POLITECNICO DI TORINO
Repository ISTITUZIONALE

Extracting a transferable rule-based controller from deep reinforcement learning agents with validation across multiple scenarios for heat pump systems with thermal storage

Original
Extracting a transferable rule-based controller from deep reinforcement learning agents with validation across multiple scenarios for heat pump systems with thermal storage / Piscitelli, M.S., Mele, A.A., Razzano, G., Capozzoli, A.. - In: APPLIED ENERGY. - ISSN 0306-2619. - 412:(2026). [10.1016/j.apenergy.2026.127661]

Availability:
This version is available at: 11583/3009354 since: 2026-03-29T18:31:39Z

Publisher:
Elsevier Ltd

Published
DOI:10.1016/j.apenergy.2026.127661

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

16 November 2026

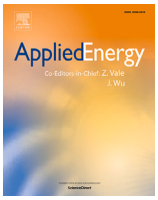
Applied Energy 412 (2026) 127661



Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apen



Extracting a transferable rule-based controller from deep reinforcement learning agents with validation across multiple scenarios for heat pump systems with thermal storage

Marco Savino Piscitelli*, Alessandro Aniello Mele, Giuseppe Razzano, Alfonso Capozzoli

Department of Energy (DENERG), TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy

HIGHLIGHTS

- A rule extraction (RE) framework distills control logic from optimized DRL agents.
- The process is tested on a heat pump equipped with a thermal energy storage.
- The RE controller was benchmarked against the best DRL agent and a standard RBC.
- The RE controller was transferred across eight scenarios without retraining.
- The RE controller outperformed the RBC and closely matched DRL performance.

ARTICLE INFO

Keywords:
Deep reinforcement learning
Rule extraction
Building energy management
Co-simulation environment
HVAC systems
Optimal control

ABSTRACT

This paper proposes a framework to derive interpretable and transferable Rule Extraction (RE) controllers from Deep Reinforcement Learning (DRL) control agents. Despite their strong performance in predictive energy management, model free and model based controllers face limited adoption due to complexity, computational cost, and low interpretability. The proposed RE approach translates DRL policies into simple, transparent, and scalable rule sets. The approach was tested on a heating system composed of an air-to-water Heat Pump (HP), a Thermal Energy Storage (TES) tank that also serves as a hydraulic separator, and fan coil units in a single-zone building, under the climate conditions of Turin (Northern Italy). The methodology follows a multi-stage process. High performing DRL agents were first identified through co-simulation and an optimization routine. A decision tree procedure then extracted if else rules, aggregated into a single interpretable RE controller. This controller was benchmarked against the best DRL agent and a standard Rule Based Controller (RBC). In the reference case, the RE outperformed the RBC and matched the DRL, retaining the original control logic. The RE controller was deployed without retraining across eight scenarios with different occupancy profiles, climates, and TES sizes. In all cases, it ensured stable indoor temperature control and outperformed the RBC, achieving at least 7% energy savings, 23% peak power reduction, and about 13% improvement in heat pump COP. The RE controller was also compared with scenario specific DRL agents optimized for each context. Without retraining, it achieved on average 74% of their energy savings and 86% of their peak power reduction relative to the RBC. The results confirmed the robustness and transferability of the framework, highlighting its potential for real-world deployment.

1. Introduction

Buildings account for nearly 30% of global final energy consumption and about 26% of energy-related CO_2 emissions, which makes them a central sector in the transition to low-carbon energy systems [1]. Once treated as passive consumers, buildings are now expected to operate as active components of future energy networks, interacting with both the power grid and local renewable generation [2,3]. Their large share of electricity demand, combined with thermal inertia and the increasing use of controllable technologies such as HPs, thermal storage, electrical storage, and responsive loads, makes them well suited to provide demand-side flexibility [4,5]. This flexibility allows buildings to adjust their energy consumption in response to external signals and system

* Corresponding author.
Email address: marco.piscitelli@polito.it (M.S. Piscitelli).

https://doi.org/10.1016/j.apenergy.2026.127661
Received 30 December 2025; Received in revised form 14 February 2026; Accepted 6 March 2026
Available online 13 March 2026
0306-2619/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Acronym

BACS	Building Automation and Control Systems
CART	Classification and Regression Trees
COP	Coefficient of Performance
CTREE	Conditional Inference Trees
DRL	Deep Reinforcement Learning
DT	Decision Tree
FMI	Functional Mock-up Interface
FMU	Functional Mock-up Unit
HP	Heat Pump
HVAC	Heating, Ventilation, and Air Conditioning
KPI	Key Performance Indicator
MPC	Model Predictive Control
OTL	Online Transfer Learning
PI	Proportional Integral
PID	Proportional Integral Derivative
RBC	Rule Based Controller
RE	Rule Extraction
SAC	Soft Actor-Critic
TES	Thermal Energy Storage
TL	Transfer Learning
XAI	eXplainable Artificial Intelligence

needs. By preheating or pre-cooling indoor spaces and by charging or discharging storage devices buildings can adapt their consumption patterns according to price signals or grid requests, contributing to overall system efficiency [3,6,7]. In addition, coordinated operation of building resources can provide ancillary services such as frequency regulation and reserve capacity, thereby enhancing the stability and reliability of the power grid itself [8]. Exploiting this flexibility poses a complex control problem. Building energy systems must operate in a predictive way, anticipating future conditions such as weather, occupancy, and signals from the grid [9]. They must optimize multiple and often conflicting objectives, including cost reduction, carbon mitigation, thermal comfort, and equipment reliability [10,11]. They must also adapt continuously to uncertainty from forecast errors, user behavior, and variable grid signals [12]. This combination of requirements results in a high-dimensional, stochastic, and time-coupled optimization task.

Traditional RBC and Proportional Integral Derivative (PID) controllers are unable to address such complexity [13–15]. Model Predictive Control (MPC) has emerged as a reference approach, since it embeds forecasts and optimization in a receding horizon framework [16]. However, its performance depends strongly on the availability of accurate and computationally efficient models, and deploying it across heterogeneous buildings requires significant engineering and calibration effort [17]. More recently, data-driven methods and DRL have been investigated as alternatives, since they can learn control policies directly from data or simulations and adapt to nonlinear dynamics without explicit models [18]. DRL has demonstrated the ability to derive near-optimal control strategies in complex building energy scenarios at different scales, achieving significant improvements in cost as energy savings and well as flexibility provision [19–21]. However, the resulting policies are encoded in high-dimensional neural networks that are difficult to interpret [22]. This lack of transparency limits trust, makes it challenging to validate safety and comfort requirements, and represents a deployment barrier in practice [23]. Moreover, DRL policies are typically tailored to the specific environment in which they are trained, which reduces their scalability and complicates their transfer to different buildings, system configurations, or operating conditions [24]. To address the transferability challenge, the literature primarily addresses this issue through Transfer Learning (TL), a machine learning method that involves leveraging a model trained on a specific task (source task) in a particular domain to address a new but related task (target task) [25]. Several studies illustrate the benefits of TL. For example, Coraci et al. [26] applied TL to transfer a DRL controller from a source building to target buildings. These target buildings, while sharing the same geometry and energy system as the source, differed in key characteristics such as weather conditions, occupancy patterns, and building envelope efficiency levels. Their Online Transfer Learning (OTL) approach demonstrated superior performance compared to an RBC but inferior performance relative to a specifically trained offline DRL, confirming its advantage in adapting to new scenarios without requiring a new model. In addition to adapting controllers, TL can also be used to reduce training time. Lissa et al. [27] introduced parallel TL to transfer control policies among multiple DRL agents during training, achieving a five-fold reduction in training time and a 10% decrease in energy consumption. However, while the TL approach can enhance model transferability and reduce training time, it does not resolve the other primary challenges of DRL controllers, namely the lack of interpretability and the high deployment effort for real-time applications. These issues create a clear mismatch with the needs of practitioners, who favor transparent and verifiable control logic, often expressed as simple if–then rules that can be understood, adjusted, and validated during operation. A promising way to overcome these limitations is the use of RE techniques, which aim to approximate complex machine learning models with simpler, human-readable representations. By translating the behavior of a trained DRL agent into a set of if–then rules, the resulting control strategy becomes transparent and interpretable. This provides a clear advantage for building energy applications, where practitioners need to understand, validate, and adjust the logic of controllers before deployment. Although RE has been explored within the broader machine learning community, its application to the control of energy systems remains limited. Bridging this gap holds the potential to combine the high performance of DRL with the interpretability and operational practicality of rule-based control, laying the foundation for scalable and trustworthy deployment of advanced control strategies in real building environments.

A key advantage of RE lies in its potential to enhance transferability. Once a DRL-derived policy is reformulated into a compact set of interpretable rules, it becomes easier to adapt to different systems, building typologies, or operating conditions, thereby reducing the need to retrain DRL agents from scratch. Additionally, presenting control logic in a rule-based format increases acceptance among building operators and engineers, who are already accustomed to supervisory control schemes and can readily integrate such rules into existing automation infrastructures.

While promising, the transferability of RE-based controllers across diverse applications has not yet been demonstrated. This remains a central objective of ongoing work, aiming to assess whether interpretable controllers extracted from a single scenario can maintain high performance under varying boundary conditions.

In this context, this paper introduces a methodological framework for developing interpretable and generalizable controllers by extracting rules from advanced DRL control policies. Rather than extracting rules from a single advanced DRL agent, the proposed framework integrates a multi-stage rule extraction from a set of sub-optimal DRL controllers. This enables an ensemble approach aimed at retaining decision patterns that are consistent across multiple independently optimized DRL agents, enhancing cross-case robustness. The approach is demonstrated on a reference European heating system layout comprising an air–water HP, a hot-water TES tank also acting as a hydraulic separator between primary and secondary loops, and fan coil units serving a single-zone building. The considered control actions are the on/off status of the HP and the setpoint of the supply water temperature.

A co-simulation environment developed in Modelica was used both to train the DRL controllers and to evaluate the performance of the extracted rule-based controllers. The process began with an optimization pipeline that generated a set of high-performing DRL agents, all trained on a single reference scenario. From each agent, a Decision Tree (DT) was extracted for each control action, resulting in a collection of if–then rules that approximated the learned policy. These rule sets were then combined using a majority-vote strategy to form a consolidated control logic. This ensemble was further distilled into a single interpretable RE controller and was benchmarked against two references: the best-performing DRL agent and a reference RBC in order to evaluate its performance and behavior in terms of energy consumption and control of indoor air temperature.

To assess its transferability, the RE controller was then deployed without any retraining or structural adjustments in eight additional scenarios, each characterized by different occupancy profiles, climate conditions, and TES sizes. Its performance was then benchmarked against two baselines, the reference RBC and a scenario-specific DRL controller retrained and optimized from scratch for each new scenario. This experimental setup allowed for a preliminary assessment of the potential transferability of the RE approach, which remains a key research objective.

In this perspective, Section 1.1 reviews the existing literature on RE processes to emulate advanced control strategies. Subsequently, Section 1.2 presents and explores the contributions of this study along with the innovative elements it introduces.

1.1. Related works

RE is a research area within the broader field of eXplainable Artificial Intelligence (XAI). In the context of control, it involves the use of surrogate models to approximate and explain the behavior of complex black-box policies [28]. In the control field, surrogate models are deliberately simpler and more interpretable representations that mimic the input–output mapping of a more sophisticated controller. When applied to DRL or MPC, this approach makes it possible to derive explicit if–then rules that capture the decision-making logic of the underlying policy. As a result, RE contributes to bridging the gap between the high performance of advanced control strategies and the transparency required for their deployment in real Heating, Ventilation, and Air Conditioning (HVAC) systems.

The application of RE processes for energy management applications in buildings and energy systems offers several significant advantages: [29–33]:

- RE mimics complex control strategies, such as MPC or DRL control, using interpretable and understandable rules;
- If-then rules can be integrated into existing Building Automation and Control Systems (BACS) without requiring enhanced computational resources;
- The simplicity of the extracted rules makes them well suited for real-time applications, overcoming the higher computational time required by MPC or DRL controllers;
- It is possible to formulate a set of rules that are more generalized, preventing overfitting to specific conditions and ensuring applicability across different zones or buildings, thereby enhancing flexibility and scalability.

The literature provides consistent evidence that RE can reproduce the performance of advanced control strategies with only marginal losses while greatly reducing computational effort. In building applications, studies have shown that DRL policies can be distilled into DT rules that achieve almost identical energy performance. For example, simplified rules derived from deep Q-learning in a cooling system led to energy savings with only a 1.2% deviation from the original DRL controller [34]. Similar results were reported in power system emergency control, where an information gain rate-based weighted oblique DT (IGR-WODT) was proposed to replace DRL policies for under-voltage load shedding, improving transparency and enabling deployment on resource-constrained devices [30].

A larger body of work has applied rule extraction to MPC controllers. Yu and Pavlak [10] demonstrated that decision-tree surrogates could match the accuracy of more complex models such as random forests in learning MPC behavior. Their extracted rules preserved 95–97% of the energy savings and 89–92% of the cost savings of a detailed MPC. Other works combined optimization and RE to derive practical control strategies. In Ref. [31,35], genetic algorithms were used to identify optimal operational parameters across climates and occupancy scenarios, and decision trees were then employed to extract interpretable rules that reproduced the original outcomes with minimal performance loss. A multi-objective MPC solution for smart glazing was also distilled into a compact set of rules, reducing complexity and reproducing control signals with up to 65% accuracy [32]. In Ref. [33], a classification learning approach was applied to detailed MPC outputs from 27 rooms of an institutional building: DTs achieved 27–33% energy savings compared to 18–42% with MPC. Similarly, Ref. [36] showed that MPC-based ventilation strategies could be approximated with multi-logistic regression, retaining 70–90% of the energy savings with a much lower computational cost.

More recently, researchers have explored hybrid strategies that integrate RE with conventional rule-based control. Zhu et al. [37] introduced a learning-based MPC for HVAC systems, where Classification and Regression Trees (CART) algorithms approximated MPC decisions. While the MPC reduced energy costs by 13.89% and the surrogate by 12.27%, the computation time required for one control step dropped from 0.55 s to 0.002 s. Extending this concept, Zhu et al. [38] proposed the machine learning-enhanced lightweight rule-based control strategy (ML-RBC), which used XGBoost to automatically tune the parameters of predefined RBCs from offline MPC simulations. The resulting controller achieved nearly the same cost savings as MPC (21.95% vs. 23.07%) with a computational burden more than one thousand times lower.

Overall, the reviewed studies demonstrate that RE is an effective process to simplify advanced control strategies and make them interpretable, with minimal loss in performance compared to the original DRL or MPC policies. Most contributions have focused on showing that simplified rule sets can reproduce the behavior of complex controllers while reducing computational costs and enabling transparency. However, it remains unclear whether rules extracted in one setting can be reliably applied to others with different climates, occupancy patterns, or system configurations. Existing works have mainly evaluated rule extraction in the same environment used to generate the original controller, with only limited evidence of generalization across boundary conditions. Addressing this gap is critical for real-world deployment, since the practical value of RE lies not only in interpretability and simplicity, but also in the ability to scale across diverse building types and operating contexts without the need for retraining or extensive recalibration.

In this perspective, the next section presents the main contributions and the novel elements that this research aims to introduce.

1.2. Novelty and motivation

The current literature increasingly addresses the challenges associated with advanced control. Among the explored strategies, RE has shown strong potential to produce interpretable rule sets that enhance transparency, support stakeholder trust, and enable practical deployment. In parallel, TL has been widely applied to advanced controllers, especially DRL, to ensure generalization across different conditions without the need for retraining in each specific case.

Although TL has been extensively explored as a means to scale complex control policies, further investigation is needed into RE, particularly to assess whether interpretable controllers derived from DRL can deliver a meaningful degree of generalizability. In particular, it remains unclear

whether rules extracted in one context can be applied in others, such as systems with different climates, occupancy patterns, or configurations, without a significant drop in performance. To demonstrate the transferability of the extracted RE controller, this study integrates an ensemble-based RE process with a multi-action formulation, enabling successful deployment across previously unseen scenarios without requiring retraining.

In this perspective the contributions of this work can be summarized as follows:

- **Conceptualization of a DRL-based controller with a generalizable state representation**. A DRL controller was developed using a standard Soft Actor-Critic (SAC) algorithm. Particular attention was given to the selection and normalization of input variables, which include external temperature, HP and TES measurements from the primary loop, and the operational schedule of the secondary loop (used as a proxy for occupancy). This state representation is designed to be as much as possible system-agnostic ensuring that the controller can be adapted to different secondary loop configurations.
- **Ensemble-based RE from a family of optimized DRL controllers**. A novel methodology was introduced for RE that does not rely on a single trained policy but leverages a family of optimal DRL controllers generated using the Optuna optimization framework. Therefore, the ensemble approach is used to identify robust decision patterns across different DRL agents, with the goal of retaining common decision patterns and filtering out scenario-specific ones to improve cross-scenario transferability rather than approximating a single agent. A final RE step is applied to this ensemble consensus to capture the collective logic, improving generalization while simplifying the extracted controller and increasing the interpretability of the control policy, yielding a single, transparent rule set.
- **Definition of a multi-action rule set through separate decision paths**. The RE controller was designed to handle multiple actions by organizing the rule set into two separate decision paths. The HP activation was formulated as a classification problem, using inputs such as the TES state and occupancy schedule, and trained on the complete DRL-generated dataset. In contrast, the supply water temperature setpoint adjustment was modeled as a regression task, restricted to time steps in which the HP was active. This decoupled formulation produces two specialized rule sets, each dedicated to a specific control objective. The separation simplifies the extraction procedure and improves rule readability, as each rule set relies only on the features relevant to its task. This structure supports both predictive accuracy and interpretability.
- **Evaluation of the extracted controller across heterogeneous scenarios**. The RE controller was evaluated in multiple scenarios featuring diverse climate conditions and occupancy patterns. Its performance was compared with a reference RBC and with a DRL controller trained and optimized specifically for each scenario to assess and highlight the transfer and robustness of the proposed method.

The rest of the paper is structured as follows. Section 2 presents the case study, providing details on the building, the heating system, and the transfer scenarios under consideration. Section 3 outlines the methodology, including the development of the DRL controllers, the RE framework, and the baseline RBC control logic. Section 4 presents the results, highlighting the performance of the DRL, RE, and RBC controllers across the different transfer scenarios. Finally, Section 5 discusses the implications of the findings, and Section 6 provides the conclusions of the study highlighting its potential benefits and related future research directions.

2. Case study

This section introduces the case study used to evaluate the proposed methodology. It focuses on a single zone building equipped with an air source HP and a hot water TES tank for space heating. This system configuration is increasingly adopted in the European building sector as part of the shift toward more efficient, electricity-based HVAC solutions. In this context, the integration of TES is widely recognized as a key enabler for optimizing system operation, enhancing demand-side flexibility, and contributing to grid decarbonization goals [7].

To evaluate the robustness and transferability of the obtained RE controller, the case study considers multiple scenarios with varying boundary conditions. These include different geographic locations, namely Turin, Bolzano, and Naples, which represent distinct climate zones, as well as diverse occupancy patterns. For each scenario, the TES tank was appropriately resized to match the demand for space heating. Specifically, Section 2.1 describes the main features of the building and outlines the weather conditions and occupancy patterns used to define the different operational scenarios for testing the RE controller. Building on this, Section 2.2 details the configuration of the HVAC system implemented across all scenarios.

2.1. Building overview and scenarios of study

The building under consideration is an open space office. Table 1 reports the total surface areas and thermal transmittance values of the envelope components. These values vary according to the climatic conditions of each site, following the Italian classification of climate zones defined by DPR 412/93 [39]. Turin and Bolzano fall within Zone E, characterized by 2100 to 3000 heating degree days, while Naples is located in Zone C, with 900 to 1400 heating degree days. Consequently, for the building envelope higher levels of thermal insulation were adopted in Turin and Bolzano compared to Naples. This differentiation reflects established construction practices that respond to the severity of local climate conditions during winter season.

After defining the building model, the evaluation of the RE controller is conducted across nine scenarios obtained by combining three climates with three occupancy schedules (Table 2). In particular the RE controller was evaluated for a reference scenario, corresponding to the climate of Turin with the Occupancy Base schedule, while the remaining eight scenarios were used to test the controller's transferability to different

Table 1
U-value and surface area of building elements.

Building Element	Surface Area [m^2]	U-value [W/m^2K]	
		Climate zone E	Climate zone C
Windows	20	1.2	1.75
External walls	210	0.28	0.68
Internal walls	55	3.15	3.15
Roof	130	0.35	1.18
Floor	130	0.7	1.32

Table 2
Scenarios considered for deriving the RE controller and evaluating its transferability.

Scenario ID	Location climate	Occupancy pattern
Turin - OB (reference)	Turin	Occupancy Base
Turin - O1	Turin	Occupancy 1
Turin - O2	Turin	Occupancy 2
Bolzano - OB	Bolzano	Occupancy Base
Bolzano - O1	Bolzano	Occupancy 1
Bolzano - O2	Bolzano	Occupancy 2
Naples - OB	Naples	Occupancy Base
Naples - O1	Naples	Occupancy 1
Naples - O2	Naples	Occupancy 2

boundary conditions. Specifically, Bolzano represents colder and more variable weather, with a temperature distribution similar to Turin (reference scenario) but with greater fluctuations; Naples, on the other hand, is characterized by a milder and more stable winter climate with higher average outdoor temperatures.

Regarding occupancy schedules, these were defined to reproduce different patterns of use:

- **Occupancy Base**: Conventional office hours from 08:00 to 18:00 on weekdays, with no weekend occupancy.
- **Occupancy 1**: Split workday with occupancy from 08:00 to 12:00 and 16:00 to 20:00 on weekdays, and from 08:00 to 14:00 on weekends.
- **Occupancy 2**: Continuous occupancy from 09:00 to 12:00 on weekdays, and from 12:00 to 18:00 on weekends.

Occupancy 1 and Occupancy 2 represent alternative patterns to the Occupancy Base, enabling the analysis of how variations in user presence affect the performance of the RE controller. In all scenarios, the secondary loop of the heating system is scheduled to start two hours before occupancy, ensuring that the indoor temperature setpoint is reached by the time occupants arrive.

2.2. Description of the heating system layout

The heating system considered in this study, illustrated schematically in Fig. 1, consists of an air to water HP connected to a hot water TES tank. The TES acts as a buffer that balances heat generation and demand while also serving as a hydraulic separator between the primary loop (HP to TES) and the secondary loop (TES to office). Heat is delivered through fan coil units, which serve as the emission subsystem and are designed to operate with low water supply temperatures (lower than 50 °C).

The system is supposed to be controlled through a rule-based strategy taken as a reference in the study:

- **Primary Loop Control System**: The primary loop is managed by controlling both the operational status of the HP (modHP) and its supply water temperature setpoint (TsetHP). This strategy is adopted as a reference in all scenarios and serves as the baseline for comparison with the DRL controllers and the obtained RE controller in order to evaluate their performance.

  The RBC logic (Fig. 2) is based on two setpoint thresholds depending on the secondary loop schedule, with a linear modulation of the supply water temperature to optimize Coefficient of Performance (COP) (Eq. (1)).

$$\mathrm{TSetHP} = 56\,^{\circ}\mathrm{C} - (T_{\mathrm{tank}} - 45\,^{\circ}\mathrm{C}) \tag{1}$$

- **Secondary Loop Control System**: The secondary loop control system is managed by a Proportional Integral (PI) controller that maintains the indoor zone temperature at the setpoint of 20 °C during occupied periods. The controller regulates a three way valve and a variable speed pump to adjust both the flow rate and the supply water temperature delivered from the storage tank to the fan coil units. The activation period begins two hours before the expected arrival time of occupants to ensure comfort conditions are met. During unoccupied periods, a setback strategy prevents the indoor air temperature from falling below 15 °C.

Despite the layout of the considered system being the same for all the considered scenarios, the TES was assumed to be differently sized according to the climatic conditions of the locations considered (i.e., winter conditions of Turin, Bolzano and Naples). The sizing of the TES and the performance modeling of the HP are designed as follows:

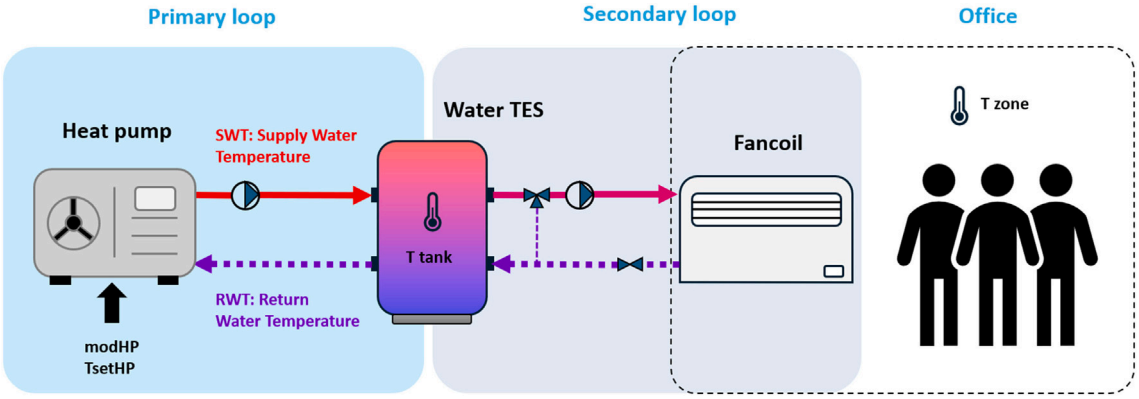


Fig. 1. Schematic representation of the considered heating system.

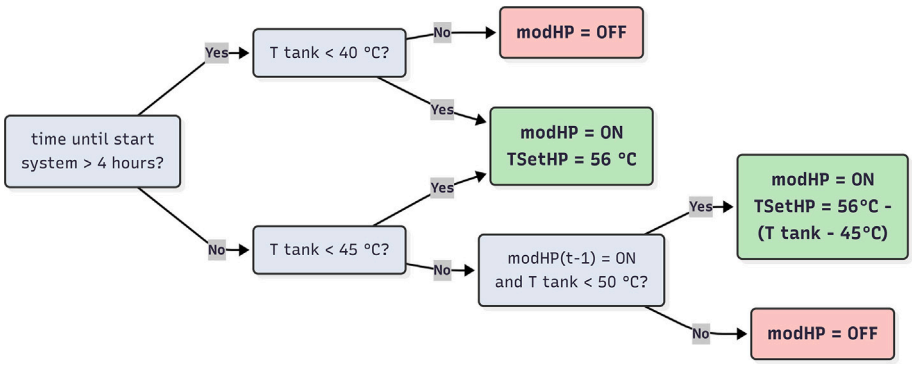


Fig. 2. RBC control logic.

Table 3
Sizes of the thermal energy storage for different climates.

City	Thermal power demand [kW]	Design outdoor air temperature [°C]	TES size [L]
Turin	10.1	−8	1300
Naples	8.5	2	1100
Bolzano	12.7	−15	1630

- **TES Sizing:** The sizes were determined based on the design external temperature specified in Ref. [40], corresponding to each location's climatic conditions. Using this data, the thermal power demand of the reference office was estimated. A 20% safety margin was then applied to account for uncertainties and ensure operational flexibility. The TES was sized to provide a continuous heat supply for a discharge duration of three hours. This corresponds to the amount of thermal energy that can be delivered by a water volume undergoing a temperature drop from 60 °C (fully charged) to 40 °C (discharged). Table 3 reports the thermal power demand, the design outdoor air temperature for each location and the obtained TES size expressed in liters of water.
- **HP COP Modeling**: The Clivet ELFOEnergy Edge - WSAN-YMi 31 HP model was used as a reference for performance calculations. The manufacturer's performance data, provided in accordance with UNI/TS 11,300 Part 4 [41], was used to model the COP. A polynomial interpolation on two variables (supply water temperature and external air temperature) was employed to derive the function used to dynamically compute the COP during the simulation.

3. Methodology and methods

This section presents the methodological framework adopted for the development and evaluation of the RE control strategy. The framework is organized into two main components: the development of the co-simulation environment, and the workflow used to extract the RE controller and benchmark it against the RBC and DRL controllers in both the reference case and transfer scenarios, as illustrated in Fig. 3.

3.1. Development of the co-simulation environment

The co-simulation framework was based on a high-fidelity Modelica model representing both the built environment and the HVAC system, developed using the open-source software OpenModelica [42]. All control strategies (DRL, RBC, RE) were implemented in Python. Co-simulation was enabled through the Functional Mock-up Interface (FMI) standard [43], which allows the complete Modelica model to be encapsulated into a standardized Functional Mock-up Unit (FMU). The `pyfmi` package [44] enabled Python to act as the master, managing the interactions between the external Python-based controllers and the FMU.

Within this setup, the OpenAI Gym framework [45] was employed to train the DRL control policy, enabling action execution, state observation, and policy updating based on the received reward. DRL controllers were trained using simulation data for the month of January. Subsequently, all comparative evaluations between the three main controllers (RBC, DRL, and RE) were conducted on a deployment episode that spans the period of January and February. All controllers operated in real-time using the simulation feedback from the FMU. This means that the evolution of the system state is directly determined by control actions, thereby exposing the controller to the dynamic evolution of the environment, which differs depending on the control logic used.

3.2. Workflow to extract and benchmark the RE controller

Once the co-simulation environment was set up, the workflow for extracting and benchmarking the RE controller was structured into the following steps:

- **Training of DRL controllers for the reference case**: The Optuna optimization framework [46] was used to identify a set of high-performing DRL controllers for the reference scenario. The best performing controller in each case was selected as a benchmark.

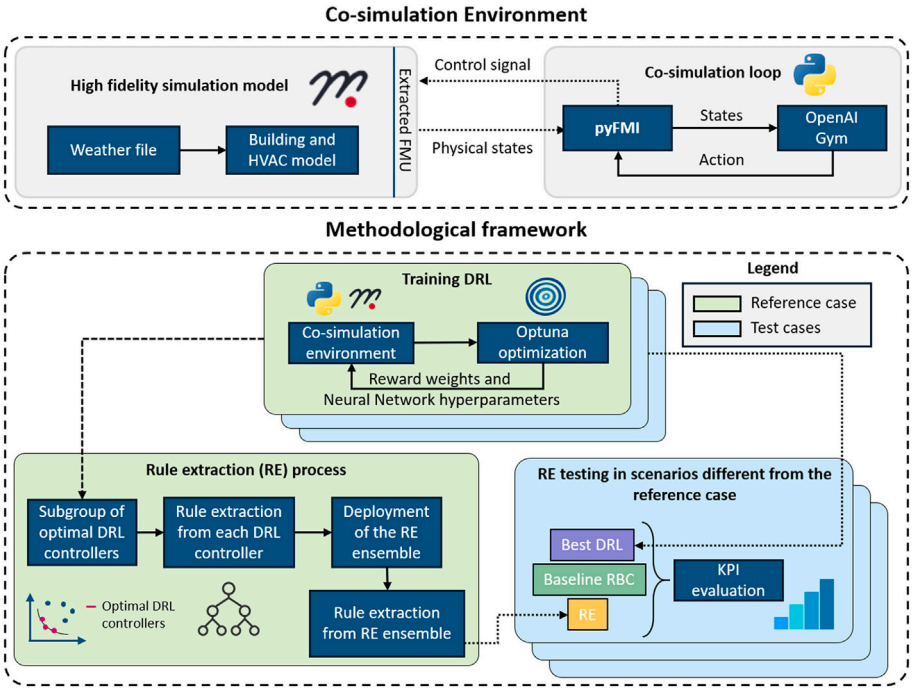


Fig. 3. Methodological framework.

- **Rule Extraction (RE)**: Control policies were extracted from the trained DRL controllers using DTs, resulting in a simplified and interpretable rule set in an `if-else` format.
- **Scenario testing**: The extracted RE controller was benchmarked against the best-performing DRL and the RBC in both the reference case and transfer scenarios, using a consistent set of Key Performance Indicators (KPIs).

3.2.1. Training of DRL controller for the reference case

DRL is a machine learning approach where an agent learns an optimal control policy by interacting with the environment and receiving feedback in the form of a reward signal [47]. This study adopts the SAC algorithm, a model-free DRL approach based on the actor-critic framework, which combines value-based and policy-based methods to optimize decision-making. The actor defines the control policy and selects actions, while the critic estimates the value function and assesses those actions [29]. A key feature of SAC is entropy regularization, which balances exploration and exploitation by encouraging stochastic policies. By jointly maximizing expected reward and policy entropy, the agent mitigates premature convergence to local optima and yields control strategies that remain robust under dynamic operating conditions [17]. SAC was selected due to its extensive adoption in continuous control applications [17,29,48], providing a stable and widely recognized benchmark for evaluating the proposed rule extraction framework. Table 4 lists the specific observations processed by the SAC algorithm to develop an optimal control policy with evidence of the variable names and their corresponding descriptions.

The learning process of the DRL agent is governed by several key hyperparameters:

- **Network architecture**: Defines the number of layers and neurons per layer in the neural network.
- **Learning rate (μ)**: Determines the rate at which the weights of the neural network are updated during the training process.
- **Batch size**: Specifies the number of samples used to perform each update to the neural network weights.
- **Discount factor (γ)**: Balances the importance of immediate versus future rewards, influencing the long-term strategy of the agent.

The controller provides a continuous action in the range [0, 1], which is decoded as both the activation state of the HP and its supply water temperature via a linear mapping (Eq. (2)).

$$\text{If } 0 < \text{action} < 0.5: \begin{cases} \text{modHP} = \text{OFF} \\ \text{TSetHP} = \text{RWT} \end{cases}$$

$$\text{If } 0.5 \le \text{action} \le 1: \begin{cases} \text{modHP} = \text{ON} \\ \text{TSetHP} = \left(\frac{60-\max(45,\text{RWT})}{1-0.5}\right)\text{action} + \max(45,\text{RWT}) \end{cases} \tag{2}$$

Specifically, when the action lies between 0 and 0.5, the HP is turned off (modHP = OFF), and the setpoint is set equal to the return water temperature (TSetHP = RWT), implying no active heating on the primary loop.

When the action is between 0.5 and 1, the heat pump is activated (modHP = ON), and the setpoint of supply water temperature (TSetHP) is linearly mapped based on the action value. The mapping scales the action to a temperature range between max(45 °C, RWT) and 60 °C, ensuring that the supply water temperature setpoint is always higher than or equal to the return water temperature (RWT), with 45 °C as a minimum threshold.

The reward function adopts a multi-objective structure consistent with approaches reported in prior studies [17,19,29]. It balances several objectives, including minimizing energy consumption, maintaining the TES temperature within safe operating limits, and reducing equipment stress as reported in Eq. (3). It is composed of three penalty components:

- **TES Penalty**: The TES penalty is applied when the water temperature of the TES T tank drops below a given thresholds (45 °C or 40°C based on system status). Specifically, the penalty increases quadratically as the temperature decreases further. This penalty component ensures that the controller keeps the TES temperature within an acceptable range, preserving thermal system's readiness to operate when needed.
- **Energy Penalty**: It accounts for both the HP and primary loop pump electric energy consumption, considering a quadratic function. This quadratic formulation penalizes high energy consumption more heavily and as a side effect tends to reduce peaks of electrical power demand.
- **Switching Penalty**: The switching penalty accounts for the frequency of on/off switching of the HP. It compares the current operational mode (0 if OFF and 1 if ON) of the HP with its mode in the previous time step (t-1). If a change in status is detected, a penalty is applied. This discourages excessive switching, which can lead to reduced equipment lifespan and unstable system operation.

$$R = -(\text{energy_penalty} + \text{tank_penalty} + \text{switch_penalty})$$

where

$$\text{TES penalty} = \begin{cases} k_1(1 + (45\ ^\circ\text{C} - T_{\text{tank}}))^2 & \text{if } 40 \le T_{\text{tank}} < 45 \\ & \text{and system on} = 1 \\ k_2(1 + (40\ ^\circ\text{C} - T_{\text{tank}}))^2 & \text{if } T_{\text{tank}} < 40 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Energy penalty} = (k_3\text{energy_el_heat_pump} + k_4\text{energy_el_hydronic_pump})^2$$

$$\text{Switch penalty} = \begin{cases} k_5|\text{modHP(t-1)} - \text{modHP}| & \text{if modHP} \neq \text{None} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

From Eq. (3) it can be seen that the reward function includes five weighting parameters: k_1, k_2, k_3, k_4 and k_5. These coefficients allow the analyst to tune the relative influence of each penalty component.

In this study, Optuna was used to optimize both the weighting parameters of the reward function and the hyperparameters of the DRL agent.

Table 4
Observations available for primary-loop control.

Observation	Description
`T ext`	Current outdoor air temperature [°C]
`T tank`	TES water temperature [°C]
`RWT`	Return water temperature to the HP [°C]
`Energy el heat pump`	Electricity consumption of the heat pump normalized to the maximum value of 1.5 kWh
`COP heat pump`	Coefficient of performance of the HP
`time until start occupancy`	Remaining time until the building occupancy begins
`time until end occupancy`	Remaining time until the building occupancy ends
`time until start system`	Remaining time until the secondary loop system is turned on
`time until end system`	Remaining time until the secondary loop system is turned off
`Tout 1step`	Expected outdoor air temperature after 15 minutes
`Tout 3step`	Expected outdoor air temperature after 45 minutes
`Tout 6step`	Expected outdoor air temperature after 90 minutes
`Tout 9step`	Expected outdoor air temperature after 135 minutes
`Tout 12step`	outdoor air external temperature after 180 minutes
`T tank (t-1)`	TES water temperature in the previous timestep [°C]
`Thermal energy flow`	Thermal energy withdrawn from the TES in the current timestep [kWh]

Specifically, Optuna allows users to dynamically define search spaces for optimization as detailed in Ref. [46]. In this work, Optuna is used to perform a two-step multi-objective optimization targeting the minimization of three KPIs evaluated over the testing period (i.e., January and February):

- **Energy Consumption [kWh]**: Total electricity consumed by the HP and the primary loop circulation pump.
- **Max Power Demand [kW]**: Maximum average electrical power demand of the HP and the primary loop circulation pump at a time step (15-min time interval).
- **Average TES water temperature violation [°C]**: Average magnitude of the temperature shortfall below the 45 °C setpoint in the thermal energy storage (TES), calculated only when the secondary loop is active.

In each optimization run, the DRL controller was trained and evaluated using a specific combination of reward function weights and model hyperparameters. The optimization process was structured as a sequential two-step procedure, with each step targeting a distinct set of parameters:

1. **Optimization of the reward function parameter weights**: In the first step, the optimization focused exclusively on tuning the weights of the reward function. During this phase, the neural network architecture and training hyperparameters were held constant, using reference values commonly adopted in the literature [26,49–51] (see Table 5) while the explored parameter search spaces are reported in Table 6. The optimization routine explored 50 configurations of weights and selected the one that led to the minimum of the optimization function among those that achieved an improvement of at least 5% in terms of energy consumption and maximum power demand in the testing episode compared to the reference RBC. The best weight configuration was then used to initialize the second optimization step.
2. **Optimization of the DRL agent hyperparameters**: In this step, the optimal reward function parameter weights identified in the previous step were fixed. The optimization then proceeded through 20 additional runs to explore and optimize the neural network architecture and the learning hyperparameters over the search spaces reported in Table 7.

At the end of this step among the obtained DRL agents that lead to an improvement of at least 5% in terms of energy consumption and maximum power demand in the testing episode compared to the reference RBC were selected. From this filtered set, the DRL controllers were ranked based on their ability to reduce energy consumption, limit maximum power demand, and minimize indoor air temperature violations. The latter metric quantifies the average deviation below the comfort threshold of 19 °C, considering only the periods when the indoor temperature falls below this lower bound.

The highest-ranked DRL controller, demonstrating the best overall trade-off among these metrics, was selected as the best performing DRL solution. This controller then served as the main benchmark for evaluating the performance of alternative control strategies.

In addition, to enable the RE in the next phase of the workflow, the largest odd-numbered subset of these high-performing DRL agents was selected.

Table 5
Hyperparameter combinations used for the optimization of reward function weights.

Neural network hyperparameters		Learning hyperparameters		
Hidden layers	Neurons per layer	Learning rate μ	Discount factor γ	Batch size
4	64	1e−4	0.99	64

Table 6
Search space for the optimization of the reward function parameter weights.

Reward function parameters	Weight search range
TES_penalty_parameter (k_1)	(1 - 200)
TES_penalty_parameter (k_2)	(1 - 200)
Energy_penalty_parameter (k_3)	(1 - 200)
Energy_penalty_parameter (k_4)	(1 - 200)
Switch_penalty_parameter (k_5)	(1 - 200)

Table 7
Search space for the optimization of the neural network architecture and the learning hyperparameters of the DRL agent.

Hyperarameters	Search space
Network architecture	*small*: (64, 64, 64, 64) *medium*: (128, 128, 128) *large*: (256, 256)
Learning rate μ	5e−4, 1e−4, 5e−3, 1e−3, 1e−2
Discount factor γ	From 0.95 to 0.99 (step 0.01)
Batch size	64, 128, 256

3.2.2. Development of the rule extraction (RE) controller from the sub-set of optimal DRL agents

This step aims to synthesize a single, interpretable RE controller that captures and emulates the collective knowledge embedded in the sub-set of optimal DRL agents. To achieve this, the control policy of each DRL agent is approximated by using DTs [52]. Specifically the DTs were developed to mimic actions of DRL where state-actions observations were collected from the simulation of the deployment episode.

In this study, the Conditional Inference Trees (CTREE) algorithm [53] was adopted and implemented using the `partykit` package in R [54]. The model is a binary and recursive DT algorithm that builds splits through a statistically grounded two-step procedure [55].

Specifically, each DRL agent was distilled by means of three DTs: a classification tree to predict the on/off state action of the HP (modHP), and two regression trees to predict the supply water temperature setpoint (TSetHP). The regression trees were trained considering only data pertaining to the operation of the HP separately for operational conditions when the secondary loop is active or inactive.

The tree growth was constrained by using stopping criteria to ensure model generalization. A maximum depth of 7 was applied to all trees to prevent overfitting [37,48]. The minimum number of observations per leaf node was set to 40 for the classification tree and 20 for the regression trees. The lower threshold for the regression trees accounts for the smaller sample size, as they were trained exclusively on data where the HP was active.

The input variables used to develop the decision trees are listed in Table 8, indicating whether each variable was used exclusively for the classification tree (modHP action), the regression trees (TSetHP action), or both.

Once the model is trained, decision rules are extracted by tracing each path from the root to a leaf node. Each path defines a conjunction of if–else conditions based on input variables, resulting in a specific prediction that captures the model's decision-making logic.

These rule sets (three per DRL agent) were then deployed collectively as an ensemble controller within the co-simulation environment as represented in Fig. 4. At each time step, a majority vote among the classification trees determined the on/off state of the HP. If the outcome was "activation", the supply water temperature setpoint was computed as the average of the values predicted by the corresponding regression trees of the active DRL agents; otherwise, it was set equal to the return temperature of the TES unit (RWT).

From the simulation of the deployment episode, the resulting state-action observations were collected to derive an RE controller capable of distilling and mimicking the behavior of the ensemble. The RE process

Table 8
Observations available to develop the classification tree (modHP action) and the regression trees (TSetHP action) as the basis of the RE controller.

Observation	Description	Decision tree
`T ext`	Current outdoor air temperature [°C]	TSetHP action
`T tank`	TES water temperature [°C]	modHP action
`time until start system`	Remaining time until the secondary loop system is turned on	TSetHP action; modHP action
`time until end system`	Remaining time until the secondary loop system is turned off	TSetHP action; modHP action
`T tank (t-1)`	TES water temperature in the previous timestep [°C]	modHP action
`Thermal energy flow`	Thermal energy withdrawn from the TES in the current timestep [kWh].	TSetHP action

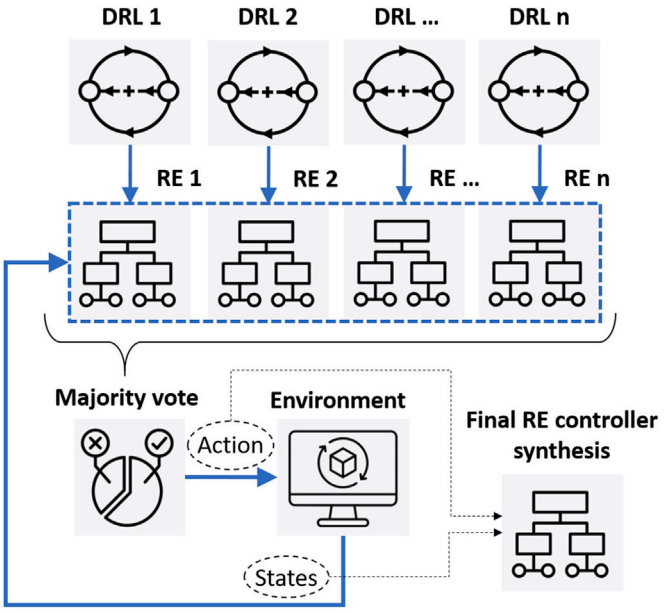


Fig. 4. Methodological framework for the development of the RE controller.

followed the same procedure used for each individual DRL agent: a classification tree was developed to model the HP activation (modHP), and two separate regression trees were trained to predict the supply water temperature setpoint (TSetHP), depending on whether the secondary loop was active or not during HP operation. The input variables used to train this final RE controller were the same as those reported in Table 8.

Once the final rule set was extracted, the rules pertaining to the supply water temperature setpoint (TSetHP) were refined through an expert-driven post-processing phase. Firstly, to improve real-world applicability, all the predicted setpoint values were rounded to the nearest 0.5 °C. Secondly, where possible, decision rules that led to the same output after rounding were merged in order to reduce complexity as much as possible. Eventually, to further improve reliability of the RE controller, an expert safeguard rule was added to the obtained rule set. This rule is activated if the RE controller fails to maintain sufficient TES temperature during specific timesteps that would compromise a proper operation of the fan coil unit.

The logic behind the safeguard rule is illustrated in Fig. 5.

3.2.3. Controller benchmarking in the reference and transfer scenarios

In the final step, the developed RE controller is benchmarked against the best-performing DRL controller and the baseline RBC in the reference scenario. This evaluation, based on a set of KPIs, begins with the reference case and is subsequently extended to the eight transfer scenarios.

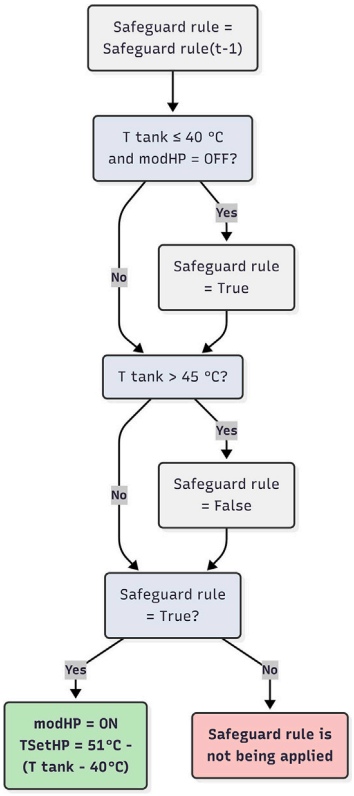


Fig. 5. Logic of the safeguard control rule.

The KPIs used for benchmarking the controllers were evaluated from the results obtained in the deployment episode and are listed as follows:

- **Energy Consumption** [kWh]: Total electricity consumed by the HP and the primary loop circulation pump.
- **Max Power Demand** [kW]: Maximum average electrical power demand of the HP and the primary loop circulation pump during a time step (15-min time interval).
- **Average COP**: The average COP of the HP evaluated during its operational hours.
- **Time Violation** [%]: The percentage of time during the occupancy period with an indoor air temperature that is lower than the lower bound of the comfort region i.e., the 19 °C.
- **Average Temperature Violation** [°C]: The average temperature deviation below the 19 °C threshold during periods of violation.

These indicators offer a comprehensive assessment of control performance across energy usage, peak demand, and indoor air temperature control.

In the reference scenario, the comparison aimed to assess the ability of the RE controller to compete against the best DRL agent identified through the optimization process, while also outperforming the simpler RBC. This evaluation served to verify whether the RE approach could retain the effectiveness of advanced control strategies in a more interpretable and practical format.

In the transfer scenarios, the focus shifted to testing the generalizability of the RE controller. Derived from the reference scenario, it was deployed without retraining and compared to two benchmarks in each new case: (i) a scenario-specific DRL agent trained from scratch using the same optimization procedure (see Section 3.2.1), and (ii) the baseline RBC. This setup was intended to evaluate the robustness and adaptability of the RE controller across varying conditions, including different occupancy profiles, climates, and TES configurations.

4. Results

This section presents the results of the study, organized according to the three main phases of the proposed methodology. First, the outcomes of the DRL training phase are discussed, highlighting the optimal hyperparameter values identified for the best-performing controllers across the nine evaluated scenarios (one reference and eight transfer scenarios). Second, the decision trees derived during the RE phase for the reference scenario are illustrated, and the corresponding control logic is analyzed. Finally, the results of the comparative benchmark are introduced, evaluating the performance of the RE, DRL, and RBC controllers across all scenarios.

4.1. Identification of the best DRL agents and hyperparameter optimization

The outcomes of the DRL training phase are analyzed by assessing the performance of the controllers optimized via Optuna across the nine evaluated scenarios, following the two-step procedure described in Section 3.2.1.

For completeness, the results of the optimization process for the reference scenario are illustrated in Fig. 6. The figure shows the performance of different DRL agents in terms of energy consumption (x-axis) and maximum power demand (y-axis) during the deployment episode. Blue points represent agents from the reward weight optimization phase, while green points correspond to those from the subsequent tuning of neural network parameters. The dashed yellow line indicates the 5% performance threshold relative to the baseline RBC, and only agents outperforming this threshold in both metrics were retained. Among them, the top performing controller (red point) was selected based on energy consumption, maximum power demand, and indoor air temperature violations.

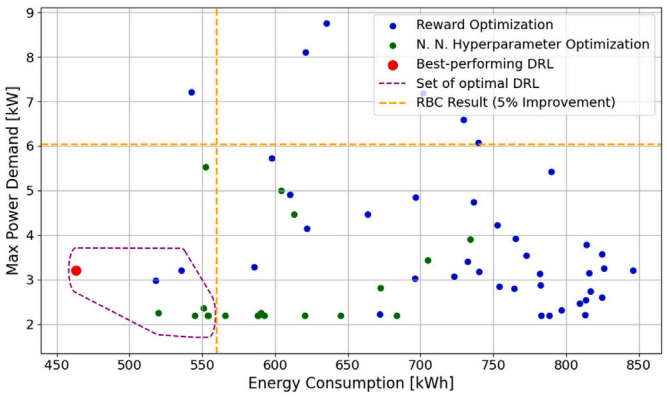


Fig. 6. Performance of different DRL agents in terms of energy consumption (x-axis) and max power demand (y-axis) during the deployment episode.

The subset of controllers used for the RE phase is highlighted by the dashed purple boundary. To support majority voting during RE, the set was reduced to the highest odd number of eligible controllers by removing the lowest-performing agent (in this case 7 DRL agents). The best DRL controller (red point) was selected as the benchmark for comparisons with both the RBC and the RE controllers.

Fig. 7 and Table 9 report the optimized reward weights and learning parameters for the best-performing DRL controller in each scenario. The significant variability of these values across scenarios underscores the strong impact of scenario-specific conditions on controller configuration and confirms the need for dedicated optimization in each case. In summary, this optimization procedure identifies suitable reward weights, hyperparameters, and network architectures for each scenario. Only DRL agents trained from scratch that outperformed the baseline by at least 5% were retained. The resulting scenario-specific controllers approximate the upper performance bound within the explored search space and provide a consistent benchmark for evaluating the transfer capabilities of the RE controller.

Table 9
Optimized hyperparameters for the best-performing DRL controller in each scenario.

Scenario	Learning rate μ	Batch size	Discount factor γ	Network architecture
Turin - OB	0.005	64	0.96	small
Turin - O1	0.001	256	0.96	small
Turin - O2	0.01	64	0.96	large
Bolzano - OB	0.01	256	0.96	medium
Bolzano - O1	0.005	64	0.95	large
Bolzano - O2	0.005	64	0.95	large
Naples - OB	0.0001	64	0.99	small
Naples - O1	0.005	64	0.95	medium
Naples - O2	0.005	128	0.97	medium

4.2. Rule extraction from DRL agents in the reference case

Based on the identified subset of seven high-performing DRL agents for the reference scenario, the RE controller was derived following the procedure detailed in Section 3. Specifically, individual RE controllers were first extracted from each agent, and a majority voting scheme was implemented to evaluate the ensemble during a deployment episode. The process concluded with the extraction of a final, consolidated rule set that distilled the common logic across all individual controllers.

As in previous cases, the RE process involved generating a classification DT to determine the on/off state of the HP, along with two

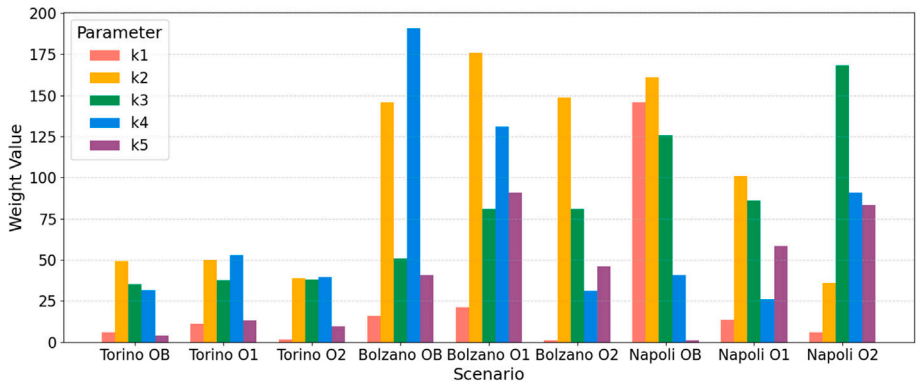


Fig. 7. Optimized reward weights for the best-performing DRL controller in each scenario.

regression DTs to select the supply water temperature that were separately developed for the active and inactive states of the secondary loop.

To enhance interpretability, the resulting RE control logic developed for the reference scenario is described and visualized in form of decision trees. Specifically, the two regression trees are combined into a single graphical representation, with the operational state of the secondary loop serving as the initial branching condition. For clarity, terminal nodes are color-coded: in the classification tree (Fig. 8), red indicates a predicted 'off' state of the HP and green indicates 'on' while in the regression tree (Fig. 9), a blue-to-red gradient reflects lower to higher predicted supply water temperature setpoints.

From Fig. 8 it can be inferred that the on/off state DT reached the maximum allowed depth of seven. Its structure reveals that schedule-related features are the most influential for predicting this action, particularly the time remaining until system deactivation (`time until end system`) and time until activation (`time until start system`). These observations, combined with current and previous TES water temperatures, form the foundation of the RE control logic. The use of past temperatures indicates that the model has captured aspects of the TES's thermal dynamics from the DRL agents, enabling it to differentiate effectively between charging and discharging patterns over time.

Fig. 9 clearly shows that the control logic adapts to the operational state of the secondary loop. When the loop is active, decisions are primarily driven by the time remaining until deactivation and the outdoor air temperature (`T ext`), along with the observed thermal energy withdrawn from the TES tank. When the loop is inactive, the logic is mainly based on the time until the next activation and `T ext`.

Eventually, by translating the DTs into a set of if-then rules, the final RE controller consists of a set of 34 decision rules, where 14 are used for the on/off operation of the HP and 20 for the control of the supply water temperature setpoint.

To understand the behavior of the RE control logic the Fig. 10 displays for the 5 KPIs introduced in Section 3 the performance achieved against the best DRL and the RBC in the reference scenario. Fig. 11 complements this by showing the cumulative energy consumption over the same evaluation period.

Eventually Table 10 presents percentage variation in energy performance indicators of the DRL and RE controllers with respect to the RBC, evaluated in the reference scenario and calculated as expressed in Eq. (4).

$$\text{Energy Saving } [\%] = \frac{E_{\text{RBC}} - E_{\text{Ctrl}}}{E_{\text{RBC}}} \times 100 \tag{4a}$$

$$\text{Max Power Reduction } [\%] = \frac{P_{\text{max, RBC}} - P_{\text{max, Ctrl}}}{P_{\text{max, RBC}}} \times 100 \tag{4b}$$

$$\text{COP Increase } [\%] = \frac{\text{COP}_{\text{Ctrl}} - \text{COP}_{\text{RBC}}}{\text{COP}_{\text{RBC}}} \times 100 \tag{4c}$$

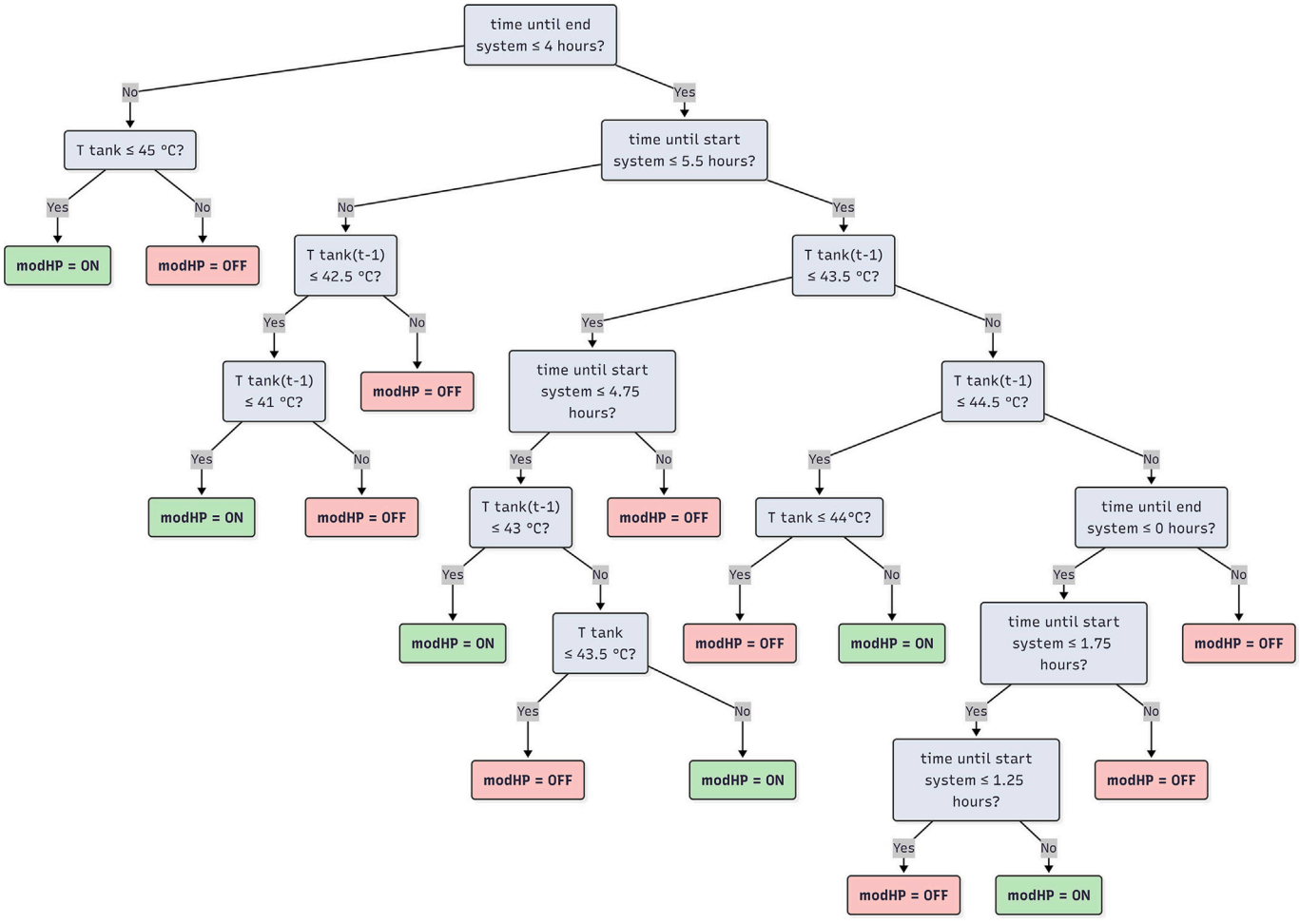


Fig. 8. Classification tree for the estimation of modHP action.

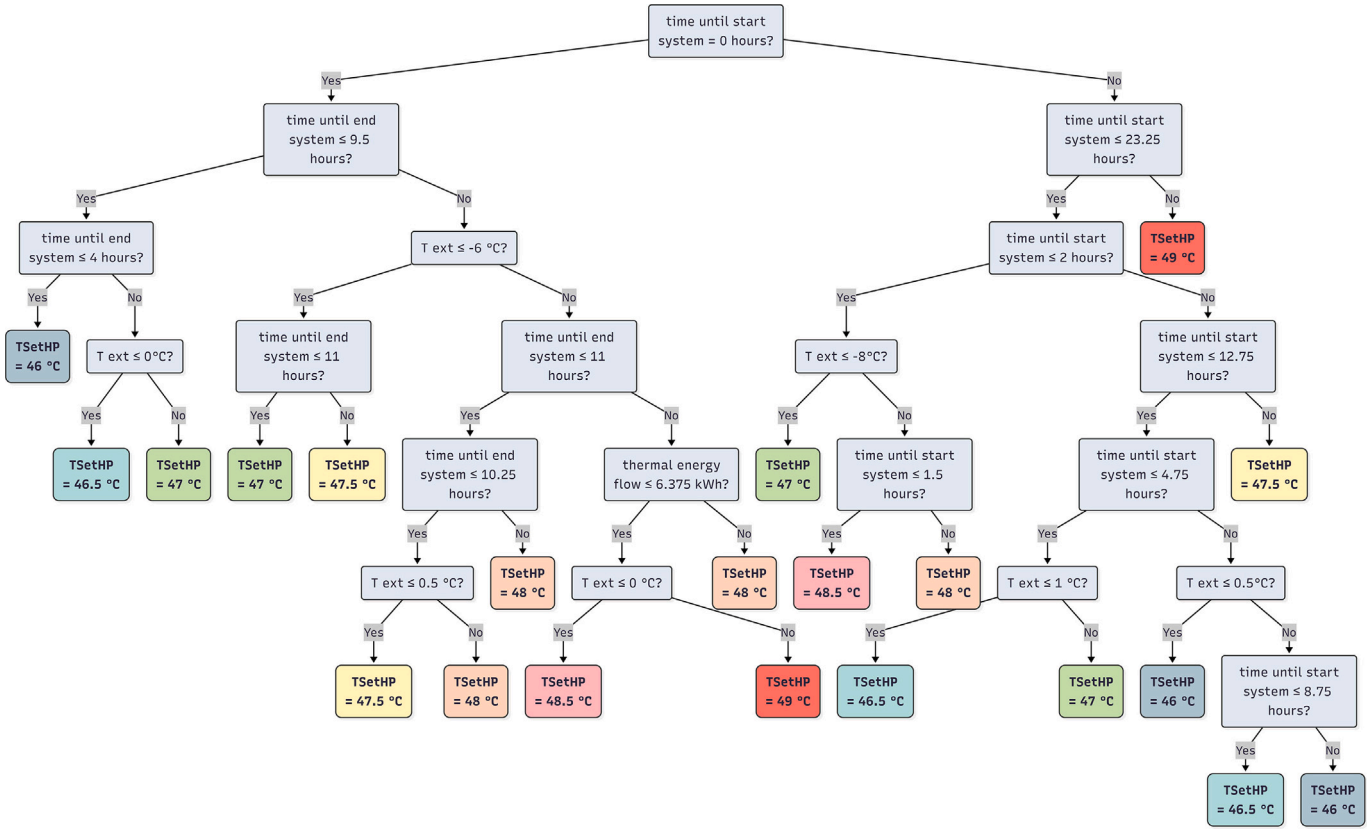


Fig. 9. Regression tree for the estimation of the TSetHP action.

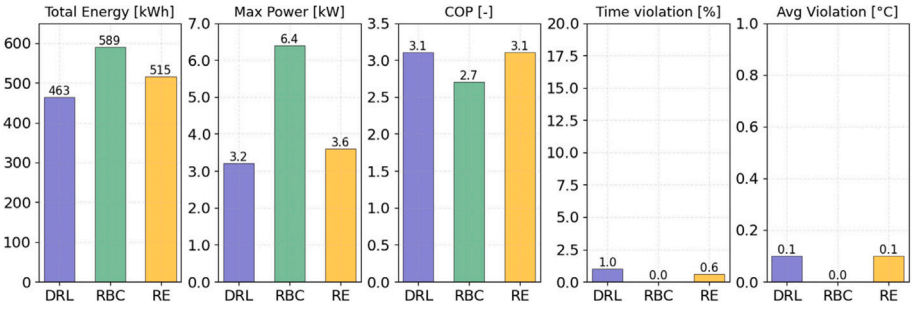


Fig. 10. Comparison between RBC, DRL and RE controller according to different KPIs in the reference scenario.

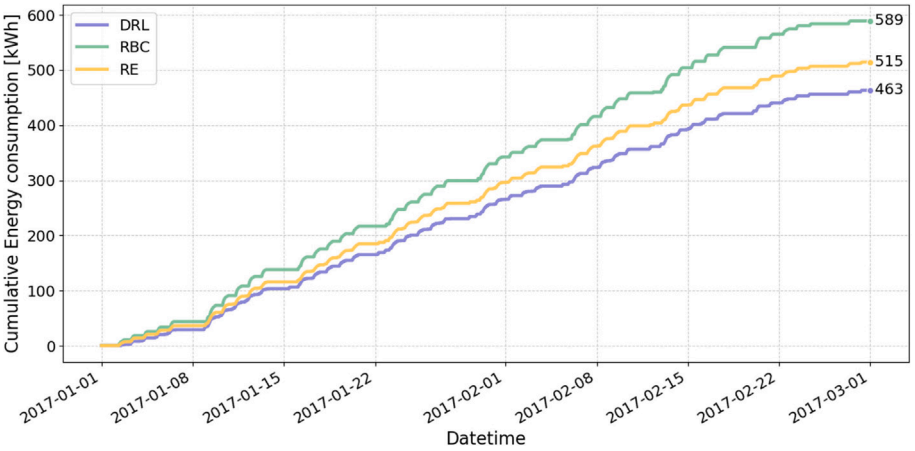


Fig. 11. Cumulative electrical energy consumption over the testing episode achieved by the RBC, DRL and RE controllers in the reference scenario.

Table 10
Percentage variation of energy performance metrics for the DRL controllers and RE controller relative to the RBC baseline for the reference scenario.

Scenario	Energy saving [%]		Max power reduction [%]		COP increase [%]	
	DRL	RE	DRL	RE	DRL	RE
Turin OB	21.4	12.6	49.6	43.9	14.8	15.6

In the equations, 'Ctrl' refers to either the DRL or RE controller.

To ensure a fair comparison, it was first necessary to confirm that all controllers maintained a similar level of indoor air temperature conditions. This was assessed by evaluating both the "Time Violation" and "Average Temperature Violation". Overall, all controllers maintained temperature violations below 1% of the secondary loop's operational time and kept the average deviation under 0.1 °C, allowing for a consistent and fair comparison of their energy-related performance. This is a key result, as it confirms that the energy savings achieved by DRL and RE, did not come at the expense of indoor air temperature control.

Both figures highlight the ability of the RE controller to approximate DRL performance across all key metrics, offering a significantly more practical implementation path. While DRL achieves the lowest total energy use (463 kWh), the RE controller follows closely (515 kWh), with a relatively small performance gap. More importantly, RE delivers a substantial reduction in energy consumption compared to the RBC baseline, demonstrating that the rule-extracted logic effectively encodes the core behavioral patterns of the source DRL agents.

Fig. 10 also highlights the ability of RE controller to reduce maximum power demand and improve the operational COP of the HP with respect to RBC baseline, closely matching the performance of the DRL controller.

To better visualize the on/off logic learned by the RE controller, Fig. 12 presents the operational state of the HP during the deployment episode of the reference scenario. The carpet plot shows simulation days on the y-axis and hours of the day on the x-axis, using a 15-minute resolution. Periods of HP activity are shown in green, while inactive periods are shown in grey, clearly illustrating the temporal patterns of system operation.

Fig. 12 reveals several key trends in the RE controller's behavior. On weekdays, a consistent pre-heating phase occurs between 4:30 and 5:00, before the secondary loop is activated. During daytime hours (6:00 to 18:00), HP operation strongly depends on outdoor temperature: as the climatic conditions become milder, afternoon activations gradually shorten by late February. At night and during weekends, the HP remains mostly inactive, with activations most likely due to the safeguard rule's activation to prevent the TES temperature from dropping below a critical limit.

A similar approach was used to visualize the control actions derived from the supply water temperature regression trees, with the water supply temperature reported in Fig. 13. The figure reports a carpet plot, where the color gradient from blue to red is associated with the supply water temperature setpoint selected by controllers. It is important to note that during periods when the HP is inactive, the plotted temperature corresponds to the tank temperature, which must be considered for accurate interpretation of the results.

Even in this case some key patterns can be inferred from the learned control strategy. During the pre-heating phase, the water supply temperature is increased to raise the TES temperature before the secondary loop starts. During daytime operation, the supply water temperature peaks in the early morning to cover the initial heating demand, then stays at a moderate level until around 14:00, before decreasing in the afternoon. At night and on weekends, the controller maintains the TES temperature and aims to keep a high value of the COP for the HP.

4.3. Performance analysis of the RE controller across the transfer scenarios

After being evaluated in the reference scenario (Turin-OB), the RE controller was also tested in the other 8 transfer scenarios described above. The goal was to assess whether it could still outperform the baseline RBC logic and achieve performance levels comparable to a DRL agent specifically trained for each case.

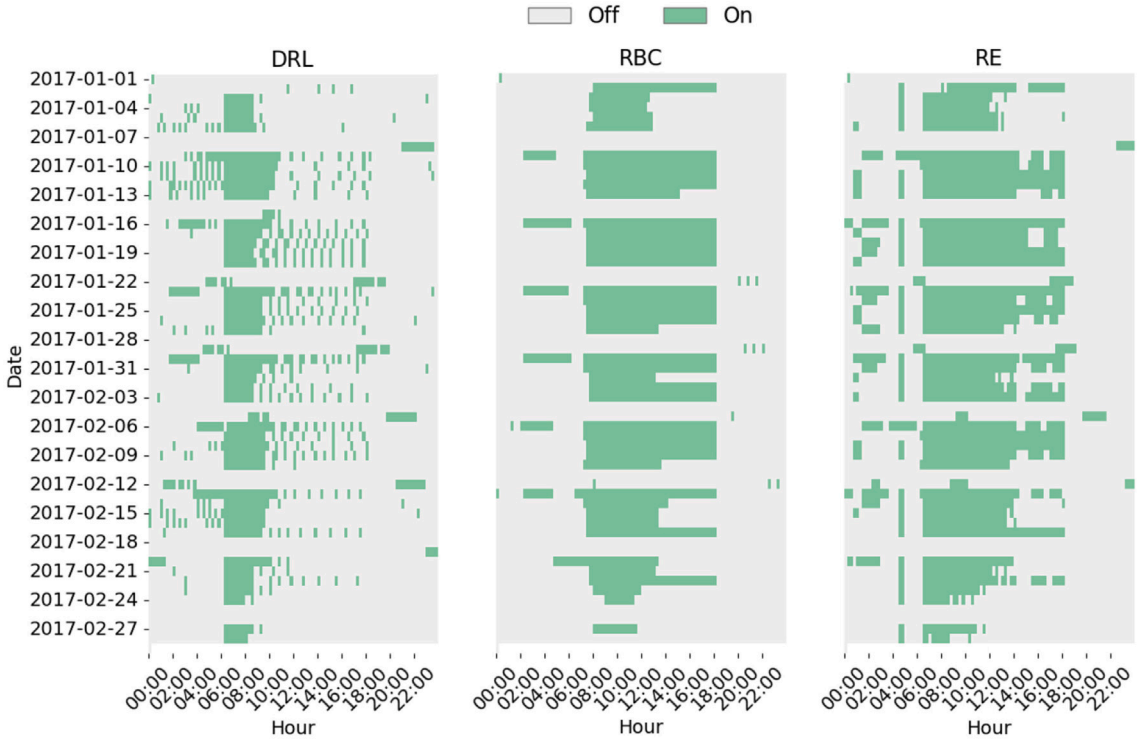


Fig. 12. Comparison of modHP actions for the RBC, DRL, and RE controllers over the testing episode in the reference scenario.

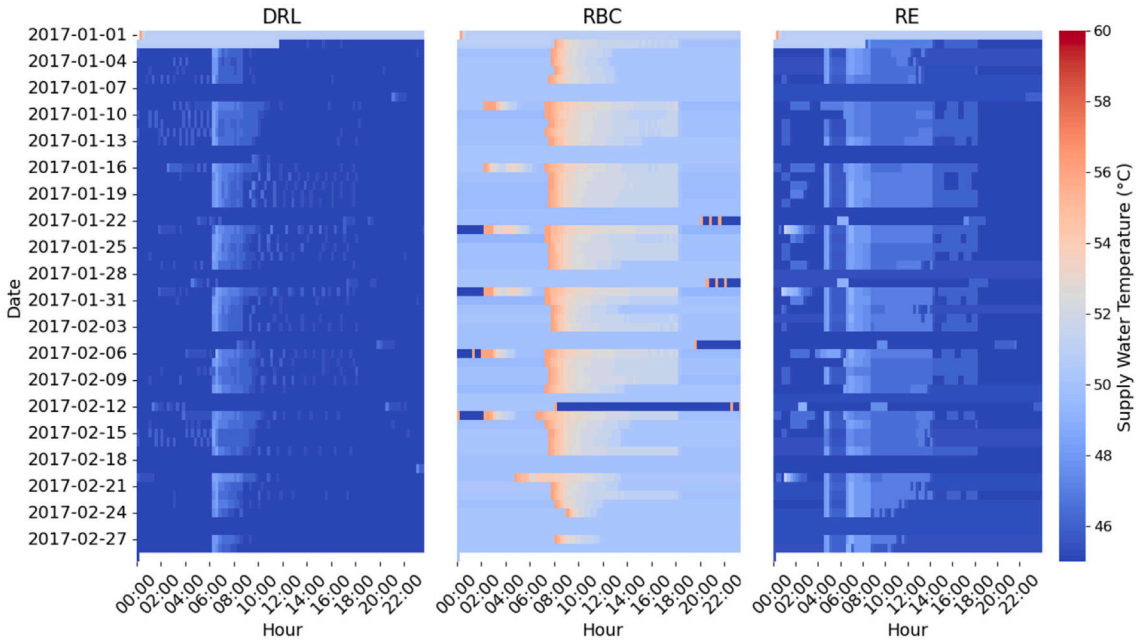


Fig. 13. Comparison of supply water temperature setpoints for the RBC, DRL, and RE controllers over the testing episode in the reference scenario.

Fig. 14 shows a comparative analysis of the three controllers across all scenarios using the defined KPIs. The DRL controller, tailored for each individual case, represents the upper bound of performance. In contrast, both RBC and RE rely on a fixed control logic. Since the RE controller was trained solely on data from the reference scenario, its results in the transfer scenarios provide a clear indication of its ability to generalize.

In all cases, the controllers achieved satisfactory indoor air temperature control. While the DRL controller exceeded the 10% violation time threshold in a few scenarios, the average temperature deviation remained very low, indicating that temperature violations were minor and short-lived. This ensured a consistent and fair basis for comparing the controllers in terms of observed energy-related performance. In terms of energy consumption, DRL performs best, with RE closely following in most cases. The RE controller notably reduces energy use compared to RBC, especially in colder scenarios where heating loads are higher. Even for the milder climate conditions (e.g., transfer scenarios of Naples), RE maintains a good trade-off, performing better than RBC and narrowing the gap with DRL.

The max power demand plots reinforce this pattern. RBC is always responsible for the highest electrical power demand. In contrast, both DRL and RE stay well below. However, in some scenarios of cold climates (e.g., Turin-O2, Bolzano-OB and Turin-O2), RE shows a noticeable increase in max power compared to DRL. For example, in Bolzano O2, RE peaks at 5.3 kW, which is about 3 kW higher than DRL. This suggests that the rule-based logic, while effective on average, may not effectively clip demand peaks under high load variability and different boundary conditions.

In contrast, both DRL and RE maintain higher and nearly identical COP values, demonstrating their ability to adapt to external climatic conditions and adjust the supply water temperature based on actual thermal demand.

Overall, the RE controller emerges as a reliable and scalable solution. While DRL remains the reference in terms of absolute performance, the marginal gains it offers over RE must be weighed against the increased complexity, particularly in real-world applications where explainability and control maintainability are critical.

Eventually Table 11 presents the percentage improvement for each scenario for the DRL and RE controllers relative to the RBC baseline, calculated as expressed in Eq. (4).

In terms of energy savings, across all the scenarios RE achieves values between 6.9% and 15.0%, while DRL ranges from 9.9% to 20.3%. Although DRL shows a higher saving potential, RE still delivers consistent and substantial reductions, particularly in colder climates (Turin and Bolzano). For max power reduction, RE displays the widest variability, ranging from 22.8% to 61.3%, while DRL spans from 27.0% to 66.2%. This indicates that DRL is generally more robust in reducing max power demand, whereas RE's effectiveness can be more scenario-dependent. Regarding COP increase, both controllers show significant and stable improvements over the baseline. RE ranges from 12.8% to 23.4%, while DRL ranges from 14.4% to 27.1%.

Having demonstrated the overall superiority of both advanced controllers over the RBC baseline, Table 12 provides a direct comparison between the RE and DRL controllers. It reports the percentage difference in performance, where positive values indicate better results for the RE controller, and negative values indicate a performance advantage for DRL. This comparison helps quantify the trade-offs between the two approaches, offering a clearer perspective on how effectively the transferred RE controller performed against DRL agents specifically trained for each scenario. Eventually, the activation frequency of the expert-designed safeguard rule was evaluated to assess its impact on the control strategy. For each scenario, the share of time during which the HP would operate under the safeguard logic was computed. In all tested cases, activations remained infrequent. Even in the most critical scenario, Bolzano - O2, the safeguard was active for less than 3% of total operating hours. Key performance indicators, including total electrical consumption, average heat pump COP, and indoor temperature violations, remained unaffected. Only peak power showed minor differences, as activations occurred mainly during night hours when the outdoor air temperature was low. Overall, the safeguard ensured safe operation without altering the core behavior of the extracted control logic.

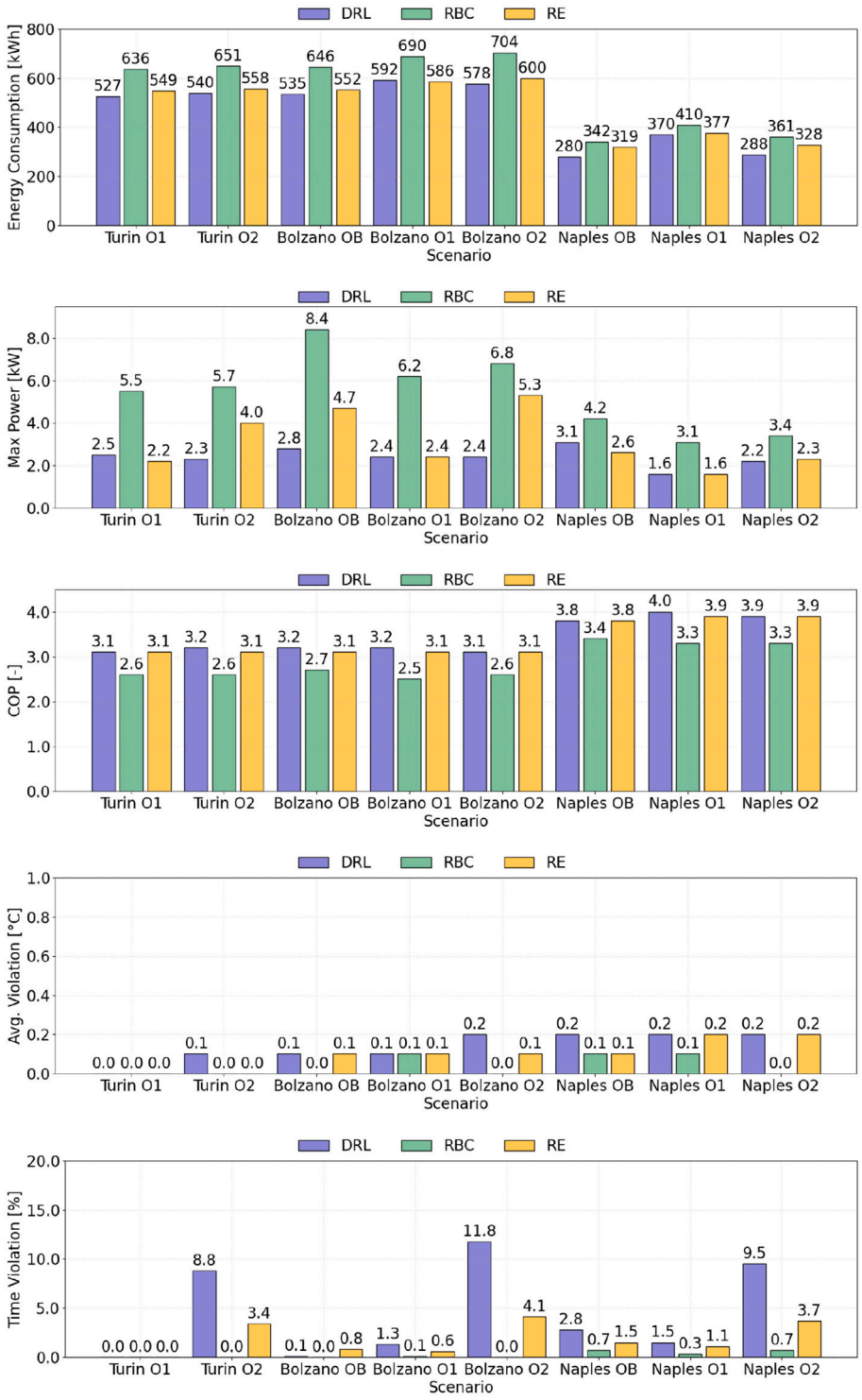


Fig. 14. Comparison between RBC, DRL and RE controller according to different KPIs in the transfer scenario.

5. Discussion

The results demonstrated that RE control strategies, when properly designed, could replicate the essential behaviors of advanced control policies while maintaining robust performance across varied operating conditions. The ensemble-based RE approach developed in this study offered a practical balance between accuracy, transparency, and transferability, supporting its potential for real-world deployment.

In all tested scenarios, the RE controller outperformed the baseline RBC. It reduced both energy consumption and maximum power demand while ensuring stable indoor temperature control. Improvements in COP closely aligned with those achieved by DRL agents, confirming that the RE logic successfully captured efficiency-oriented behaviors. Although DRL maintained superior performance in managing maximum power demand, the RE controller represented a more stable and interpretable solution with only limited performance degradation.

Table 11
Percentage variation of energy performance metrics for the DRL controllers and RE controller relative to the RBC baseline for each transfer scenario.

Scenario	Energy saving [%]		Max power reduction [%]		COP increase [%]	
	DRL	RE	DRL	RE	DRL	RE
Turin O1	17.1	13.7	54.0	60.0	18.7	21.0
Turin O2	17.1	14.3	60.5	29.7	22.3	18.8
Bolzano OB	17.2	14.6	66.2	44.1	18.4	15.4
Bolzano O1	14.1	15.0	60.9	61.3	27.1	23.4
Bolzano O2	17.8	14.8	65.2	22.8	20.1	19.0
Naples OB	18.2	6.9	27.0	37.1	14.4	12.8
Naples O1	9.9	8.2	49.6	48.1	20.3	16.4
Naples O2	20.3	9.2	36.9	31.8	17.5	16.2

Table 12
Differences between the percentage improvements of the DRL and RE controllers compared to the RBC baseline.

Scenario	Δ Energy saving [%]	Δ Max power reduction [%]	Δ COP increase [%]
Turin OB	−8.7	−5.7	+0.8
Turin O1	−3.5	+6.0	+2.2
Turin O2	−2.8	−30.8	−3.5
Bolzano OB	−2.6	−22.2	−3.1
Bolzano O1	+0.9	+0.5	−3.8
Bolzano O2	−3.0	−42.4	−1.1
Naples OB	−11.4	+10.1	−1.6
Naples O1	−1.7	−1.5	−4.0
Naples O2	−11.1	−5.1	−1.3

A key contribution of the study was the ensemble-based structure of the RE process. Rather than depending on a single DRL agent, control rules were extracted from a family of independently optimized agents. This approach enabled the final controller to incorporate a broader range of decision patterns, which helped prevent overfitting and improved robustness under diverse conditions. The use of a majority voting mechanism filtered out noisy actions and ensured a smoother and more consistent control output.

The architecture of the RE controller, which separates binary activation logic from temperature modulation, proved fundamental for generalization. The DTs prioritized variables linked to system schedules and system thermal dynamics. Features such as time until start system, time until end system and TES water temperature emerged as dominant, allowing the controller to anticipate demand and adjust its operation according to external and internal conditions.

Transferability assessments confirmed that the RE controller maintained acceptable comfort levels even in scenarios with occupancy patterns and climates that differed substantially from the training case. In warmer climates, where operational conditions diverged more significantly, the controller exhibited a moderate decline in performance. However, it continued to outperform the RBC baseline. In cases with more similar climate profiles, such as Bolzano, the RE controller closely approached the DRL benchmark, with an energy savings gap of approximately 3% compared to DRL when measured against the RBC.

Max power demand remained the metric where the RE controller differed most from DRL. This gap reflected the DRL agents' capacity to fine-tune control actions over short time intervals, a capability that static rule sets could only partially replicate. Nonetheless, the RE controller preserved the core operational principles of the original agents and proved resilient across a variety of boundary conditions.

Based on these findings, future work should extend the methodology beyond its current scope and address its main limitations. A primary step is to move beyond validation on a single system layout. The approach should be tested across a set of archetypal HVAC configurations to assess scalability and general applicability under structurally different system layouts. In addition, the framework should be evaluated on more complex energy systems. Relevant cases include hybrid HVAC configurations integrating thermal and electrical storage, combined heating and cooling technologies, and on site renewable generation. Such scenarios introduce additional operational constraints and strengthen the need for explicit multi-objective control formulations.

In parallel, the integration of real-time grid-side signals should be considered. Moving beyond fixed price profiles to incorporate flexibility requests from transmission or distribution system operators would enable the RE controllers to align local objectives with system-level constraints, allowing buildings to act as intelligent, grid-responsive assets.

A further development pathway involves coordinating RE controllers across multiple buildings. In such a setup, individual RE units could serve as modular elements within a decentralized or hierarchical architecture, where supervisory agents dispatch signals tailored to each site. This would allow building clusters to deliver a coordinated response to external inputs. Thanks to their compatibility with existing automation systems, minimal reliance on cloud infrastructure, and low deployment complexity, RE controllers appear well suited for large-scale adoption, particularly in retrofit contexts or digitally limited environments.

6. Conclusion

This work introduced an innovative methodology to derive an advanced, interpretable, and transferable RE controller from a family of DRL controllers. The Optuna framework was first used to identify a set of high performing DRL controllers for a reference scenario. DTs rules were extracted for each of these controllers and combined through a majority voting scheme to build a robust, consolidated control policy. A final RE process was then applied to this ensemble policy to obtain a single RE controller. This controller was deployed both on the reference case and on eight additional transfer scenarios with different occupancy profiles and climate conditions, without requiring any structural adaptation. Its performance was evaluated against both a baseline RBC and a tailored, scenario-specific DRL controller. As reported in Section 4 and discussed in Section 5, the RE controller delivered solid energy performance without compromising indoor air temperature control. Across all tested scenarios and in comparison with the RBC baseline, it consistently reduced energy consumption by at least 7%, lowered max power demand by 23%, and increased the HP COP during operation by 13%. The comparison with the DRL controller revealed a performance gap, with the RE controller capturing on average 74% of the energy savings and 86% of the max power reduction. This gap remains acceptable once the advantages of the proposed approach are considered. The if-else structure of the RE controller provides full interpretability and allows operators to understand the rationale behind each decision. This level of transparency strengthens confidence in the controller and makes it possible for experts to intervene directly to refine its behavior, as shown in this study through the modification and pruning applied to the regression tree responsible for adjusting the supply water temperature setpoint. Although its performance is slightly lower than that of a case-specific DRL agent, the RE controller would be easier to implement in real applications and its demonstrated transferability reduces the need to design a custom controller for each operational scenario. It also shows strong tolerance to variations in building usage, climate conditions, and TES sizes. The results confirm that the controller can adapt reliably to challenging and previously unseen boundary conditions while ensuring stable and effective system operation.

However it applies only to cases sharing the same system layout. Future work will therefore assess the process across a set of archetypal HVAC configurations to verify scalability under structurally different systems. Further developments will follow two directions. First, the approach will be extended to more complex multi-zone buildings and integrated energy systems including renewable generation. Second, it will be scaled from single building applications to clusters of buildings, enabling coordinated energy management strategies.

CRediT authorship contribution statement

Marco Savino Piscitelli: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Alessandro Aniello Mele:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Giuseppe Razzano:** Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alfonso Capozzoli:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Data availability

Data will be made available on request.

References

[1] International Energy Agiency (IEA). Developing a global energy efficiency workforce in the buildings sector; 2024. https://www.iea.org/reports/developing-a-global-energy-efficiency-workforce-in-the-buildings-sector.
[2] International Energy Agiency. Energy efficiency 2023; 2023. https://www.iea.org/reports/energy-efficiency-2023.
[3] Ramos JS, Moreno MP, Delgado MG, Álvarez Domínguez S, Cabeza LF. Potential of energy flexible buildings: evaluation of DSM strategies using building thermal mass. Energy Build 2019;203:109442. https://doi.org/10.1016/J.ENBUILD.2019.109442
[4] Østergaard Jensen S, Marszal-Pomianowska A, Lollini R, Pasut W, Knotzer A, Engelmann P, Stafford A, Reynders G. IEA EBC annex 67 energy flexible buildings. Energy Build 2017;155:25–34. https://doi.org/10.1016/J.ENBUILD.2017.08.044
[5] Brandi S, Piscitelli MS, Martellacci M, Capozzoli A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energy Build 2020;224:110225. https://doi.org/10.1016/J.ENBUILD.2020.110225
[6] Arteconi A, Hewitt NJ, Polonara F. Domestic demand-side management (DSM): role of heat pumps and thermal energy storage (TES) systems. Appl Therm Eng 2013;51:155–65. https://doi.org/10.1016/J.APPLTHERMALENG.2012.09.023
[7] Wu P, Wang Z, Li X, Xu Z, Yang Y, Yang Q. Energy-saving analysis of air source heat pump integrated with a water storage tank for heating applications. Build Environ 2020;180:107029. https://doi.org/10.1016/j.buildenv.2020.107029
[8] Müller D, Monti A, Stinner S, Schlösser T, Schütz T, Matthes P, Wolisz H, Molitor C, Harb H, Streblow R. Demand side management for city districts. Build Environ 2015;91:283–93. https://doi.org/10.1016/J.BUILDENV.2015.03.026. https://www.sciencedirect.com/science/article/pii/S0360132315001432?via.
[9] Xin X, Zhang Z, Zhou Y, Liu Y, Wang D, Nan S. A comprehensive review of predictive control strategies in heating, ventilation, and air-conditioning (HVAC): model-free VS model. J Build Eng 2024;94:110013. https://doi.org/10.1016/J.JOBE.2024.110013. https://www.sciencedirect.com/science/article/pii/S235271022401581X?ref=pdf_download&fr=RR-2&rr=951bfbe4da37ea56.
[10] Yu MG, Pavlak GS. Extracting interpretable building control rules from multi-objective model predictive control data sets. Energy 2022;240:122691. https://doi.org/10.1016/J.ENERGY.2021.122691
[11] Rohrer T, Frison L, Kaupenjohann L, Scharf K, Hergenröther E. Deep reinforcement learning for heat pump control. [ArXiv] arXiv:2212.12716. 2022. https://api.semanticscholar.org/CorpusID:255125400.
[12] Hou J, Li H, Nord N, Huang G. Model predictive control under weather forecast uncertainty for HVAC systems in university buildings. Energy Build 2022;257:111793. https://doi.org/10.1016/J.ENBUILD.2021.111793
[13] Wang W, Hu B, Wang RZ, Luo M, Zhang G, Xiang B. Model predictive control for the performance improvement of air source heat pump heating system via variable water temperature difference. Int J Refrig 2022;138:169–79. https://doi.org/10.1016/J.IJREFRIG.2022.03.001
[14] Zhang Z, Lam KP. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In: BuildSys 2018 - proceedings of the 5th conference on systems for built environments. Association for Computing Machinery; 2018. p. 148–57. https://doi.org/10.1145/3276774.3276775
[15] Wang Z, Hong T. Reinforcement learning for building controls: the opportunities and challenges. Appl Energy 2020;269:115036. https://doi.org/10.1016/J.APENERGY.2020.115036. https://www.sciencedirect.com/science/article/abs/pii/S0306261920305481?via.
[16] Lu X, Fu Y, O'Neill Z. Benchmarking high performance HVAC rule-based controls with advanced intelligent controllers: a case study in a multi-zone system in modelica. Energy Build 2023;284:112854. https://doi.org/10.1016/J.ENBUILD.2023.112854
[17] Buscemi G, Cuomo FP, Razzano G, Cappiello FL, Brandi S. Deep reinforcement learning-based control of thermal energy storage for university classrooms: co-simulation with trnsys-python and transfer learning across operational scenarios. Energy Rep 2025;14:1349–67. https://doi.org/10.1016/J.EGYR.2025.07.003. https://www.sciencedirect.com/science/article/pii/S2352484725004172.
[18] Francois-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J. An introduction to deep reinforcement learning. Found Trends Mach Learn 2018;11:219–354. https://doi.org/10.1561/2200000071. http://arxiv.org/abs/1811.12560http://dx.doi.org/10.1561/2200000071.
[19] Fu Y, Xu S, Zhu Q, O'Neill Z, Adetola V. How good are learning-based control v.s. Model-based control for load shifting? Investigations on a single zone building energy system. Energy 2023;273:127073. https://doi.org/10.1016/J.ENERGY.2023.127073
[20] Wang X, Kang X, An J, Chen H, Yan D. Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. Energy Build 2023;301:113696. https://doi.org/10.1016/J.ENBUILD.2023.113696. https://www.sciencedirect.com/science/article/pii/S037877882300926X?via.
[21] Quang TV, Phuong NL. Using deep learning to optimize HVAC systems in residential buildings. J Green Build 2024;19:29–50. https://doi.org/10.3992/jgb.19.1.29
[22] Dulac-Arnold G, Mankowitz D, Hester T. Challenges of real-world reinforcement learning. arXiv:1904.12901. 2019. https://arxiv.org/abs/1904.12901.
[23] Li X, Liu H, Li C, Chen G, Zhang C, Dong ZY. Deep reinforcement learning-based explainable pricing policy for virtual storage rental service. IEEE Trans Smart Grid 2023;14:4373–84. https://doi.org/10.1109/TSG.2023.3253140. https://ieeexplore.ieee.org/document/10068282.
[24] Coraci D, Brandi S, Hong T, Capozzoli A. An innovative heterogeneous transfer learning framework to enhance the scalability of deep reinforcement learning controllers in buildings with integrated energy systems. Build Simul 2024 17:5 2024;17:739–70. https://doi.org/10.1007/S12273-024-1109-6. https://link.springer.com/article/10.1007/s12273-024-1109-6.
[25] Pinto G, Wang Z, Roy A, Hong T, Capozzoli A. Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives. Adv Appl Energy 2022;5:100084. https://doi.org/10.1016/J.ADAPEN.2022.100084
[26] Coraci D, Brandi S, Hong T, Capozzoli A. Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. Appl Energy 2023;333:120598. https://doi.org/10.1016/J.APENERGY.2022.120598
[27] Lissa P, Schukat M, Keane M, Barrett E. Transfer learning applied to drl-based heat pump control to leverage microgrid energy efficiency. Smart Energy 2021;3:100044. https://doi.org/10.1016/J.SEGY.2021.100044
[28] Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency; 2022. p. 2239–50. https://api.semanticscholar.org/CorpusID:249872684.
[29] Razzano G, Brandi S, Piscitelli MS, Capozzoli A. 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. Appl Energy 2025;381:125046. https://doi.org/10.1016/j.apenergy.2024.125046. https://www.sciencedirect.com/science/article/pii/S0306261924024309.
[30] Dai Y, Chen Q, Zhang J, Wang X, Chen Y, Gao T, Xu P, Chen S, Liao S, Jiang H, zhong Gao DW. Enhanced oblique decision tree enabled policy extraction for deep reinforcement learning in power system emergency control. Electr Power Syst Res 2022;209. https://doi.org/10.1016/j.epsr.2022.107932
[31] Choi Y, Lu X, O'Neill Z, Feng F, Yang T. Optimization-informed rule extraction for HVAC system: a case study of dedicated outdoor air system control in a mixed-humid climate zone. Energy Build 2023;295:113295. https://doi.org/10.1016/j.enbuild.2023.113295
[32] Piscitelli MS, Brandi S, Gennaro G, Capozzoli A, Favoino F, Serra V. Advanced control strategies for the modulation of solar radiation in buildings: MPC-enhanced rule-based control. In: Proceedings of building simulation 2019: 16th conference of IBPSA, volume 16 of building simulation. Rome, Italy: IBPSA; 2019. p. 869–76. https://doi.org/10.26868/25222708.2019.210609. https://publications.ibpsa.org/conference/paper/?id=bs2019_210609.
[33] Bursill J, O'Brien L, Beausoleil-Morrison I. Multi-zone field study of rule extraction control to simplify implementation of predictive control to reduce building energy use. Energy Build 2020;222:110056. https://doi.org/10.1016/J.ENBUILD.2020.110056
[34] Cho S, Park CS. Rule reduction for control of a building cooling system using explainable AI. J Build Perform Simul 2022;15:832–47. https://doi.org/10.1080/19401493.2022.2103586

[35] Gunay B, Ouf M, Newsham G, O'Brien W. Building performance optimization for operational rule extraction. In: Proceedings of building simulation 2019: 16th conference of IBPSA, volume 16 of building simulation. Rome, Italy: IBPSA; 2019. p. 2819–26. https://doi.org/10.26868/25222708.2019.210271. https://publications.ibpsa.org/conference/paper/?id=bs2019_210271.

[36] May-Ostendorp PT, Henze GP, Rajagopalan B, Corbin CD. Extraction of supervisory building control rules from model predictive control of windows in a mixed mode building. J Build Perform Simul 2013;6:199–219. https://doi.org/10.1080/19401493.2012.665481;ISSUE:ISSUE:DOI. https://www.tandfonline.com/doi/abs/10.1080/19401493.2012.665481.

[37] Zhu J, Niu J, Zhan S, Tian Z, Chong A, Wang H, Zhou H. A learning-based model predictive control method for unlocking the potential of building energy flexibility. Energy Build 2025;330:115299. https://doi.org/10.1016/J.ENBUILD.2025.115299. https://linkinghub.elsevier.com/retrieve/pii/S0378778825000295.

[38] Zhu J, Tian Z, Niu J, Lu Y, Cheng B, Zhou H. Machine learning-enhanced lightweight rule-based control strategy for building energy demand response. Build Simul 2025:1–20. https://doi.org/10.1007/S12273-025-1275-1/METRICS. https://link.springer.com/article/10.1007/s12273-025-1275-1.

[39] Stato Italiano. D.p.r. 26 agosto 1993, n. 412; 1993. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.del.presidente.della.repubblica:1993-08-26;412.

[40] Ente Italiano di Normazione (UNI). Riscaldamento e raffrescamento degli edifici - dati climatici - parte 2: dati DI progetto, Technical Report UNI/TR 10349-2, UNI, 2016.

[41] Ente Italiano di Normazione (UNI). UNI/TS 11300-4:2016; 2016. https://store.uni.com/uni-ts-11300-4-2016.

[42] Fritzson P, Pop A, Aronsson P, Lundvall H, Nyström K, Saldamli L, et al. The openmodelica modeling, simulation, and development environment; 2005. https://www.researchgate.net/publication/252264811[accessed: 14 July 2025].

[43] Blockwitz T, Otter M, Akesson J, Arnold M, Clauss C, Elmqvist H, et al. Functional mockup interface 2.0: the standard for tool independent exchange of simulation models. In: Proceedings of the 9th international modelica conference, vol. 76. Linköping University Electronic Press; 2012. p. 173–84. https://doi.org/10.3384/ecp12076173. http://dx.doi.org/10.3384/ecp12076173.

[44] Andersson C, Åkesson J, Führer C. Pyfmi: a Python package for simulation of coupled dynamic models with the functional mock-up interface; 2016. https://api.semanticscholar.org/CorpusID:218002023.

[45] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W. Openai gym. arXiv:1606.01540. 2016. https://arxiv.org/abs/1606.01540.

[46] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. arXiv:1907.10902. 2019. https://arxiv.org/abs/1907.10902.

[47] Fu Q, Han Z, Chen J, Lu Y, Wu H, Wang Y. Applications of reinforcement learning for building energy efficiency control: a review. J Build Eng 2022;50:104165. https://doi.org/10.1016/J.JOBE.2022.104165

[48] Chen H, Guo W, Bao W, Cui M, Wang X, Zhao Q. A novel interpretable decision rule extracting method for deep reinforcement learning-based energy management in building complexes. Energy Build 2025;334:115514. https://doi.org/10.1016/J.ENBUILD.2025.115514. https://www.sciencedirect.com/science/article/pii/S0378778825002440?via.

[49] Silvestri A, Coraci D, Brandi S, Capozzoli A, Borkowski E, Köhler J, Wu D, Zeilinger MN, Schlueter A. Real building implementation of a deep reinforcement learning controller to enhance energy efficiency and indoor temperature control. Appl Energy 2024;368:123447. https://doi.org/10.1016/J.APENERGY.2024.123447

[50] Coraci D, Brandi S, Piscitelli MS, Capozzoli A. Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings. Energies 2021;14. https://doi.org/10.3390/en14040997. https://www.mdpi.com/1996-1073/14/4/997.

[51] Coraci D, Silvestri A, Razzano G, Fop D, Brandi S, Borkowski E, Hong T, Schlueter A, Capozzoli A. A scalable approach for real-world implementation of deep reinforcement learning controllers in buildings based on online transfer learning: the hilo case study. Energy Build 2025;329:115254. https://doi.org/10.1016/J.ENBUILD.2024.115254. https://www.sciencedirect.com/science/article/pii/S0378778824013707?via.

[52] Song Y-Y, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015;27:130–5. https://doi.org/10.11919/j.issn.1002-0829.215044

[53] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 2006;15:651–74. https://doi.org/10.1198/106186006X133933

[54] Hothorn T, Zeileis A. Partykit: a modular toolkit for recursive partytioning in r. J Mach Learn Res 2015;16:3905–9. https://jmlr.org/papers/v16/hothorn15a.html.

[55] Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerg Themes Epidemiol 2017;14. https://doi.org/10.1186/s12982-017-0064-4