Heating, Ventilation, and Air Conditioning system management in medium-sized office buildings, employing a smart RL controller based on a model-

free Proximal Policy Optimization algorithm. Our framework integrates EnergyPlus building modeling, which facilitates a precise and dynamic description of system behavior within different control tasks, specifically focusing on supply air temperature regulation. We developed a Gymnasium co-simulation environment to refine the custom control strategy using RL algorithms. The developed DRL controller's performance was benchmarked against a conventional controller employing an outdoor air temperature reset method. Results demonstrated a significant 27.8 % enhancement in energy savings achieved while maintaining comfortable indoor temperature, humidity, and CO_2 concentration levels. According to [25] authors used fuzzy logic for similar smart home optimization task which give efficiency gain in 25.7%, which indicates that our RL approach with PPO algorithm provides similar or better results.

RL methodology presents the potential to make sophisticated control strategies more approachable and easier to deploy in real HVAC systems. Compared to traditional model-based optimization, DRL is model-free and does not require precise mathematical representations of the physical system. Instead of relying on intricate equations, control decisions are derived directly from the observed connections between actions taken and their resulting effects on the system's condition. The study highlights the promising synergy between the RL approach and HVAC control, making the employment of smart energy management applications in real-world scenarios.

The limitation of the proposed model is it does not take into account the frequency of turning the controller on/off, since frequent switching of the heating element can lead to its failure. This can be overcome by extending reward function to adding penalty to frequent changes to the actuator variable that will provide smoother behaviour without abrupt controller utilization.

Future research will explore multi-agent reinforcement learning approaches to coordinate multiple HVAC zones simultaneously, potentially improving system-wide energy efficiency and occupant comfort. Additionally, investigating the scalability of the RL approach to large commercial buildings with complex HVAC architectures and exploring integration with smart grid systems for demand response optimization represent promising avenues for advancing this technology.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] O. Vyshnevskyy, L. Zhuravchak, Forecasting the Electricity Consumption for energy management software using an Ensemble model, in: Proceedings of the 2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, Oct. 2024, pp. 1-5. doi: 10.1109/CSIT65290.2024.10982553.

[2] O. Vyshnevskyy, L. Zhuravchak, Semantic Models for Buildings Energy Management, in: Proceedings of the 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, Oct. 2023, pp. 1–4.

[3] H. T. Walnum, I. Sartori, P. Ward, S. Gros, Demonstration of a low-cost solution for implementing MPC in commercial buildings with legacy equipment, Applied Energy, vol. 380, p. 125012, Feb. 2025, doi: 10.1016/j.apenergy.2024.125012.

[4] A. T. D. Perera, P. Kamalaruban, Applications of reinforcement learning in energy systems, Renewable and Sustainable Energy Reviews, vol. 137, p. 110618, Mar. 2021.

[5] W. Cai, S. Sawant, D. Reinhardt, S. Rastegarpour, S. Gros, A Learning-Based Model Predictive Control Strategy for Home Energy Management Systems, IEEE Access, vol. 11, pp. 145264–145280, 2023, doi: 10.1109/ACCESS.2023.3346324.

[6] D. Bayer, M. Pruckner, Enhancing the Performance of Multi-Agent Reinforcement Learning for Controlling HVAC Systems, in 2022 IEEE Conference on Technologies for Sustainability (SusTech), Corona, CA, USA: IEEE, Apr. 2022, pp. 187–194. doi: 10.1109/SusTech53338.2022.9794179.

[7] G. Razzano, S. Brandi, M. S. Piscitelli, A. Capozzoli, Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal

management of Heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings, Applied Energy, vol. 381, p. 125046, Mar. 2025, doi: 10.1016/j.apenergy.2024.125046.
[8] A. Heidari, F. Maréchal, D. Khovalyg, An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach, Applied Energy, vol. 312, p. 118833, Apr. 2022, doi: 10.1016/j.apenergy.2022.118833.
[9] Z. Wen, D. O'Neill, H. Maei, Optimal Demand Response Using Device-Based Reinforcement Learning, IEEE Trans. Smart Grid, vol. 6, no. 5, pp. 2312–2324, Sep. 2015, doi: 10.1109/TSG.2015.2396993.
[10] X. Huang, D. Zhang, X. Zhang, Energy management of intelligent building based on deep reinforced learning, Alexandria Engineering Journal, vol. 60, no. 1, pp. 1509–1517, Feb. 2021, doi: 10.1016/j.aej.2020.11.005.
[11] K. L. Mugumya, J. Y. Wong, A. Chan, C.-C. Yip, S. Ghazy, Indoor haze particulate control using knowledge graphs within self-optimizing HVAC control systems, IOP Conf. Ser.: Earth Environ. Sci., vol. 489, no. 1, p. 012006, Apr. 2020, doi: 10.1088/1755-1315/489/1/012006.
[12] S. Kim, H. Lim, Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings, Energies, vol. 11, no. 8, p. 2010, Aug. 2018, doi: 10.3390/en11082010.
[13] P. Korkidis, A. Dounis, P. Kofinas, Computational Intelligence Technologies for Occupancy Estimation and Comfort Control in Buildings, Energies, vol. 14, no. 16, p. 4971, Aug. 2021, doi: 10.3390/en14164971.
[14] C. Mahapatra, A. Moharana, V. Leung, Energy Management in Smart Cities Based on Internet of Things: Peak Demand Reduction and Energy Savings, Sensors, vol. 17, no. 12, p. 2812, Dec. 2017, doi: 10.3390/s17122812.
[15] E. Mocanu et al., On-Line Building Energy Optimization Using Deep Reinforcement Learning, IEEE Trans. Smart Grid, vol. 10, no. 4, pp. 3698–3708, Jul. 2019, doi: 10.1109/TSG.2018.2834219.
[16] D. Coraci, A scalable approach for real-world implementation of deep reinforcement learning controllers in buildings based on online transfer learning: The HiLo case study, Energy and Buildings, vol. 329, p. 115254, Feb. 2025, doi: 10.1016/j.enbuild.2024.115254.
[17] L. Spangher et al., Prospective Experiment for Reinforcement Learning on Demand Response in a Social Game Framework, in: Proceedings of the Eleventh ACM International Conference on Future Energy Systems, Virtual Event Australia: ACM, Jun. 2020, pp. 438–444. doi: 10.1145/3396851.3402365.
[18] A. Forootani, M. Rastegar, M. Jooshaki, An Advanced Satisfaction-Based Home Energy Management System Using Deep Reinforcement Learning, IEEE Access, vol. 10, pp. 47896–47905, 2022, doi: 10.1109/ACCESS.2022.3172327.
[19] B. Park, A. R. Rempel, A. K. L. Lai, J. Chiaramonte, S. Mishra, Reinforcement Learning for Control of Passive Heating and Cooling in Buildings, IFAC-PapersOnLine, vol. 54, no. 20, pp. 907–912, 2021, doi: 10.1016/j.ifacol.2021.11.287.
[20] O. Almughram, S. Abdullah Ben Slama, B. A. Zafar, A Reinforcement Learning Approach for Integrating an Intelligent Home Energy Management System with a Vehicle-to-Home Unit, Applied Sciences, vol. 13, no. 9, p. 5539, Apr. 2023, doi: 10.3390/app13095539.
[21] J. Wang, Y. Wang, D. Qiu, H. Su, G. Strbac, Z. Gao, Resilient energy management of a multi-energy building under low-temperature district heating: A deep reinforcement learning approach, Applied Energy, vol. 378, p. 124780, Jan. 2025, doi: 10.1016/j.apenergy.2024.124780.
[22] A. Tortorelli, G. Sabina, B. Marchetti, A Cooperative Multi-Agent Q-Learning Control Framework for Real-Time Energy Management in Energy Communities, Energies, vol. 17, no. 20, p. 5199, Oct. 2024, doi: 10.3390/en17205199.
[23] J. Vazquez-Canteli, J. Kämpf, Z. Nagy, Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration, Energy Procedia, vol. 122, pp. 415–420, Sep. 2017, doi: 10.1016/j.egypro.2017.07.429.
[24] M. Cordeiro-Costas, D. Villanueva, P. Eguia-Oller, E. Granada-Alvarez, Intelligent energy storage management trade-off system applied to Deep Learning predictions, Journal of Energy Storage, vol. 61, p. 106784, May 2023, doi: 10.1016/j.est.2023.106784.
[25] I. Lytvinchuk, B. Savenko, S. Danchuk, Heating optimization system in a smart home based on fuzzy logic and integration with cloud services, Computer Systems and Information Technologies(1), 2025, 16–28, doi: 10.31891/csit-2025-1-2.