Build Simul (2024) 17: 739–770
https://doi.org/10.1007/s12273-024-1109-6

An innovative heterogeneous transfer learning framework to enhance the scalability of deep reinforcement learning controllers in buildings with integrated energy systems

Davide Coraci[1], Silvio Brandi[1], Tianzhen Hong[2], Alfonso Capozzoli[1] (✉)

1. Politecnico di Torino, Department of Energy, TEBE research group, BAEDA Lab., Corso Duca degli Abruzzi 24, Torino, 10129, Italy

2. Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

Abstract

Deep Reinforcement Learning (DRL)-based control shows enhanced performance in the management of integrated energy systems when compared with Rule-Based Controllers (RBCs), but it still lacks scalability and generalisation due to the necessity of using tailored models for the training process. Transfer Learning (TL) is a potential solution to address this limitation. However, existing TL applications in building control have been mostly tested among buildings with similar features, not addressing the need to scale up advanced control in real-world scenarios with diverse energy systems. This paper assesses the performance of an online heterogeneous TL strategy, comparing it with RBC and offline and online DRL controllers in a simulation setup using EnergyPlus and Python. The study tests the transfer in both transductive and inductive settings of a DRL policy designed to manage a chiller coupled with a Thermal Energy Storage (TES). The control policy is pre-trained on a source building and transferred to various target buildings characterised by an integrated energy system including photovoltaic and battery energy storage systems, different building envelope features, occupancy schedule and boundary conditions (e.g., weather and price signal). The TL approach incorporates model slicing, imitation learning and fine-tuning to handle diverse state spaces and reward functions between source and target buildings. Results show that the proposed methodology leads to a reduction of 10% in electricity cost and between 10% and 40% in the mean value of the daily average temperature violation rate compared to RBC and online DRL controllers. Moreover, online TL maximises self-sufficiency and self-consumption by 9% and 11% with respect to RBC. Conversely, online TL achieves worse performance compared to offline DRL in either transductive or inductive settings. However, offline Deep Reinforcement Learning (DRL) agents should be trained at least for 15 episodes to reach the same level of performance as the online TL. Therefore, the proposed online TL methodology is effective, completely model-free and it can be directly implemented in real buildings with satisfying performance.

Keywords

transfer learning
reinforcement learning
building control
building energy management

Article History

Received: 21 November 2023
Revised: 09 January 2024
Accepted: 18 January 2024

1 Introduction

With the introduction of the European Green Deal (European Commission 2019), the European Commission has established the penetration of Renewable Energy Sources (RES) in buildings as an ambitious countermeasure to achieve net-zero carbon emissions by 2050 (Pinto et al. 2022a), considering that buildings have a pivotal role in the ongoing energy transition process accounting for 40% of the primary energy consumed worldwide (IEA 2019; Pinto et al. 2021). In this framework, incentive programs support the integration of Photovoltaic (PV) and Battery Energy Storage System (BESS) in building energy systems, which traditionally consist of Heating, Ventilation and Air Conditioning (HVAC) systems (Coraci et al. 2023a). HVAC systems are the most energy-intensive source in building operations, therefore substantial efforts contributed in recent years to enhance their energy efficiency through improved energy management

E-mail: alfonso.capozzoli@polito.it

List of symbols

α	Boltzmann temperature coefficient
β	temperature term weight of reward function
γ	discount factor
δ	electricity cost term weight of reward function
η_{rte}	round-trip efficiency of battery
θ	peak term weight of reward function
μ	learning rate
χ_i	internal heat capacity [kJ/(m²·K)]
C_{BESS}	nominal capacity of battery [kWh]
$c_{el,sell}$	electricity price for selling [€/kWh]
c_{el}	electricity buying price [€/kWh]
D_S	source domain
D_T	target domain
$E_{BESS,b}$	total energy supplied to the building from BESS [kWh]
$E_{ch,max}$	maximum battery charging energy [kWh]
$E_{CHILLER}$	chiller energy consumption [kWh]
$E_{cost,source}$	electricity cost for source building [€]
E_{cost}	electricity cost for target buildings [€]
$E_{dis,max}$	maximum battery discharging energy [kWh]
E_{LOAD}	non HVAC loads electrical consumption [kWh]
E_{PUMP}	circulation pumps energy consumption [kWh]
$E_{PV,b}$	total energy supplied to the building from PV [kWh]
$E_{PV,tot}$	total energy produced by PV [kWh]
E_{PV}	energy production from PV [kWh]
$E_{TOT,b}$	total building electricity consumption [kWh]
$f(\cdot)$	objective predictive function
g	solar heat gain coefficient
I_E	income from selling the excess of energy produced by PV to the grid [€]
$P_{BESS,ch,max}$	maximum battery charging power [kW]
$P_{BESS,ch}$	battery charging power [kW]
$P_{BESS,dis,max}$	maximum battery discharging power [kW]
$P_{BESS,dis}$	battery discharging power [kW]
Q_{cap}	capacity of chiller [kW]
R_E	electricity cost term of reward function
R_P	peak term of reward function
R_T	temperature term of reward function
r_t	reward at control time step t
RBC_{CF}	rule-based controller part choosing whether to supply cooling energy to the building
RBC_{el}	rule-based controller managing the operation of the electrical part of energy system
RBC_{OM}	rule-based controller part choosing the operation mode of the energy system
RBC_{th}	rule-based controller managing the operation of the thermal part of energy system
SOC_{BESS}	state-of-charge of the BESS
SOC_{TES}	state-of-charge of the water storage
SP_{INT}	indoor air temperature setpoint [°C]
$\Delta T_{viol,daily}$	mean value of the daily average temperature violation rate
T_{ch}	chiller supply temperature [°C]
T_{INT}	indoor air temperature [°C]
T_{LOW}	lower threshold limit of temperature comfort range [°C]
$T_{s,max}$	storage temperature upper boundary [°C]
$T_{s,min}$	storage temperature lower boundary [°C]
T_S	source task
T_T	target task
T_{UPP}	upper threshold limit of temperature comfort range [°C]
T_{viol}	temperature violation [°C]
U_{OP}	thermal transmittance of the opaque envelope [W/(m²·K)]
U_{TR}	thermal transmittance of the transparent envelope [W/(m²·K)]

Abbreviations

AC	alternate current
AHUs	air handling units
AI	artificial intelligence
BESS	battery energy storage system
BCVTB	building control virtual test bed
CF	cooling fraction
COP	coefficient of performance
DC	direct current
DNNs	deep neural networks
DoC	depth of charge
DoD	depth of discharge
DR	demand response
DRL	deep reinforcement learning
DTW	dynamic time warping
FDD	fault detection and diagnosis
HVAC	heating, ventilation and air conditioning
IES	integrated energy systems
IL	imitation learning
IRL	inverse reinforcement learning
KPIs	key performance indicators
LfD	learning from demonstration
MILP	mixed-integer linear programming
ML	machine learning
MPC	model predictive control
OM	operation mode
OTL	online transfer learning
PID	proportional-integrative-derivative
PV	photovoltaic
RBC	rule-based controller
RES	renewable energy sources
RL	reinforcement learning
SAC	soft-actor-critic
SC	self-consumption
SOC	state-of-charge
SS	self-sufficiency
TES	thermal energy storage
TL	transfer learning
TOU	time-of-use
VAV	variable air volume

(Coraci et al. 2021; Piscitelli et al. 2021). Moreover, buildings have the potential to leverage flexibility sources on both the thermal side (e.g., Thermal Energy Storage (TES) and building thermal inertia) and the electrical side (e.g., PV and BESS) to shift or reduce their energy demand.

Currently, HVAC systems in buildings are managed by easy to implement control strategies, as Rule-Based Controller (RBC) or Proportional-Integrative-Derivative (PID), developed by experts and reported in ASHRAE Guideline 36-2021 (ASHRAE 2021). However, RBC and PID controllers are unable to address conflicting objectives (Salsbury 2005) and to automatically adapt their control policy to dynamic boundary conditions such as weather conditions, electricity prices, and grid requirements (Finck et al. 2018). In this context, Reinforcement Learning (RL) emerged as a promising solution to overcome the limitations of traditional control strategies (Nagy et al. 2023). RL is a model-free control approach that involves a control agent learning the optimal control policy by means of interaction with the controlled environment through a reward mechanism (Sutton and Barto 2018). RL-based control strategies avoid the complex modelling efforts associated with model-based control methods, as Model Predictive Control (MPC) (Wei and Calautit 2023). The mainly implemented algorithm belonging to the RL family for control purposes is DRL, which combines RL with Deep Neural Networks (DNNs) as function approximators for control policies. The implementation of DRL has been successfully employed to address complex control problems as in real buildings, achieving performance levels approaching human capabilities (Mnih et al. 2015).

In the framework of building energy management, DRL results as an effective control strategy when implemented to manage HVAC system for selecting the operation mode of thermal energy systems (Wang et al. 2023b), supply water temperature (Zhang et al. 2019) and supply water flow at generation level (Wang et al. 2023a), indoor air temperature setpoint (Elehwany et al. 2024), water pump speed (Xiong et al. 2023) and fan speed (Fulpagare et al. 2022). Moreover, several contributions in literature evaluated the implementation of DRL to manage the operation of thermal storage (by managing the charge/discharge process (Deltetto et al. 2021) or storage temperature (Vázquez-Canteli et al. 2019)), PV systems coupled with BESS (Anvari-Moghaddam et al. 2017) and hybrid photovoltaic-thermal panels (Yang et al. 2015).

The training approaches for enabling the RL-based agent to interact with the controlled environment and retrieve the optimal control policy have two primary forms: online DRL and offline DRL. The offline DRL strategy is widely explored in literature since it guarantees a stable control policy and near-optimal performances in buildings through an offline pre-training carried out on surrogate models emulating the building dynamics. These models could be more detailed as physics-based engineering models developed in Modelica (Modelica Association 2000) and EnergyPlus (Crawley et al. 2001), or data-driven models based on monitored building data (Zou et al. 2020).

Despite the excellent performance demonstrated by the pre-trained offline DRL agents, a critical issue arises concerning the limited scalability and generalisability of this process. Given the uniqueness of each building, the definition of a tailored surrogate model is required. Furthermore, the development of a data-driven surrogate model requires a minimum amount of monitoring data for the controlled building, while the definition of physics-based models could be a time-consuming task, demanding access to detailed building information (which may not always be accessible) and domain expertise.

To address these practical challenges, Transfer Learning (TL) emerges as a viable solution to enhance the scalability and to enable the implementation of RL-based controllers in real buildings. Transfer learning is a Machine Learning (ML) technique where a model initially trained to solve a specific task (i.e., source task) in a particular domain, is transferred to address a new task (i.e., target task) that shares similarities with the original task, either within the same domain or across different domains (Pan and Yang 2010). The development of a transfer learning strategy in the framework of energy management in buildings can offer multiple advantages, such as enabling the direct implementation of DRL-based in real buildings with acceptable performance from the early stage of deployment and enhancing the scalability of such controllers in buildings avoiding the development of surrogate models.

The following subsection offers an overview of relevant studies concerning the integration of TL in the context of building control applications, while theoretical principles of TL are extensively discussed in Appendix A1.

1.1 Related works on TL applications for reinforcement learning control agents in buildings

In the framework of applications for smart buildings, transfer learning ensures excellent performance in the knowledge sharing of ML-based model employed for emulating building dynamics (Grubinger et al. 2017; Pinto et al. 2022b), load forecasting (Li et al. 2021; Li et al. 2022), occupancy detection (Mosaico et al. 2019), activity recognition (Chiang et al. 2017) and Fault Detection and Diagnosis (FDD) (Li et al. 2023). In the control field, TL is applied mainly in the automotive (Wang et al. 2019) and robotics domains (Zelinka et al. 2022), particularly for controllers based on RL.

Implementing transfer learning strategies for smart building control is a topic of recent interest since the earliest TL applications in this context are relatively recent compared to others within the context of machine learning. Fang et al. (2023) developed a TL methodology to investigate the cross-temporal and spatial transferability of a DRL controller within an HVAC system, consisting of a chiller and three Air Handling Units (AHUs), to improve indoor temperature conditions while reducing energy consumption. This study demonstrated that effective transfer occurs when the DRL agent is moved between buildings situated in similar climatic conditions, mainly when two out of the five layers of the neural network approximating the control policy in the source building were shared between source and target control agents. Zhang et al. (2022a) introduced a strategy aimed at transferring a multi-agent RL from a source building with multiple zones to a target building. In this work, the authors developed a methodology to choose the most suitable pre-trained RL policy obtained from the source building prior to the transfer learning process, managing zone temperature setpoints in a Variable Air Volume (VAV) system. This approach was effective since the HVAC system in the target building saved 40% of energy consumption compared to the baseline controller and 50% when compared to an RL controller trained from scratch over 5000 episodes. Esrafilian-Najafabadi and Haghighat (2023) developed an occupancy-based TL methodology exploiting the K-means algorithm and Dynamic Time Warping (DTW) to match similar occupancy patterns in 26 residential units. This approach enhanced the control performance for the HVAC system compared to a state-of-the-art model-free controller, minimising thermal discomfort during the learning process and increasing by 25% and 5% the jumpstart and asymptotic performances. Lissa et al. (2021) introduced parallel transfer learning, an intra-transfer learning technique enabling the knowledge transfer among five distinct agents throughout their training without the need to wait until the end of the training process. This approach has been applied within a microgrid comprising five individual homes, leading to a five-fold reduction in training time and a 10% decrease in energy consumption compared to scenarios with no knowledge transfer. Each home is equipped with its energy system, consisting of a PV system and a heat pump, managed by a DRL controller that optimises heat pump operation to minimise energy costs.

The applications discussed so far have been evaluated on target buildings after transferring the controller and training it across multiple episodes (i.e., temporal segments representative of the particular control problem). However, running multiple episodes corresponds to several months or years in reality before achieving acceptable performance for DRL-based controllers. Furthermore, the training or fine-tuning process across multiple episodes would require the definition of a surrogate building model, a task requiring a certain level of expertise.

In light of these challenges, recent advancements in the literature have introduced TL-based methodologies to achieve robust DRL controller performance right from the initial stages of implementation in target buildings. Nweye et al. (2023) developed a TL strategy implementing Imitation Learning (IL), by observing for 5 months an expert RBC) and weight-initialisation for training, evaluating and deploying DRL-based controllers within an energy community comprising 17 family houses, which operation is simulated using the CityLearn environment (Vázquez-Canteli et al. 2020). Each building was equipped with high-efficiency appliances, electric heating, a water heating system and PV panels, while six of these buildings featured 6.4 kWh BESS managed by independent DRL controllers to minimise electricity costs and carbon emissions from the grid electricity supply. The results suggested that transferring the DRL control policy from one building to another within the energy community yielded comparable performance while reducing the training costs. Coraci et al. (2023b) developed an online transfer learning approach that exploits two knowledge-sharing techniques, weight-initialisation and IL, to transfer a DRL controller pre-trained on a source office building that minimises electricity cost while enhancing indoor temperature conditions by managing a cooling system. The proposed online transfer learning approach aims to replicate real-world implementation by simulating the transferred DRL agent in the target buildings for a single episode. Source and target buildings have the same energy system and geometry but differ in several other aspects such as weather conditions, electricity price, occupancy schedules and building envelope characteristics. The results show that the online transfer learning strategy performed significantly better than the RBC and online DRL strategies, enhancing indoor temperature conditions of 50% and 80%, respectively, while improving energy system operation and achieving electricity cost savings ranging from 20% to 40%. In their works, Dey et al. (2023b) introduced respectively IL and Inverse Reinforcement Learning (IRL) strategies for a DRL agent managing indoor temperature setpoints within a 5-zone office building model created using Spawn (Wetter et al. 2023). The building was equipped with five AHUs and operated within the framework of a Demand Response (DR) program, to reduce energy consumption and minimise thermal discomfort. In Dey et al. (2023b), an IL strategy is designed to leverage knowledge extracted from a RBC to expedite DRL training and mitigate the unstable behaviour often encountered by DRL agents during initial exploration phases. The results demonstrated the effectiveness of the imitation learning approach, resulting in a 6% reduction

in average cost and an average score improvement of 7% during the testing period when compared to a rule-based heuristic policy. In Dey et al. (2023a), the authors develop an IRL framework consisting of two steps: inverse learning and forward learning. Results showed that when compared with three DRL-based training strategies (i.e., direct training, offline training and meta-model training), the IRL approach achieved a performance level better or comparable to those of RBC before the DRL agent begins its learning through direct interaction with the building while mitigating the erratic exploratory behaviour typically exhibited by an untrained DRL agent during the initial training phase when directly implemented in a building environment.

1.2 Novelty and contributions of the paper

From the literature review, it emerges that the development and implementation of TL-based methodologies have increased in recent years. To the best of the authors' knowledge, these applications refer to transfer procedures of DRL-based control agents in homogeneous and transductive settings. In detail, the homogeneous TL occurs between controllers having the same action and state-spaces, since they manage the operation in different buildings (i.e., domain) having similar geometric properties and the same energy system, while transductive TL indicates that source and target controllers optimise the same objective function.

However, buildings are not unique entities, since they are often equipped with distinct energy systems controlled in different ways to meet specific objectives. Consequently, there is a growing demand to develop a systematic procedure that facilitates the effective transfer of pre-trained DRL controllers in buildings where the energy systems and objectives differ from those in the source building. Furthermore, the transfer learning approach to be developed should enable a transferred agent to achieve acceptable performance from the early stage of deployment in the target building, enhancing the scalability of DRL controllers in a real-world context.

In this scenario, a possible solution could involve the development of an Online Transfer Learning (OTL) strategy. According to this strategy, the weights of the neural networks are initialised to values obtained for the source building and then further fine-tuned on the target building while the DRL agent actively controls the system. The weight-initialisation helps retain the acquired knowledge from the source building in cases where certain boundary conditions match those of the target building. Subsequently, fine-tuning the control policy would be necessary to adapt to the new conditions (e.g., energy systems) present in the target buildings. However, this procedure is not always feasible as the presence of different energy systems leads to different state and/or action spaces among DRL controllers implemented in different buildings, thereby limiting the application of the weight-initialisation technique.

To address these limitations, this paper explores the implementation of online heterogeneous transductive and inductive transfer learning strategies for DRL controllers having different state-spaces (i.e., heterogeneous TL), since they operate in buildings having different energy systems to optimise different objective functions (i.e., inductive TL).

The developed online TL strategy includes the model slicing technique and imitation learning in the knowledge-sharing process. The model slicing technique, commonly used in training and deploying neural networks (Zhang et al. 2022b), enables the weightinitialisation for the target controller by employing the pre-trained weights from the source controller, even when the state-space of source and target controllers differs. Imitation learning is employed to initialise the memory buffer of the target controller with transitions obtained from a warm-up phase conducted with a RBC.

The proposed approach is designed to transfer a pre-trained DRL control policy based on the discrete formulation of Soft-Actor-Critic (SAC) algorithm, developed by Christodoulou (2019) to allow the selection of discrete actions. Theoretical aspects regarding DRL and the discrete SAC controllers can be found in the literature (Bellman 1966; Sutton and Barto 2018; Christodoulou 2019; Haarnoja et al. 2019). EnergyPlus and Python were coupled in a simulation environment to test the effectiveness of the developed online transfer learning strategy. The SAC controller is pre-trained on a source building to minimise electricity cost while enhancing indoor temperature conditions by effectively managing the operation of a cooling system consisting of a chiller and a TES. In detail, the SAC controller chooses the optimal operation mode of the cooling system and the fraction of nominal cooling energy to be delivered to the building. The pre-training phase of the DRL agent in the source building also involved an automated optimisation procedure of DRL controller hyperparameters carried out by means of the Python library Optuna (Akiba et al. 2019), considering that the values of hyperparameters significantly influence the performance of DRL agents.

Then, the DRL agent pre-trained on the source building was transferred in heterogeneous setting to multiple target buildings that shared similar geometric properties but with a more integrated energy system compared to the source building, since PV panels and BESS are installed in addition to the thermal components (i.e., chiller and TES). PV and BESS are managed in all target buildings by a RBC strategy. In both the source and target buildings, DRL controllers share the same action space, enabling DRL agents to take similar actions in both domains. However, the state space

differs between these buildings. This difference arises from the integration of PV and BESS into the energy system of target buildings, resulting in a more complex and expanded state space compared to the source building. Including information about PV and BESS operation within the state space of DRL controllers enables the DRL agent to make optimal decisions regarding the sequence of control actions for operating the cooling system. Three control strategies based respectively on offline DRL, online DRL and RBC were defined to benchmark the performance of the online TL strategy in terms of electricity cost and temperature violations in heterogeneous transductive setting. Moreover, a performance comparison for online TL is carried out in terms of peak violations compared with the offline DRL strategy when the source DRL controller is transferred in heterogeneous inductive setting.

Based on existing research on the application of TL for DRL controllers, this paper includes a set of novel contributions that could be summarised as follows:

- An online transfer learning strategy was developed to transfer a pre-trained DRL controller from a source office building for minimising electricity cost and improving indoor temperature conditions. To the best of the authors' knowledge, the implementation of transfer learning in heterogeneous settings for building control systems has not been explored in the existing literature. The effectiveness of the online transfer learning methodology was tested in heterogeneous transductive settings by including differences, in addition to the different energy systems, in climatic conditions, electricity price schedules, occupancy schedules, and building thermophysical properties.
- The performance of the developed online transfer learning methodology was assessed in the heterogeneous inductive setting, evaluating a target building with the same boundary conditions as the source but including in the objective function a term related to peak shaving in addition to the electricity cost minimisation and the improvement of indoor temperature conditions for occupants. To the best of our knowledge, TL applications for DRL controllers operating in different building environments to address different objective functions have not been explored in literature.
- The current state-of-the-art applications for TL in the context of building control have primarily been limited to homogeneous settings, where both source and target agents share an identical number of states and actions through a strategy combining weight-initialisation and fine-tuning. Nevertheless, this conventional scenario does not align with the specific case study analysed in this work, given that the state-space of the source controller includes fewer states than that of the target controllers. In this framework, the model slicing technique is included within the developed knowledgesharing strategy including imitation learning, weight-initialisation and fine-tuning to effectively transfer a DRL agent to several target controllers with different state-space.

The remaining paper is structured as follows: Section 2 outlines the methodological framework and provides details about the formulation of the control problem. Implementation details about source and target buildings, as well as information on online transfer learning and the controllers are elaborated in Section 3. Section 4 and 5 respectively provide an overview of the results obtained and discuss them. To conclude, Section 6 offers concluding remarks and outlines future directions.

2 Methodology

This section describes the methodology used in this paper to evaluate the implementation of transductive and inductive transfer learning techniques in heterogeneous settings. The framework is organised into three main parts, as shown in Figure 1:

1) Design of RBC and DRL controllers and training of the source DRL controller.
2) Transfer of DRL control policy in heterogeneous inductive and transductive settings.
3) Performance benchmarking of online TL strategy on target buildings.

The developed methodological framework enables the sharing of a control policy between source and target buildings that have the same geometric design and end-use (i.e., office buildings). However, they differ in terms of the configuration of the energy system.

The experiments are carried out during the cooling season, as a consequence the heating system is not considered. The cooling system consists of an air-to-water chiller, operating at a constant cold water temperature setpoint (denoted as T_{ch}), and a TES, acting as a buffer between chiller and building and operating at constant water flow rate. The TES system operates within a temperature range defined between a minimum $T_{s,min}$ and a maximum temperature $T_{s,max}$. These temperatures correspond respectively to the maximum and minimum State-Of-Charge (SOC) of the TES system. This study evaluates the effectiveness of thermostatic control within the building, whereby the cooling system modulates the supply of cooling energy according to building dynamics to ensure indoor temperature control for the occupants. In detail, the cooling system can be operated by a high-level controller that:

- optimal selects the cooling system operation mode, determining the most appropriate considering relevant influencing factors (e.g., outdoor conditions, occupancy patterns, electricity price schedules);

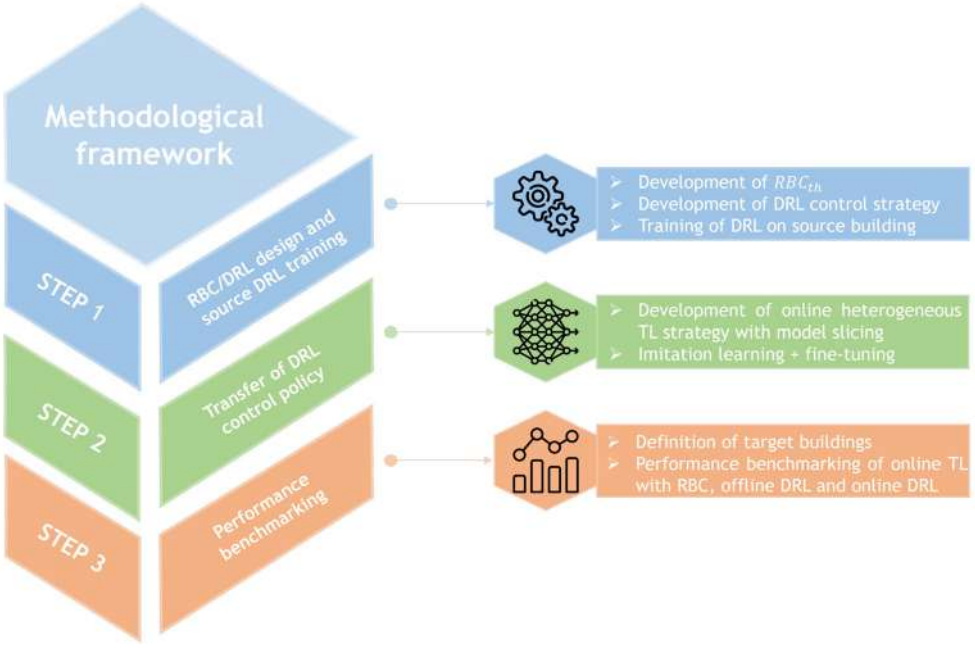

Fig. 1 Methodological framework developed in this work

– chooses the fraction of nominal cooling energy (i.e., cooling fraction) to be delivered to the building based on the desired indoor temperature and the measured temperature. By dynamically adjusting the cooling fraction, the controller ensures that the indoor temperature remains within the defined acceptability range.

Regarding the operation mode of the cooling system, the controller can choose between three different operation modes:

– **Discharging mode** (operation mode = −1): the cooling energy needed by the building is provided by discharging the TES. In this mode, the cooling system operates by supplying variable water temperature (i.e., constant water mass flow rate) to meet the cooling demands.
– **Chiller mode** (operation mode = 0): the cooling energy required by the building is provided by the chiller. In this setting, the cooling system operates at constant supply water temperature.
– **Charging mode** (operation mode = 1): the cooling system operates at constant supply water temperature to simultaneously supply cooling energy to both the thermal storage and the building (if required).

The cooling system configuration is the same for source and target buildings. The configuration of the electrical system differs from source to target buildings. The total building electrical load is determined by the electrical demand of the chiller, circulation pump, appliances and lighting services. The electrical system of the building consists of a Direct Current (DC) bus and an Alternate Current (AC) bus, which are interconnected through a unidirectional AC/DC inverter. The PV system and the BESS are installed both on the DC bus and they are connected to the AC bus via a DC/DC converter. Modelling features for the PV system and BESS are provided in Section 3. According to Italian regulations, the BESS charge from grid is not allowed. However, the grid operates to assist in balancing the electricity demand of the building with the renewable power generation from the PV system. This ensures that the energy needs of the building are met efficiently employing a combination of grid power and RES, while complying with the specified requirements. The BESS system is always managed through a RBC controller identified as RBC_{el} and detailed in Section 3. A simplified scheme of the energy system investigated for target buildings is presented in Figure 2. This case study is designed to assess the performance of online TL for a DRL-based controller.

The objective of the high-level controller managing the cooling system is to minimise the electricity cost associated with the operation of the chiller and circulation system while ensuring that the indoor temperature remains within an acceptable range. The temperature range is defined with a deviation of [−1, +1] from the desired indoor temperature setpoint of 26 °C (i.e., [25, 27] °C). To this purpose, in target buildings the controller can leverage renewable energy directly produced by the PV or stored in the BESS. Despite the controller is not capable of directly managing the BESS, its actions can influence the utilisation patterns of energy produced by the PV since the chiller is the most demanding electrical equipment of the considered system. Furthermore, in the framework of heterogeneous inductive transfer learning it was tested the capability of the controller to

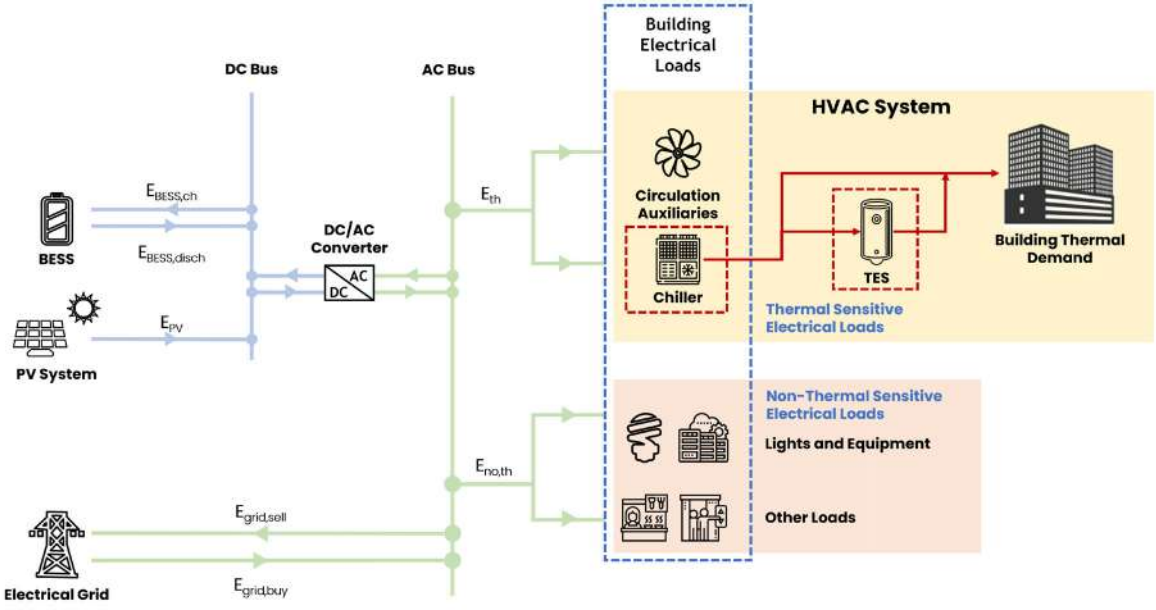

Fig. 2 Scheme of the integrated energy system installed in target buildings

perform peak shaving (i.e., power withdrawal from the grid less than a defined threshold) in addition to the planned objectives for electrical cost and indoor temperature requirements.

2.1 Design of baseline rule-based and DRL control agents and training of the source DRL controller

The initial phase of the proposed methodology involves the development of RBC and DRL controllers for source buildings for the management of the cooling system.

The RBC controller RBC_{th} is divided into two agents: the first one determines the provisioning of cooling energy to the building (identified as RBC_{CF}), and the second one determines the operational mode of the energy system (identified as RBC_{OM}). These two agents are correlated since the operation mode of the energy system depends on the supply of cooling energy to the building. Comprehensive insights into the design of the RBC can be found in Section 3.

The primary objective of the source DRL agent lies in minimising electricity expenses while ensuring suitable indoor temperature conditions during occupancy hours. Developing a DRL controller requires the definition of its pivotal elements: action space, state space and the reward function. The DRL controller undergoes offline training within the source building. During this phase, an automated process is executed through Optuna (Akiba et al. 2019) to establish the best configuration of hyperparameters for the DRL control algorithm. Complete details on the DRL training phase can be found in Section 3.

2.2 Transfer of DRL control policy in heterogeneous inductive and transductive settings

The second phase of the framework aimed at executing TL to share the optimal control policy of the source controller with the agents to be deployed in the target buildings.

The TL strategy employed in this paper is categorised as heterogeneous TL since DRL controllers implemented on the source and target buildings have different state-space due to the integration of PV and BESS in energy systems implemented in target buildings in addition to chiller and TES. Nevertheless the energy system is different between source and target buildings, the action-space remains the same as in source DRL controller, managing the operation of chiller and TES. In this framework, the operation of PV and BESS in target buildings is managed by RBC_{el}. Thus, the size of the output layer corresponding to the action chosen by the DRL controller remains the same, therefore all the layers of the neural network approximating the DRL control policy are transferred. The experiments are conducted to evaluate heterogeneous TL in both transductive and inductive settings. In the transductive setting, the DRL controller transferred to target buildings optimises the same objective function as in the source building, while in the inductive setting the objective function to be optimised is different compared to the one considered in the source building. An online TL strategy is developed to enhance the scalability of the transfer procedure.

The knowledge transfer strategy employed is weight-initialisation, which entails the exploration of the exchange of neural network parameters between the source and target

controllers. In detail, the weights of the Actor and Critic networks within the target controllers are initialised using the weights of the pre-trained source DRL controller. Then, a fine-tuning procedure is conducted to update the neural network weights and adjust the control policy in response to the unexplored boundary conditions inherent in the target buildings.

However, this knowledge transfer strategy that combines weight-initialisation and fine-tuning would operate effectively only if the number of states and actions is the same for the source agent and the target agents. This scenario does not align with the case study of this work, as the number of states in the state-space of the source controller is lower compared to that of the target controllers. In fact, new variables related to the operation of the PV system and BESS are included in the state-space of target controllers to achieve an optimal control policy.

In this framework, the traditional approach would involve initialising the weights of all layers in the neural networks approximating the control policy of the target building with those pre-trained on the source building, except for the weights corresponding to the input layer, which would be initialised from scratch. However, this approach is not compatible with the online TL methodology adopted in this study, since the single training episode employed for the fine-tuning process (as expected by the online TL strategy), would be insufficient to retrieve an optimal control policy for the target building. The solution adopted in this work to overcome the aforementioned challenges is related to the implementation of the model slicing technique.

Model slicing is a technique frequently employed in the training and inference of neural networks to address various objectives such as model size reduction or deployment on resource-constrained devices (Zhang et al. 2022b). This strategy mainly involves partitioning a neural network model into smaller, independent segments, each of which can be treated as a self-contained model with a subset of the original parameters. However, the model slicing technique should be adapted for the specific application and available resources.

In this paper, model slicing was implemented to enhance the performance of the online TL process. Specifically, for each controller implemented on the target building, the corresponding Actor and Critic neural networks are initialised. These networks have a larger number of neurons in the input layer (i.e., 86 inputs) compared to the source controller (i.e., 59 inputs). Afterwards, the input layer is divided into two parts as shown in Figure 3: one with an input layer of the same dimensions as that of the source agent (i.e., 59 × 64 neurons), and the other part corresponding to the new inputs introduced to obtain an optimal control policy in the target controllers (i.e., 27 × 64 neurons). This approach allowed the weight-initialisation process for the target controller to employ the pre-trained weights of the source controller, preserving the pre-existing knowledge of the source controller regarding the operation of the components that remained unchanged in the target building (i.e., chiller and TES). Then, the target controller is fine-tuned while operating within the building implementing a different energy system and new environmental conditions (e.g., different weather conditions).

However, the fine-tuning process is preceded by an imitation learning phase where the RBC strategy is implemented during the first week of the deployment period (i.e., from 1 June to 7 June) to initialise the memory buffer of the OTL agent with transitions derived at each

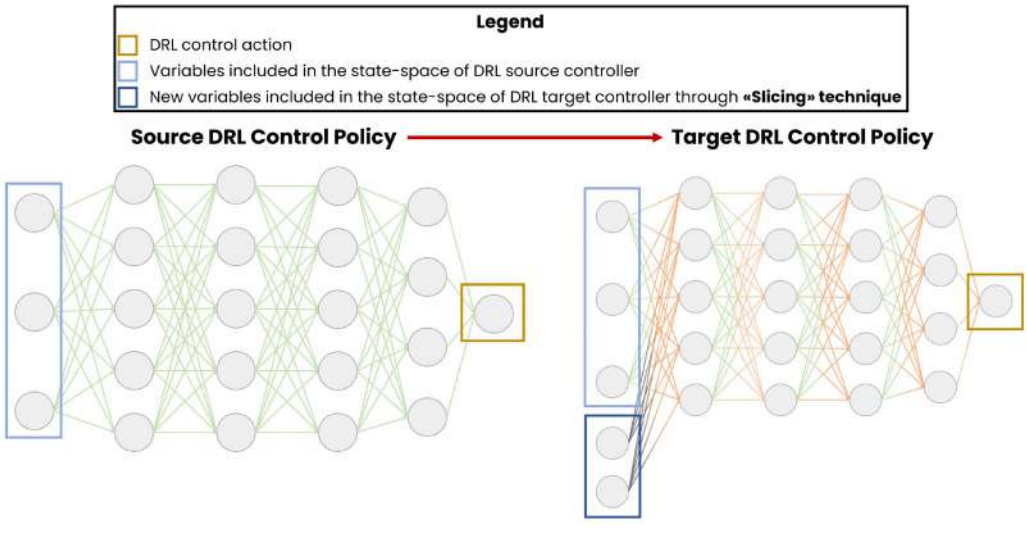

Fig. 3 Implementation of slicing technique to transfer control policy for DRL agents with different numbers of inputs (in a heterogeneous setting)

control time step from the RBC operation. This process has proven effective in enhancing the capability of OTL agent in learning the relationship between selected actions, states and the reward function during the initial days of deployment.

Throughout the fine-tuning process, the values of two DRL hyperparameters, learning rate and number of gradient steps, should be properly selected to enhance controller performance. The learning rate determines the extent of network adjustment in response to the computed error each time the network weights are updated, while the number of gradient steps indicates how many randomly drawn batches from the memory buffer experience gradient updates at each control time step (Brandi et al. 2022a). According to Li et al. (2020), the learning rate adjustment is carried out to ensure that the target agent can update its control policy without entirely discarding the pre-existing knowledge obtained from the source building. Therefore, the learning rate is reduced by half compared to the value adopted during the training phase of the DRL source control agent. Additionally, a number of gradient steps higher than the standard values typically employed for offline DRL controllers is adopted to prevent performance degradation in the OTL controller. For detailed information regarding the values of the specific parameter employed in the OTL strategy, please refer to Section 3.

2.3 Performance benchmarking of online TL strategy on target buildings

In the final phase of the methodology, the performance of OTL is benchmarked with that of offline and online DRL controllers, as well as the RBC. The inclusion of RBC serves as a benchmark to offer a comparison with a conventional control strategy frequently employed in real buildings. The effectiveness of these controllers is evaluated in the context of different target buildings, established by integrating PV and BESS into the energy systems. This enables a comprehensive evaluation of online TL performances in heterogeneous transductive settings, characterised by changing weather conditions, electricity price schedules, occupancy patterns and thermal properties of both opaque and transparent building envelopes with respect to the source building. Moreover, the implementation of PV and BESS allows the evaluation of the heterogeneous inductive TL where the objective function of the transferred DRL controller is modified compared to the source building scenario.

The offline DRL training strategy envisages a recursive process in which the control agent interacts with the environment to be controller over multiple training episodes to achieve a stable control policy. The recursive training is followed by the static deployment of the agent in which any further updates of the control policy occur since the agent does not modify its behaviour according to the feedback received from the environment after taking an action (Brandi et al. 2022a). This approach ensures a certain level of stability for the control policy but exhibits a notable drawback: in the event of modification in the controlled environment, the controller must be retrained. Additionally, its practical implementation proves challenging as it mandates several episodes to enhance the control policy and requires modelling effort to derive a reliable model of the building to be controlled.

The online DRL training strategy foresees that the control agent learns the parameters of the optimal policy while actively managing the system and without any prior understanding of the dynamics inherent to the controlled environment (Brandi et al. 2022a). To replicate a real-time implementation scenario, the training of the online DRL strategy is executed within a single episode, as opposed to multiple episodes in the case of offline DRL. The strength of the online DRL strategy lies in its model-free nature, eliminating the need to develop a model of the building. Nevertheless, in the early stages of the training period, the agent possesses limited knowledge about the control problem, and there exists a significant risk that the chosen controller actions yield suboptimal performance. In this framework, the memory buffer of the online DRL agent is initialised with transitions acquired from the operation of the RBC, which is essentially an imitation learning approach (Coraci et al. 2023b). The performance of the online DRL strategy depends strongly on the value of the number of gradient steps and learning rate. In the offline approach, a constant value for both the learning rate and the number of gradient steps is employed, considering that the agent benefits from the collection of experiences in the replay buffer garnered across numerous episodes. In contrast to the offline DRL strategy, within the online DRL approach a higher number of gradient steps is employed, facilitating the exploration process and expediting learning following the initial week of online DRL implementation. However, employing an extensive number of gradient steps could lead to the potential convergence of the control agent to an optimal deterministic control policy as the training progresses. To address this issue, the time step in which the learning process happens is increased relative to the offline DRL strategy. Further details about offline and online DRL control strategies are provided in Section 3.

3 Implementation

This section describes the implementation details for the analysed case study and for the methodological framework described in Section 2.

3.1 Implementation details on case study

As stated in Section 2, source and target buildings are identical from a geometrical point of view and they are served by the same cooling system consisting of a chiller and a TES. The case study is a prefabricated building with a rectangular layout and it covers a floor area of 196 m^2 and a net conditioned floor area of 97 m^2. The facility comprises two 10-person office rooms, one 3-person control room and a technical room. The technical room is not connected to the HVAC system and it houses the storage tank and the battery. The window-to-wall ratio is 7% while the average transmittance values for the opaque and transparent envelope components of the source buildings are 0.16 W/(m^2·K) and 0.55 W/(m^2·K), respectively. The source building is located in Turin (Italy), and it is occupied at maximum capacity from Monday to Friday between 8:00 and 18:00. In addition to the load associated with the HVAC system, in the two office rooms a total electric load (E_{LOAD}) of 0.9 kW for lighting services, appliances and office equipment is measured during the occupancy hours.

The building and the cooling system are modelled in EnergyPlus. Terminal units within buildings were not modelled. Consequently, the cooling energy is technically provided to the building by means of the *OtherEquipment* object in EnergyPlus. The use of the *OtherEquipment* object allows for a simplified representation of the cooling energy supply. The cooling system is sized considering as an external disturbance the cooling power to meet the demand of the building demand and employing the ideal-load EnergyPlus calculation for the reference weather file available in EnergyPlus for a specific location.

For the source building the reference weather file for Turin, Italy (ITA-TORINOCASELLE-IGDG.epw) is employed. The sizing process results in a design cooling power required to maintain an indoor temperature of 26 °C and relative humidity of 55% during the occupancy period of 1.8 kW and 1 kW respectively for each office zone and the control room, while the chiller is designed with a Q_{cap} capacity of 4.7 kW. Moreover, the reference Coefficient of Performance (COP) of the chiller is equal to 2.7, determined according to a reference leaving and entering fluid temperatures equal to 6.7 °C and 35 °C, while the supply water temperature at chiller outlet is 7 °C. TES is designed to have a size of 3 m^3, corresponding to three times the maximum ideal hourly cooling demand of the building. The minimum $T_{s,min}$ and maximum $T_{s,max}$ operative temperatures of TES are 10 °C and 18 °C. These temperatures correspond respectively to the maximum state of charge (SOC_{TES} = 1) and the minimum state of charge (SOC_{TES} = 0) of the TES. During the charging phase, the design water mass flow rate is set at 0.2 kg/s, while during the discharging phase, it corresponds to the sum of the design mass flow rates of the office rooms and control room, equal to 0.35 kg/s.

The same sizing approach is employed to determine the design features of the cooling system for each target building, according to the specific weather conditions.

A time-based tariff structure commonly implemented in Italy, known as Time-Of-Use (TOU), determines the electricity price for the electricity withdrawn from the grid to operate the chiller unit and auxiliary equipment. The TOU structure divides the electricity price into low price, medium price and high price as a function of the day of the week and the time of the day. These tariff rates are designed to differentiate the values for the optimisation application, starting from a real value of the high price period. In detail, the medium price is derived from the average electricity price of 0.143 €/kWh observed during the period June–September 2021, as indicated by the Italian grid regulating authority (ARERA 2022), while the low price and high price are set respectively to 1/2 (i.e., 0.071 €/kWh) and 3/2 (i.e., 0.214 €/kWh) of the medium price. Moreover, since the energy system designed for target buildings includes PV and BESS, it is required to define the price $c_{el,sell}$ at which the excess electrical energy generated by the PV system is sold to the grid, assumed to be 0.02 €/kWh based on data obtained from the Italian regulator (Brandi et al. 2022b).

The main difference between the source building and the target buildings lies in the integration of PV system and BESS on the energy system. A 3-kW nominal power for the PV system is selected to align with the peak power requirement of the overall electrical demand of the building. The PV system is modelled in the simulation environment by means of a Python class as in Brandi et al. 2022(b), and employing solar position data from the Python library *pvlib* (Holmgren et al. 2018). The PV system consists of mono-crystalline silicon modules, each having a nominal power of 200 W/m^2 and efficiency (η) defined in Equation (1) and equal to 15%.

$$\eta = f(G, AM, T_{out}) \tag{1}$$

The efficiency of PV modules is evaluated under standard conditions (i.e., solar irradiance G_{STC} = 1000 W/m^2, cell temperature T_{STC} = 25 °C, air mass AM_{STC} = 1.5) as described in Durisch et al. (2007). To compute the PV power production P_{PV} as the product of efficiency and incident solar radiation is required to define the tilt and azimuth angles. The inclination angle of the PV panels is retrieved from the global dataset presented by Jacobson and Jadhav (2018). Consequently, the tilt angle is fixed at 33° while the azimuth angle depends on the testing facility orientation and is equal to 116°.

A modular Lithium-ion BESS is implemented in the

simulation environment by employing a Python class as defined in the literature (Amato et al. 2021; Brandi et al. 2022b) that simply emulates the BESS operation by estimating the SOC_{BESS} since the BESS degradation is not considered. However, this approach is sufficiently accurate for conducting an initial assessment of the effects of BESS installation. Therefore, SOC_{BESS} is defined as follows:

$$\begin{cases} SOC_{BESS}(t) = SOC_{BESS}(t-1) + \eta_{rte}\dfrac{P_{BESS,ch}(t)\cdot\Delta t}{C_{BESS}} & \text{(charge)} \\ SOC_{BESS}(t) = SOC_{BESS}(t-1) - \dfrac{P_{BESS,dis}(t)\cdot\Delta t}{C_{BESS}} & \text{(discharge)} \end{cases} \quad (2)$$

The SOC of the BESS is computed per each time step t as a function of the SOC at the previous time step $SOC_{BESS}(t-1)$ and the battery capacity C_{BESS}. In detail, during the BESS charging process, the SOC is evaluated by including a round-trip efficiency η_{rte} of 96% and the power charged from PV to BESS between two consecutive time steps (i.e., Δt, equal to 15 minutes). Otherwise, during the BESS discharging process the $SOC_{BESS}(t)$ is computed as a function of the power discharged from BESS in Δt. The BESS capacity is 2.4 kWh, defined according to the building electrical peak demand on hourly basis. To ensure the lifespan of the battery and prevent excessively rapid charging and discharging operations, two limits are imposed on the charging and discharging processes. These limits are defined as $P_{BESS,ch,max}$ and $P_{BESS,dis,max}$ corresponding to 25% and 50% of the nominal capacity of the (BESS). Moreover, to safeguard the health of the battery and optimise its performance, SOC levels are bounded by minimum and maximum values, as indicated by the manufacturer. For a Li-ion BESS technology, these values are set respectively at 10% and 90% of the total capacity, allowing for a total Depth of Charge (DoC)/Depth of Discharge (DoD) of 80% (Amato et al. 2021).

As introduced in Section 2, the electrical part of the energy system is always managed by an RBC strategy identified as RBC_{el} across the different experiments carried out in the paper. Figure 4 shows a flowchart of RBC_{el} strategy which is inspired by state of the art BESS management strategies (Ruusu et al. 2019; Amato et al. 2021). The RBC_{el} operates according to the energy production from PV (i.e., E_{PV}), the electricity cost (i.e., c_{el}) and the state of charge of the BESS (i.e., SOC_{BESS}). According to Italian regulations, BESS charging from the grid is not permitted. Therefore, the BESS can only be charged when the energy generated by the PV system exceeds the energy demand of the building to satisfy the energy required by the cooling system and non-HVAC electrical loads. At each control time step, if the local energy production from PV system is not zero, RBC_{el} injects energy into the building/grid environment following this priority order: (i) PV energy production is employed to meet building electrical demand, (ii) PV energy production is employed to charge the BESS, (iii) PV energy production is sold to the grid. The BESS charging process complies with the limits defined for the maximum chargeable energy $E_{ch,max}$ and SOC_{BESS}. When the PV energy production exceeds the total electrical demand of the building and the capacity of the BESS, the excess energy is sold to the grid. Based on information obtained from the Italian regulator, the price for the electricity sold to the grid is 0.02 €/kWh. Otherwise, when the PV energy production is not sufficient to meet the total building electrical demand,

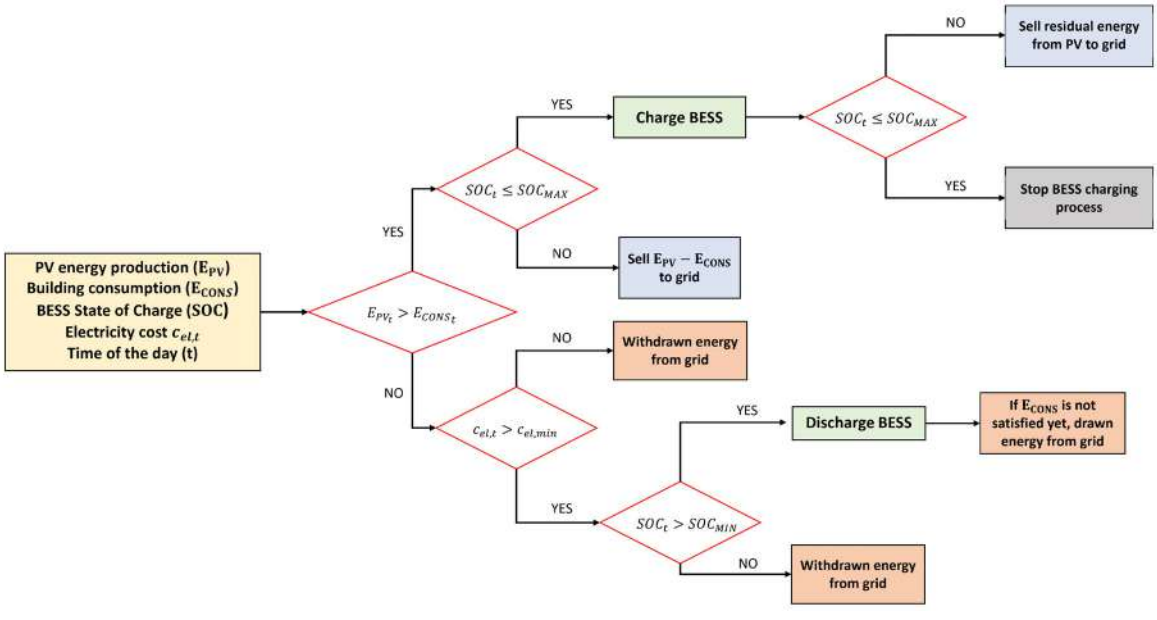

Fig. 4 Flowchart of the action selection process of the rule-based controller for the electrical part of the energy system

the BESS should be discharged. The discharge process is carried out according to the electricity purchasing price from the grid c_{el}, the prescribed limits for the maximum amount of dischargeable energy by the BESS $E_{dis,max}$ and SOC_{BESS}. The BESS is discharged when the electricity purchasing price from the grid is greater than the minimum electricity price of the implemented TOU electricity price schedule of the building to maximise cost savings associated with energy purchasing from the grid. However, if the energy discharged from the BESS is not sufficient to meet the electrical demand of the building, the energy is purchased from the grid.

The experiments are executed by means of a two-side co-simulation environment shown in Figure 5. The co-simulation environment employs a Building Control Virtual Test Bed (BCVTB) middleware and the *ExternalInterface* function of EnergyPlus to establish a connection between EnergyPlus (Crawley et al. 2001), where the building dynamics and the thermal part of the energy system (i.e., air-to-water chiller and TES) are modelled, and a Python interface built upon OpenAI Gym (Brockman et al. 2016) employed to develop PV, BESS and RBC/DRL controllers models. In this paper, the simulation time step is set to 15 minutes, ensuring convergence of the EnergyPlus simulation process (Coraci et al. 2023b). The control time step is set to 30 minutes, as a consequence, each control action is executed across a span of two simulation time steps.

The EnergyPlus side of the co-simulation environment receives at each simulation time step control actions from the Python side. The EnergyPlus model implements the control action and generates outputs related to the thermal energy system (i.e., TES SOC), indoor environmental parameters (i.e., indoor air temperature), and supplementary data (i.e., total building electrical consumption) integrated within the state-space of the controller. Furthermore, the EnergyPlus model provides supplementary details, such as direct and diffuse solar irradiation, employed by the PV model to compute the PV power generation.

The outputs from the EnergyPlus and information about the electricity price schedule are supplied as inputs to the Python side. In detail, information about electricity price and building total electrical consumption is provided as input to the RBC strategy managing the BESS. The updated BESS SOC value is obtained from the BESS model, which receives directives regarding whether to charge or discharge the battery. Then, BESS SOC, PV electricity production and electricity price are additionally relayed as state variables. These variables are used by DRL controller or RBC to determine the cooling mode and the cooling fraction during each control time step. To conclude, the controller learns the optimal control strategy through the assessment of a reward function, formulated as a function of the analysed scenario (e.g., inductive or transductive TL).

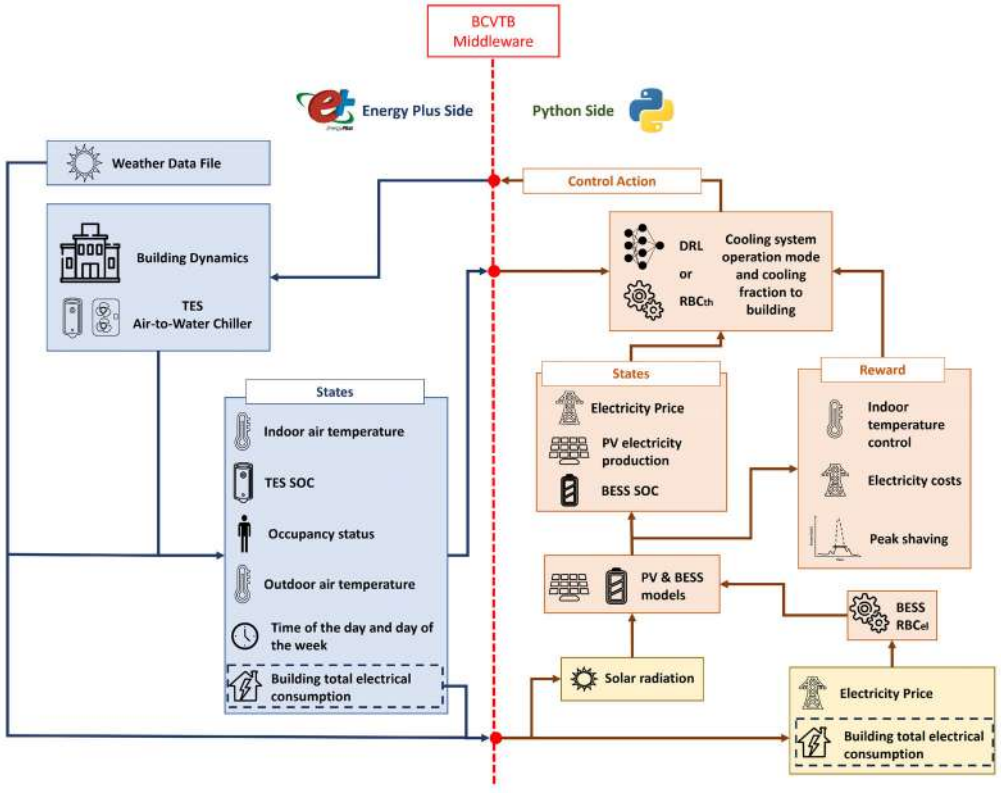

Fig. 5 Co-simulation environment architecture (modified from Brandi et al. (2022b) and Coraci et al. (2023b))

3.2 Implementation of baseline rule-based controller

The baseline controller consists of a RBC, defined as RBC_{th}, managing the operation of the chiller and TES and serves as the baseline controller. The RBC_{th} consists of two parts responsible for determining the fraction of nominal cooling energy (i.e. cooling fraction) to be delivered to the building (RBC_{CF}) and for determining the operational mode of the cooling system (RBC_{OM}) respectively. The control strategy of RBC_{CF} is activated on weekdays when the building is occupied and it is divided into two stages: pre and post-initial switch ON phase. In the pre-switch ON phase, the agent chooses the starting time to supply cooling energy to the building based on specific indoor temperature conditions and the time of day. After the switch ON, RBC_{CF} supplies cooling energy to the thermal zone as per the conditions outlined in Table 1. Once the initial phase is completed, cooling energy is provided to the building until the indoor temperature drops below the lower acceptability threshold, T_{LOW} (i.e., 25 °C). Conversely, when the indoor temperature exceeds the upper temperature acceptability threshold T_{UPP} (i.e., 27 °C) the RBC_{CF} agent resumes the cooling energy supplying to the building. The cooling energy is provided until the occupants leave the building (i.e., 18:00). The second part of the RBC_{th}, defined as RBC_{OM}, is developed to manage the operation mode of the cooling energy system. In detail, when the electricity price is low and the state of charge (SOC_{TES}) is below 0.75 (i.e., TES temperature of 12 °C), RBC_{OM} operates the cooling system in charging mode until the electricity price rises above the minimum value or SOC_{TES} reaches the maximum value (i.e., SOC_{TES} equals 1 or TES temperature equals 10 °C). If RBC_{CF} decides to supply cooling energy to the building and the electricity price is not low, RBC_{OM} operates the cooling system in discharging mode as long as SOC_{TES} is non-zero, and in chiller mode if the storage is empty.

3.3 DRL controller design

This section describes the design process of the DRL control agent, providing details about the definition of its main features: action-space, state-space and reward function.

Table 1 Starting time conditions followed by RBC_{CF} to start providing cooling energy to the building

Combination	Time period	Indoor temperature
1	$4{:}00 \le t < 5{:}00$	$T_{INT} - T_{UPP} \ge 3$ °C
2	$5{:}00 \le t < 6{:}00$	$T_{INT} - T_{UPP} \ge 2$ °C
3	$6{:}00 \le t < 7{:}00$	$T_{INT} - T_{UPP} \ge 1$ °C
4	$t \ge 7:00$	$T_{INT} - T_{UPP} \ge 0$ °C

3.3.1 Design of action-space

The action-space is discrete, as a discrete version of the SAC is employed. The action-space remains unchanged between the DRL agents implemented in the source and target buildings, as the electrical part of the energy system (i.e., PV+BESS) serving the target buildings is always controlled by the RBC_{el} strategy. In this framework, the action-space is defined as indicated in Equation (3).

$$A = OM \times CF = (om, cf) : om \in OM, cf \in CF \tag{3}$$

The action-space includes the combination of five feasible control actions, expressed in the range [0, 4], that can be executed by the DRL control agent to manage the operation mode of the cooling system (i.e., Operation Mode (OM)) and decide whether or not to supply cooling energy to the building (i.e., Cooling Fraction (CF)). The five combinations of feasible control actions are included in Table 2. Furthermore, safety constraints are implemented to prevent the system from operating in charging mode when the storage is fully charged (i.e., $SOC_{TES} = 1$) and in discharging mode when the storage is empty (i.e., $SOC_{TES} = 0$). In such scenarios, the system operates in chiller mode.

3.3.2 Design of state-space

The state-space consists of a collection of observations provided as inputs to the agent, influencing its decision-making process regarding the action selection. In this work, the state-space varies between the source and target DRL controllers, as per the definition of heterogeneous TL, since the target agents should receive information concerning the operation of the electrical side of the energy system (implemented in the target buildings). This aspect allows the DRL agent to effectively determine the optimal sequence of control actions for the cooling system operation, leveraging the information on the PV system and BESS. Figure 5 reports in *States* boxes the variables included in the state-space for the DRL controller. In detail, three supplementary variables associated with the operation of the electrical system and the building have been integrated for the controllers implemented in target buildings in addition

Table 2 Correspondence between action picked by DRL controller and the two control actions required by cooling system

Action	Operation mode	Cooling fraction
0	0	0
1	1	0
2	−1	−1
3	0	−1
4	1	−1

to the variables included in the state-space of DRL source controller.

Including the *TES SOC* in the state-space is crucial to provide the agent with sufficient information to enhance the cooling system management. This variable is observed at the current time step t and the two previous timesteps (i.e., $t - 1$ and $t - 2$) having insights about the stored cooling energy amount and its temporal evolution. Information about *Indoor temperature* conditions is included in the state-space for the same timesteps by employing the difference between the indoor setpoint and the actual indoor air temperature. This allows the agent to account for the temperature fluctuations within the building over time and consider the thermal dynamics impact of the building and energy system, as discussed in Brandi et al. (2022b). Including *Electricity price* in the state-space is crucial due to its impact on the reward. As the electricity price schedules are assumed to be known, the state-space incorporates perfect predictions of the electricity price for the next 12 hours. This allows the agent to optimise the operation mode of the cooling system. By combining three features included in the state space, namely *Day of the week*, *Time of the day* and the (0, 1) binary variable *Occupants' presence status*, the controller can recognise the occupancy schedule. This latter variable is included in the state-space for the current time step and the following 12 hours. The last variable evaluated in the state-space for both the source and target controllers is the *Outdoor air temperature*, as it affects the operation and energy consumption of the cooling system and indoor temperature. The three variables added to the state-space for the target DRL controllers are the *BESS SOC*, *PV energy production* and the *Building total electrical consumption*. The *Building total electrical consumption* and *PV energy production* play a pivotal role in the optimal management of the controlled system. Moreover, the information related to the PV operation is provided to the agent from the current control time step t to the subsequent time steps $t + 24$ (i.e., the next 12 hours), with the assumption that perfect predictions are known in advance. Solar radiation is not included in the state-space despite the PV energy production is inherently dependent on this variable. However, the agent indirectly accounts for the impact of solar irradiation on the system dynamics by incorporating the PV energy generation. The *BESS SOC* is another necessary information provided to the agent for the optimal management of the cooling system. The BESS is employed to supply electricity to the chiller, pumping system, lighting services and appliances in building during periods when the electricity price is not low. However, unlike the TES that exhibits greater inertia, the BESS SOC is provided only at the current timestep t due to its lower inertia. The state-space variables are scaled using a min-max normalisation technique to ensure that the observations are converted within a range of (0, 1) before being fed into the neural network.

3.3.3 Design of reward function

The formulation of the reward function should be aligned with the objectives of the control problem. In this work, two different formulations of the reward function are developed, based on the type of TL experiments evaluated. In the heterogeneous transductive TL setting the source and target controllers optimise the same objective function. Otherwise, in the case of heterogeneous inductive TL the controller is trained on a source building to optimise a specific objective function and then transferred to a target building with a different objective function.

In the case of transductive TL, the reward function for both source and target DRL controllers is defined as a linear combination of two terms: an electricity cost-related term and a temperature-related term. The objective of the DRL control agents is to minimise the costs associated with energy withdrawn from the grid while also improving indoor temperature conditions. Furthermore, this formulation of the reward function does not consider the excess of energy generated by the PV system and sold to the grid, as the goal is also to maximise Self-Sufficiency (SS) and Self-Consumption (SC). The importance of the two reward terms is determined by two weights, δ for the electricity cost term and β for the temperature term. The general formulation of the reward function is as follows:

$$R = -\left(\delta \times R_{\mathrm{E}} + \beta \times R_{\mathrm{T}}\right) \tag{4}$$

The electricity cost-related term corresponds to the cost incurred for the energy withdrawn from the grid to supply the chiller, circulation systems and non HVAC electrical loads and it is defined as:

$$R_{\mathrm{E}} = c_{\mathrm{E}} \times \left(E_{\mathrm{CHILLER}} + E_{\mathrm{PUMP}} + E_{\mathrm{LOAD}}\right) \tag{5}$$

where c_{E} [€/kWh] represents the electricity price for buying energy from the grid and depends on the implemented price schedule. The variables E_{CHILLER} [kWh] and E_{PUMP} [kWh] correspond to the energy consumption of the chiller and pumping system respectively, while E_{LOAD} [kWh] represents the energy consumption of lighting services, appliances and building equipment. However, E_{LOAD} is not included in R_{E} for source agent since the energy required to satisfy the energy demand of lighting services and building equipment can be purchased only from the grid and the control action of the DRL agent can not influence their operation.

The temperature-related term is determined by evaluating the presence of occupants and the indoor temperature

conditions. When the building is not occupied, the temperature-related term is zero (i.e., $R_T = 0$). Conversely, during working hours the temperature-related term is formulated differently based on the indoor temperature values, as reported in Equation (6).

$$R_T = \begin{cases} (\mid SP_{INT} - T_{INT} \mid)^3 & \text{if } T_{LOW} - 2 \le T_{INT} < T_{LOW} \text{ and } T_{UPP} < T_{INT} \le T_{UPP} + 2 \\ 50 & \text{if } T_{INT} < T_{LOW} - 2 \text{ and } T_{INT} > T_{UPP} + 2 \\ 0 & \text{if } T_{LOW} \le T_{INT} \le T_{UPP} \end{cases} \quad (6)$$

The reward was computed using the third power ($n = 3$) of the $|SP_{INT} - T_{INT}|$ term after a sensitivity analysis where three different values (i.e., $n = 1,2,3$) were evaluated, since this formulation of the reward ensured better performance concerning the control objectives. To mitigate convergence issues in the learning process related to the high magnitude of the reward, a fixed value was assigned to the reward if the indoor temperature is below 23 °C or above 29 °C. This approach avoids convergence issues since the learning process of the SAC algorithm influences the definition of the Boltzmann temperature coefficient α, defined as a function of the reward magnitude (Coraci et al. 2023b).

The implementation of heterogeneous inductive TL involved the introduction of a new term within the reward function designed for the source DRL controller. This term is related to the reduction of peak energy consumption from the grid. In this case, the objective function is defined as follows:

$$R_{inductive} = -(\delta \times R_E + \beta \times R_T + \theta \times R_P) \quad (7)$$

R_P is the peak-related term, defined as $R_P = 1$ when the power absorption from the grid is greater than 2 kW (i.e., 0.5 kWh since the simulation time step is equal to 15 minutes). The R_P is combined with the other two reward terms (defined as above) by employing θ as peak-related term weight.

3.4 Implementation details on design of target buildings

In this paper, a series of experiments are conducted to compare the performance of heterogeneous online TL in the context of both inductive and transductive scenarios, against three benchmark controllers: RBC, online DRL, and offline DRL. Inductive transfer learning involves modifying the objective function between source and target building controllers. Specifically, it simulates a scenario where the controller pre-trained in the source building and designed to minimise energy cost and optimise indoor temperature control, remained implemented in the same building after integrating PV and BESS in the energy system in addition to chiller and TES and modifying the objective function to accommodate the inductive transfer learning process. The objective function used in the source building R is represented by Equation (4), while the one considered for the building after the implementation of PV and BESS $R_{inductive}$ is presented in Equation (7), including the peak shaving term. Otherwise, the performances related to transferring the source controller to various target buildings with the same geometrical features but improved energy system and different boundary conditions are evaluated in the case of heterogeneous transductive TL. The modification of boundary conditions between the source building and target buildings involved adjustments to various factors, including weather conditions, electricity price schedules, occupancy schedules, and building thermophysical properties.

The experiments are carried out in simulative way by developing for each target building an EnergyPlus model employed as a proxy of the real building. The target building models are retrieved by changing the features of the source model and employing those indicated in Figure 6.

To assess the influence of climate on the control policy transfer process, various scenarios are evaluated based on the classification provided by the European Commission, accounting for Cooling and Heating Degree Days to categorise areas with distinct climate characteristics (Tsikaloudaki et al. 2012; PVSites 2016). These scenarios included the use of the same (i.e., Turin) or similar (i.e., Paris) climatic conditions as the source building, as well as significantly different conditions in warmer (i.e., Palermo) or colder (i.e., Helsinki) locations.

Additionally, the target building configurations are characterised by the adoption of different electricity price and occupancy schedules. Specifically, two price schedules are considered: the first schedule is based on TOU similar to that of the source building. The second schedule followed an on-off peak price scheme, utilising the electricity tariffs from Austin, Texas (Austin Energy 2023) and considering the price for the electricity sold to the grid equal to 0.008 €/kWh. Details about the two price schedules are provided in Table 3.

The two implemented occupancy schedules differ in terms of weekdays and occupancy hours. The first occupancy schedule assumes that the building is occupied during the period from Monday to Friday between 8:00 and 18:00. The other schedule accounted for the presence of occupants throughout the week, from Monday to Sunday from 7:00 to 19:00. In conclusion, different combinations of envelope efficiency for each target building matching the thermophysical properties of both the opaque envelope (i.e., opaque thermal transmittance U_{OP} and internal heat capacity χ_i) and the transparent envelope (i.e., transparent thermal transmittance

Same as source building Different from source building

Building	Energy system	Reward function	Weather conditions	Price schedule	Occupancy schedule	Envelope efficiency
Source	Chiller+TES	$R = -(\delta * R_E + \beta * R_T)$	Turin	Time of Use	Mon-Fri 8:00-18:00	Configuration 0
$T_{inductive}$	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T + \theta * R_P)$	Turin	Time of Use	Mon-Fri 8:00-18:00	Configuration 0
T_0	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Turin	Time of Use	Mon-Fri 8:00-18:00	Configuration 0
T_1	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Turin	On-off peak	Mon-Fri 8:00-18:00	Configuration 0
T_2	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Turin	Time of Use	Mon-Sun 7:00-19:00	Configuration 0
T_3	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Turin	On-off peak	Mon-Sun 7:00-19:00	Configuration 0
T_4	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Turin	On-off peak	Mon-Sun 7:00-19:00	Configuration 1
T_5	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Paris	Time of Use	Mon-Fri 8:00-18:00	Configuration 0
T_6	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Paris	On-off peak	Mon-Fri 8:00-18:00	Configuration 0
T_7	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Paris	Time of Use	Mon-Sun 7:00-19:00	Configuration 0
T_8	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Paris	On-off peak	Mon-Sun 7:00-19:00	Configuration 0
T_9	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Paris	On-off peak	Mon-Sun 7:00-19:00	Configuration 2
T_{10}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Helsinki	Time of Use	Mon-Fri 8:00-18:00	Configuration 0
T_{11}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Helsinki	On-off peak	Mon-Fri 8:00-18:00	Configuration 0
T_{12}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Helsinki	Time of Use	Mon-Sun 7:00-19:00	Configuration 0
T_{13}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Helsinki	On-off peak	Mon-Sun 7:00-19:00	Configuration 0
T_{14}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Helsinki	On-off peak	Mon-Sun 7:00-19:00	Configuration 3
T_{15}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Palermo	Time of Use	Mon-Fri 8:00-18:00	Configuration 0
T_{16}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Palermo	On-off peak	Mon-Fri 8:00-18:00	Configuration 0
T_{17}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Palermo	Time of Use	Mon-Sun 7:00-19:00	Configuration 0
T_{18}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Palermo	On-off peak	Mon-Sun 7:00-19:00	Configuration 0
T_{19}	Chiller+TES and PV+BESS	$R = -(\delta * R_E + \beta * R_T)$	Palermo	On-off peak	Mon-Sun 7:00-19:00	Configuration 4

Fig. 6 Overview of the main features of the target buildings analysed during the performance benchmarking phase

Table 3 Comparison of electricity price schedules [€/kWh] implemented in target buildings

(a) Time-of-use

Hour of the day	Day		
	Mon–Fri	Sat	Sun
00:00–07:00	0.071		
07:00–08:00	0.143		0.071
08:00–19:00	0.214	0.143	0.071
19:00–23:00	0.143		0.071
3:00–24:00	0.071		

(b) On-off peak

Hour of the day	Day		
	Mon–Fri	Sat	Sun
00:00–07:00	0.029		
07:00–0:00	0.063		0.029
0:00–24:00	0.029		

U_{TR} and solar heat gain coefficient g) are evaluated. Table A1 in Appendix A2 provides details about the five envelope configurations for the target buildings.

3.5 Implementation details on training of DRL source controller and benchmarking strategies

This section offers a comprehensive overview of the DRL control agent training phase in the source building, as well as the deployment in the target buildings of the OTL and DRL controllers trained both offline and online.

The training phase of the DRL controller in the source building involves an automated procedure to identify the best set of hyperparameters between twenty configurations of the DRL control agent. Each configuration is trained for 30 episodes consisting of 90 days (from 1 June to 29 August). On average, the simulation process of each episode required approximately 35 minutes to complete on a machine equipped with an 8th Generation Intel® Core™ i7-8550U processor operating at 4.0 GHz and 16.0 GB of RAM. The optimisation process is carried out by means of the open-source Python library Optuna (Akiba et al. 2019). Optuna operates by either minimising or maximising an objective function over a set of hyperparameters defined within an acceptability range. In this paper, the DRL source controller hyperparameters involved in the optimisation are *Reward weights* δ and β, *Learning rate* μ, *Discount factor* γ, *Number of hidden layers* and *Number of neurons per hidden layer*, while the *Batch size* is fixed at 128. The objective of the optimisation process is to identify the best configuration that enables the agent to effectively minimise electricity cost ($E_{cost,source}$), measured in €, and the cumulative sum of temperature violations (T_{viol}), measured in °C. $E_{cost,source}$ is the total cost of electricity withdrawn from the grid since the implemented energy system does not include a PV system.

$$E_{cost,source} = c_E \times (E_{CHILLER} + E_{PUMP}) \quad (8)$$

T_{viol} is computed as the cumulative of $T_{viol,i}$ (i.e., i-th temperature violation). $T_{viol,i}$ is determined by computing the absolute difference between the indoor temperature T_{INT} and either the lower limit T_{LOW} or upper limit T_{UPP} of the temperature acceptability range, which is defined as [25, 27] °C, when the indoor air temperature exceeds these boundaries during the occupancy period.

$$T_{viol} = \sum_{i=0}^{N} T_{viol,i} \tag{9}$$

where N is the number of simulation time steps in a cooling season. Details about the search domain and the best values for DRL hyperparameters are included in Table A2 in the Appendix A2. The magnitude of the search domain was defined according to the values adopted for *Learning rate* μ, *Discount factor* γ, *Number of hidden layers* and *Number of neurons per hidden layer* adopted in previous works reported in the literature (Brandi et al. 2020), while the magnitude of the search domain for the reward weights was designed to avoid any convergence issues of the DRL algorithm related to high reward magnitude values.

DRL-based control agents implemented in target buildings during the performance benchmarking phase have different values for some hyperparameters, such as batch size, learning rate, learning step and gradient steps, while δ, β and α remain consistent with those of the transferred source DRL agent since the re-optimisation process is not carried out during the benchmarking phase. The automated optimisation of hyperparameters performs better when the controllers are trained over multiple episodes (e.g., multiple seasons in real building applications) as in source DRL agent or offline DRL strategy implemented in target buildings. However, this procedure is in contrast with the online DRL and OTL strategies developed in this study, implemented for only one episode during the cooling season (from 1 June to 29 August) and aiming to represent their direct implementation in real buildings. The offline DRL strategy has a control time step of 30 minutes, a batch size of 128 and a single gradient step. Conversely, for the online DRL and OTL strategies the batch size is 32, the learning step is extended to every three days while the number of gradient steps and learning rate is respectively 30 and 0.001 for online DRL and 15 and 0.0005 for OTL. As suggested in Smith et al. (2017), the batch size is reduced to 32 for both online strategies due to the limited data volume available for training the online controllers and to speed up the convergence process towards an optimal solution. Increasing the number of gradient steps to 30 provides significant benefits to the online DRL controller, expediting the training process during the initial weeks of implementation where the agent has a limited amount of transition data (i.e., consisting of state, action, new state and reward) stored in the memory buffer to properly train the control agent. Moreover, a learning step of three days degrades the online DRL strategy performances in the early stages of training, but it ensures that the control agent accumulates a larger number of transitions before proceeding to the next learning step and improves the performance throughout the training period. In the case of OTL, the increase of the number of gradient steps to 15 coupled with the reduction of the batch size (i.e., 32) and of the learning rate (i.e., 0.0005) with respect to the source controller ensures that the pre-trained source control strategy is not entirely overwritten.

The learning rate adjustment allows the control policy to be optimised according to the different boundary conditions in the target building while preventing excessive exploration of the action space, which could lead to deviations from the optimal control policy learned at the beginning of the training phase (Li et al. 2020). The learning rate value is reduced by half compared to that used for the source DRL controller following a sensitivity analysis in which different learning rate values are evaluated while keeping other DRL hyperparameters the same. The sensitivity analysis is carried out by evaluating the online TL in target building T_{19}, representing the building with the highest degree of modifications in boundary conditions compared to source building.

In this framework, Table 4 provides a performance comparison in terms of electricity cost and the cumulative sum of temperature violations considering different learning rate values, suggesting that a value of 0.0005 for the learning rate ensures the best performance for the developed online transfer learning process. It is worth noting that the learning rate is a hyperparameter depending on the specific task and the similarity between source and target domains. Therefore, it might be useful to provide a guideline in future research to assist energy managers and system integrators in choosing the correct learning rate value.

Moreover, the Mahalanobis distance (Kaya and Bilge 2019) is computed to demonstrate the effectiveness of learning rate adjustment during the online TL process. The

Table 4 Performance comparison considering different learning rate values for online TL implemented in T_{19}

Learning rate	E_{cost} [€]	T_{viol} [°C]
0.0001	31.9	194.6
0.0002	32.2	185.9
0.0003	34.1	179.0
0.0004	33.1	172.5
0.0005	32.1	146.6
0.0006	36.0	193.5
0.0007	39.3	216.9
0.0008	40.0	225.6

learning rate adjustment allowed the transferred controller buildings to adapt the DRL control policy to the new boundary conditions in target buildings without completely overwriting the knowledge pre-acquired in the source building. In this framework, the Mahalanobis distance is computed for the target building T_{19}, considering the difference for the weights of neural network representing the DRL control policy before and after fine-tuning in online TL and online DRL (with weights randomly initialised). The average Mahalanobis distance for the weights differences of all layers is lower for the online TL approach (i.e., 0.96) compared to the case of online DRL (i.e., 1.42), indicating that the weights distribution for online TL is more similar than the case of online DRL. Therefore, employing a lower learning rate ensures that the pre-existing knowledge from source building is not entirely discarded.

To conclude, when the price schedule implemented in the target building follows an on-off peak pattern, the reward weight factor δ for the electricity cost term is doubled compared to that of the source agent for offline DRL, online DRL and OTL controllers. This adjustment is required to maintain a balance between the two terms in the reward function since TOU price tariff implemented in the source building features a higher average weekly electricity price compared to the on-off peak tariff. The performances of OTL in target building are benchmarked with those of offline DRL, online DRL and RBC in terms of E_{cost} and T_{viol}. E_{cost} is computed as the product of the electricity cost withdrawn from the grid and the sum of the energy consumption in kWh of the chiller, auxiliaries (similarly to the $E_{cost,source}$) and non-HVAC electrical loads (i.e., associated to appliances and lighting services) reduced by revenues derived from surplus of PV energy sales to the grid (I_E).

$$E_{cost} = c_E \times \left(E_{CHILLER} + E_{PUMP} + E_{LOAD} \right) - I_E \quad (10)$$

4 Results

This section summarises in two separate subsections the results of the pre-training phase of the DRL controller on the source building and those related to the implementation of the online TL strategy.

4.1 DRL controller pre-training process on source building

Before implementing the online TL process, the DRL controller was offline pre-trained on the source building. The offline pre-training process, as defined in Section 3, included the optimisation of its hyperparameters, for which values are included in Table A2 in the Appendix. The objective of the DRL agent was to optimise both electricity cost $E_{cost,source}$ and indoor temperature conditions (by minimising T_{viol}) by managing the operation mode of the cooling system and choosing the fraction of cooling energy to be supplied to the thermal zones compared to the RBC baseline.

Figure 7 compares DRL and RBC during the last week of July (starting from Sunday) in terms of the electrical energy consumption of chiller and circulation pump, also providing details regarding the utilisation of TES by showing the SOC time series. RBC and DRL controllers exhibited similar energy consumption profiles during high-price periods (in dark blue), since they operated the cooling system in discharging mode by activating only the circulation pump to supply cooling energy to the environment from the TES.

During weekends, RBC and DRL controllers employed a different strategy to manage the cooling system. In detail, the RBC charged the TES starting from Friday night and during Saturday morning (i.e., 07/30–07/31), while the DRL controller charged the TES during Sunday (i.e., 07/25). As a result, the DRL approach minimised TES losses and maintained the maximum SOC at the beginning of the occupancy period, in contrast to the RBC strategy. Moreover, during weekdays DRL controller charged the TES during low-price periods (in white) by operating the system in charging mode, supplying at the same time cooling energy to the environment through the chiller. This strategy allowed to reduce the energy consumption from the chiller during medium or high electricity price periods by operating the cooling system in discharging mode. Conversely, RBC strategy operated the cooling system in chiller mode during the last stages of high-price periods, as the TES was discharged before the end of the occupancy period, to meet indoor temperature requirements by providing cooling energy to the building. The management of the energy system for RBC and DRL controllers followed a recurrent weekly pattern throughout the entire 90-day cooling season, resulting in a 20% electricity cost saving (i.e., $E_{cost,source,DRL}$ = 56.0 € vs $E_{cost,source,RBC}$ = 70.0 €). Moreover, DRL controller ensured better performance in terms of indoor temperature control (i.e., $T_{viol,source,DRL}$ = 54.0 °C), since it reduced the cumulative sum of temperature violation by 70% compared to RBC (i.e., $T_{viol,source,RBC}$ = 176.0 °C).

Figure 8 compares the indoor temperature profiles related to the operation of DRL and RBC strategies during the same week analysed in Figure 7. The reward function formulation encouraged the DRL agent to supply cooling energy to the building by activating the chiller during low-cost hours, as illustrated in Figure 7. Since the low-cost electricity price tariff occurred during the night, over this period the DRL controller pre-cooled the indoor environment in advance compared to RBC to ensure that the temperature remained closest to the acceptable range as the occupancy

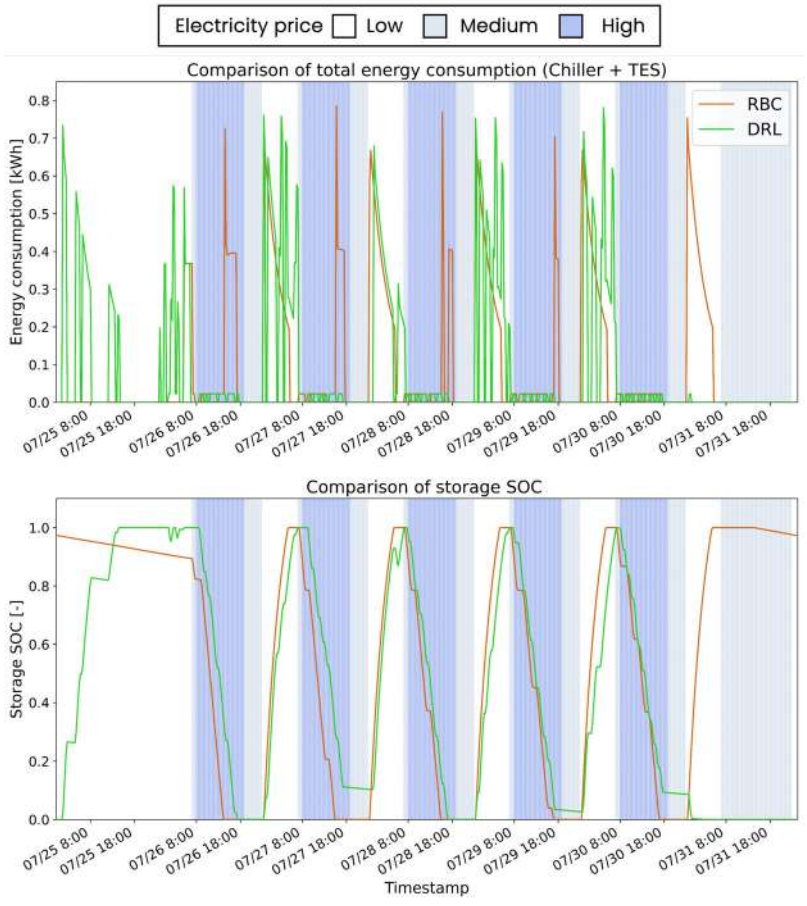

Fig. 7 Source building total electricity use (Chiller+TES) and SOC profiles for RBC and DRL controllers

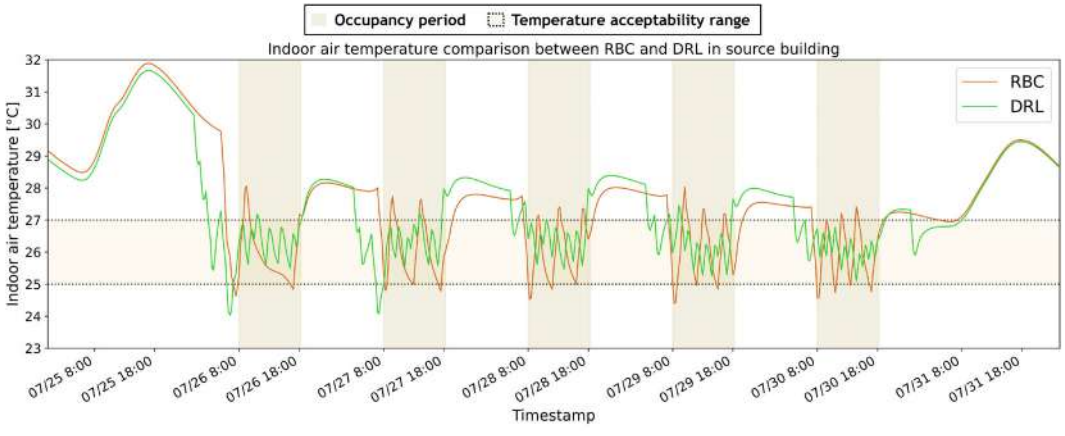

Fig. 8 Source building indoor air temperature profiles for RBC and DRL controllers

period begins. This pattern was more evident on Mondays (i.e., 07/26) since the building must be pre-cooled earlier compared to the other weekdays, as on Saturdays and Sundays there were no occupants and temperature control is not required given the temperature term formulation in the reward function (see Equation (6)). Overall, during the occupied hours the DRL controller exhibited a more homogeneous temperature profile around the temperature setpoint of 26 °C, minimising temperature violations related to exceeding the upper value of the indoor temperature acceptability range (i.e., 27 °C).

4.2 Benchmarking of heterogeneous online transfer learning performance on target buildings

The results related to the implementation of the heterogeneous

online TL strategy on the target buildings are summarised in this section, providing a benchmark of its performance compared to RBC, offline DRL and online DRL strategies. The benchmarked controllers were implemented over a single cooling season (i.e., one episode) lasting 90 days, while the offline DRL strategy involved pre-training on 30 episodes for each target building before its static deployment on the same building to test controller performance.

Figure 9 and Table 5 provide an overview of the performance obtained from the implementation of the investigated control strategies across all target buildings in heterogeneous transductive settings, respectively in terms of total electricity cost (E_{cost}) and indoor temperature control. Indoor temperature control performances are assessed in terms of the mean value of the daily average temperature violation rate $\overline{\Delta T_{viol,daily}}$, computed as the daily average of the ratio between the cumulative daily sum of temperature violations $T_{viol,daily}$ and the daily temperature violations occurrences $T_{viol,occ,daily}$. Hence, the buildings are clustered from the upper to the lower side of Figure 9 and Table 5 in ascending order based on the degree of deviation in weather conditions compared to the source building. The locations include those with equal (Turin), similar (Paris), colder (Helsinki), and warmer (Palermo) climates compared to the source building. Each target building is identified by a specific ID, whose main features are indicated in Figure 6.

As emerged from Figure 9 and Table 5 the offline DRL controller provided the best performances compared to the other strategies due to a more refined control policy retrieved from a 30 episodes training period. Conversely, online TL and online DRL were directly implemented on target buildings while actively controlling the energy system to emulate their direct implementation in physical buildings, and then its implementation relies on a single simulation episode. As a result, offline DRL outperformed online TL within a range of 1% (i.e., T_4) and 58% (i.e., T_{18}) in terms of electricity cost, even reaching differences exceeding 80% in target buildings with different weather conditions compared to source building (e.g., Helsinki T_{11} and Palermo T_{16}). Moreover, offline DRL enhanced indoor temperature conditions by reducing on average the mean value of the daily average temperature violation rate by 40% compared

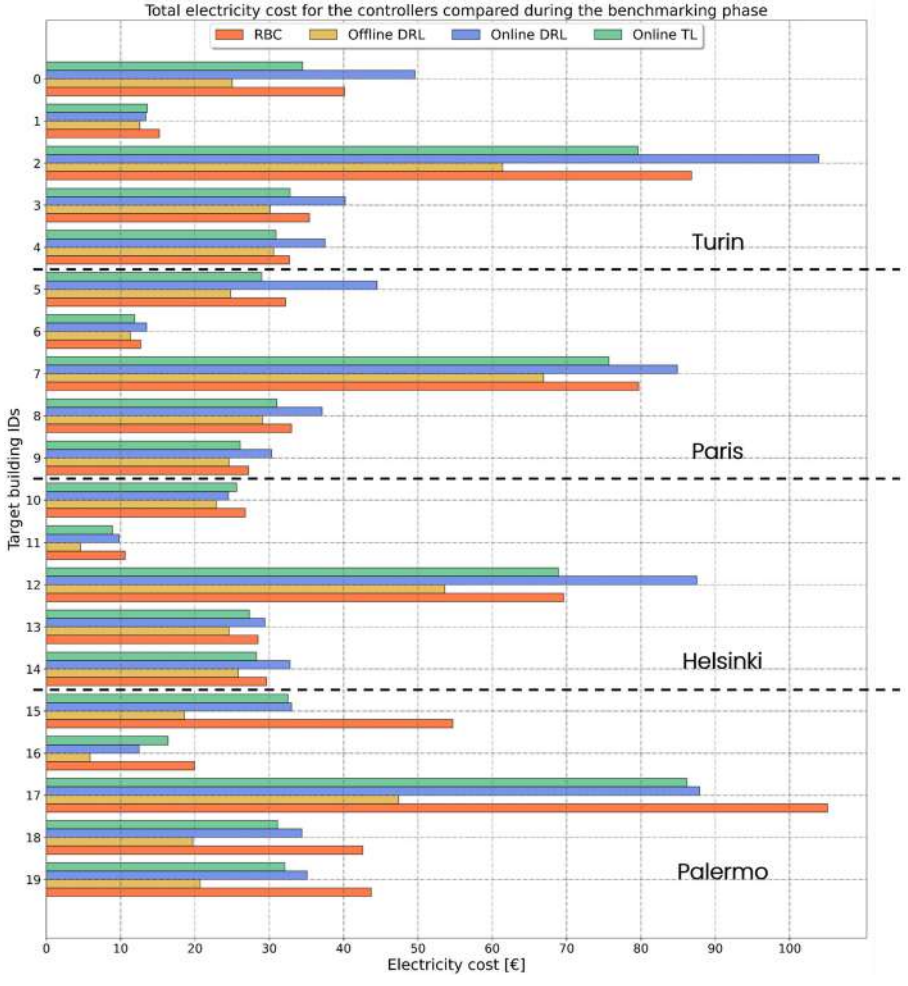

Fig. 9 Comparison of total electricity cost for the controllers implemented in target buildings during performance benchmarking phase

to the online TL strategy. However, the online TL agent achieved better performance in terms of both total electricity cost and indoor temperature control when compared to RBC and online DRL controllers. In detail, the online TL agent achieved significant electricity cost savings, ranging from 5% (i.e., Paris T_7) to 40% (i.e., Palermo T_{15}), and ensured an average reduction of 10% for $\overline{\Delta T_{viol,daily}}$ across all experiments compared to the RBC. Moreover, the online TL strategy outperformed online DRL by reducing the electricity cost between 5% (i.e., target buildings T_{10}) and 31% (i.e., target building T_{16}) and on average $\overline{\Delta T_{viol,daily}}$ by 41%.

From an accurate analysis of the results shown in Figures 9 and Table 5, climatic conditions emerged as the primary driver that affected the performances of the developed TL methodology. The weather-related differences played a pivotal role since thermal and electrical load patterns within the analysed building were influenced by weather conditions. Specifically, the overall performance achieved by online TL outperformed on average those of RBC and online DRL when climatic conditions are similar to those of source building, while also being more similar to those obtained by implementing offline DRL.

For target buildings located in Turin or Paris, online TL reached electricity cost savings of approximately 9% and 18% compared to RBC and online DRL, even though its performance was approximately 13% worse than offline DRL. Similarly, a better indoor temperature control was established by implementing online TL since $\overline{\Delta T_{viol,daily}}$ is reduced by 15% and up to 40% T_{viol} compared to RBC and online DRL and achieved an average value of $\overline{\Delta T_{viol,daily}}$ that is approximately 25% higher than that of offline DRL.

Conversely, the effectiveness of the TL process was reduced for the target buildings located in Helsinki (i.e., colder climate) and Palermo (i.e., warmer climate). In these buildings, the savings achieved through online TL decreased respectively to 6% and 9% in terms of electricity cost and approximately to 8% and between 20% and 35% in terms of $\overline{\Delta T_{viol,daily}}$ compared to RBC and online DRL. Similarly, the negative gap in performance between online TL and offline DRL was 50% greater both in terms of electricity cost and $\overline{\Delta T_{viol,daily}}$.

Lastly, from Figure 9 and Table 5 emerged that changing the price schedule under the same weather conditions has not affected the relative performance level achieved by online TL when compared to the benchmark controllers. Contrarily, the effectiveness of the knowledge-sharing process experienced a slight decrease when the occupancy schedule was changed, but not to the same extent observed when changing climatic conditions. This outcome might be related to the transferred controller requiring more engagement with the target building for adapting the control policy to the new occupancy schedule where the building was occupied also during weekends and for a larger period compared to the source building.

Table 5 Comparison of the mean value of the daily average temperature violation rate $\overline{\Delta T_{viol,daily}}$ over the testing period for RBC, offline DRL, online DRL controllers with online TL implemented in heterogeneous transductive setting in target buildings

ID target building	Average temperature violation rate [°C]			
	RBC	Offline DRL	Online DRL	Online TL
T_0	0.18	0.09	0.3	0.16
T_1	0.18	0.11	0.21	0.16
T_2	0.26	0.21	0.39	0.24
T_3	0.25	0.18	0.36	0.24
T_4	0.27	0.19	0.37	0.24
T_5	0.17	0.1	0.41	0.12
T_6	0.17	0.12	0.33	0.17
T_7	0.26	0.18	0.33	0.24
T_8	0.24	0.12	0.38	0.26
T_9	0.27	0.27	0.32	0.27
T_{10}	0.17	0.15	0.35	0.17
T_{11}	0.18	0.09	0.31	0.16
T_{12}	0.24	0.14	0.38	0.25
T_{13}	0.25	0.15	0.52	0.21
T_{14}	0.24	0.15	0.34	0.22
T_{15}	0.22	0.15	0.35	0.16
T_{16}	0.23	0.1	0.35	0.17
T_{17}	0.26	0.15	0.44	0.22
T_{18}	0.26	0.15	0.34	0.26
T_{19}	0.29	0.17	0.3	0.24

As defined in Section 3, the electricity-related term reward function in target buildings does not include the income from selling energy to the grid to maximise self-sufficiency and self-consumption. Self-sufficiency and self-consumption are computed taking into account that PV system and BESS can both provide electrical energy to the building to activate chiller and the circulation pump and to feed lighting services and appliances in building. Specifically, the formulations for SS and SC are as follows:

$$SS = \frac{E_{PV,b} + E_{BESS,b}}{E_{TOT,b}} \tag{11}$$

$$SC = \frac{E_{PV,b} + E_{BESS,b}}{E_{PV,tot}} \tag{12}$$

where $E_{PV,b}$ and $E_{BESS,b}$ are the total energy provided respectively by PV and BESS to the building, $E_{TOT,b}$ is the total building electrical energy consumption and $E_{PV,tot}$ is the PV total energy production.

In this framework, Table 6 provides a performance benchmarking in terms of SS and SC achieved by the RBC,

offline DRL, and online TL controllers for the buildings in which the transfer methodology was evaluated in the transductive heterogeneous setting. The online DRL strategy has not been included in this comparison, as demonstrated by the results presented in Figure 9 and Table 5, suggesting that its implementation leads to significantly adverse performance in terms of E_{cost} and $\Delta T_{viol,daily}$ compared to the other controllers. Table 6 demonstrates that the results obtained for SS and SC were consistent with those found for E_{cost} and $\Delta T_{viol,daily}$ for RBC, offline DRL and online TL strategies. In detail, the online TL strategy effectively managed the energy system to harness the flexibility provided by the PV and BESS to meet the building energy demand in all target building configurations, resulting in an increase of 9% for SS and 11% for SC compared to RBC. In contrast, online TL values for SS and SC were respectively 24% and 19% lower than those obtained by implementing the offline DRL agent.

To conclude, Figures 10 and 11 show respectively for offline DRL and online TL, an overview of electrical and thermal energy flows, as well as indoor temperature and TES/BESS SOC profiles. These figures allow for a comparative analysis of how these controllers manage the energy system when the source controller was transferred to the target building with ID $T_{inductive}$ (see Figure 6 for more details) in the heterogeneous transductive setting.

Table 6 Comparison of self-sufficiency and self-consumption for RBC and offline DRL controllers with online TL implemented in heterogeneous transductive settings in target buildings

	Self-sufficiency (SS) [%]			Self-consumption (SC) [%]		
ID target building	RBC	Offline DRL	Online TL	RBC	Offline DRL	Online TL
T_0	48.9	68.5	56.2	47.8	66.8	56.5
T_1	49.2	75.9	52.9	48.1	72.7	50.5
T_2	48.6	62.7	53.7	75.5	94.9	85.4
T_3	46.7	60.5	50.3	72.7	92.5	79.5
T_4	48.5	60.6	51.1	72.5	92.5	88.4
T_5	51.1	63.8	55.7	47.2	59.4	51
T_6	51.1	76.7	53.5	47.4	72.9	59.4
T_7	51.1	55.2	51.8	76.5	82.2	77.3
T_8	48.9	62.4	50.6	73.4	93.8	76.3
T_9	52.2	66.3	52.5	72	89.6	72.5
T_{10}	52.6	83.9	56.8	43.5	66.1	62.7
T_{11}	52.5	80.5	61.4	43.5	66.7	50.3
T_{12}	53.8	67.3	54	72.6	89.7	72.4
T_{13}	51.7	65.9	53.7	70.1	90.1	72.2
T_{14}	51	65.9	51.1	70	92	70.5
T_{15}	46.8	85.6	62.6	41.4	73.3	56
T_{16}	47	86.3	54	41.8	75.7	48.6
T_{17}	47.1	74.5	50.7	62.5	95.8	65.1
T_{18}	44.6	73	51.5	59.4	93.8	67.5
T_{19}	45.4	79.8	52.3	61.4	84.6	68.6

In this case, the RBC was used to assess whether the online TL strategy could maintain acceptable performance in terms of electricity cost and indoor temperature control while maintaining below 0.5 kWh the energy purchased from the grid (or the power absorption from the grid below 2 kW, since it is evaluated during each simulation time step of 15 minutes). Despite the cost reduction (i.e., −3%) and the enhancement in indoor temperature conditions compared to a potential implementation of RBC (i.e., −21% in T_{viol}), the online TL strategy exceeded the 0.5 kWh limit for grid-supplied energy on 132 occurrences during the first cooling season with a peak demand of 0.8 kWh, demanding the extension of the implementation period to assess the period required (i.e., number of episodes or cooling seasons) to achieve adequate performance also in terms of peak shaving. Thus, it emerges that two episodes are sufficient for transferring to the target building $T_{inductive}$ a DRL controller in heterogeneous inductive setting, since during the second cooling season the peak shaving limit was exceeded only 12 times out of 8640 observations, reducing the maximum peak value of energy withdrawn from the grid to 0.55 kWh. Furthermore, the online TL strategy revealed excellent performance in both electricity cost ($E_{cost,TL}$ = 34.3 €) and cumulative sum of temperature violations ($T_{viol,TL}$ = 17.2 °C). However, the offline DRL controller exceeded the peak shaving limit only two times out of 8640 observations since it relies on a more refined control policy as being trained on 30 episodes, reducing electricity cost by 50% ($E_{cost,off-DRL}$ = 16.5 €) but having poorer performance in terms of internal temperature control ($T_{viol,Off-DRL}$ = 96.3 °C) compared to online TL.

Figures 10 and 11 reveal several similarities in how the energy system was managed by the offline DRL controller and the online TL controller after updating its control policies while actively managing the system two successive cooling seasons. Offline DRL and online TL approaches managed the system to withdraw energy from the grid during low-electricity price periods. The energy import from the grid was used to activate the chiller to pre-cool the indoor environment and to charge the TES (especially closer to the occupancy period to minimise TES losses). This operational pattern led to an increased electricity demand during nighttime hours when there was no PV electricity generation. During the occupancy period in working days, both offline DRL and online TL agents maximised self-consumption by leveraging the energy production from PV system and minimised the energy selling to the grid. In detail, the chiller was activated using the energy produced by the PV system, both to meet the building energy demand and to charge the TES if the SOC_{TES} was

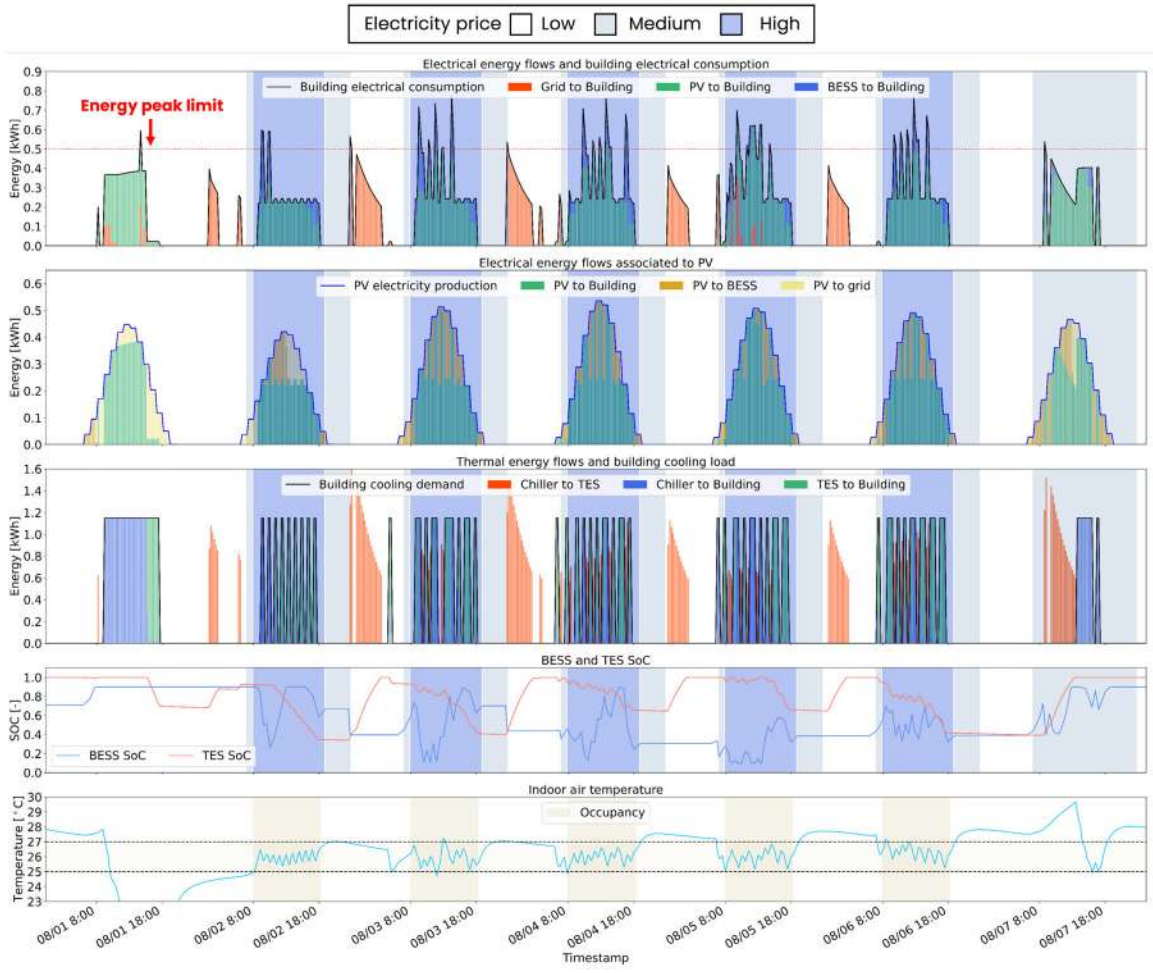

Fig. 10 Overview of electrical/thermal energy flows associated to building and energy system, BESS/TES SOC and indoor air temperature for offline DRL strategy (after 30 cooling seasons)

not at the maximum charge level. If there was surplus energy generated by the PV system and the SOC_{BESS} was not at the maximum value (i.e., SOC_{BESS} = 0.9), the BESS was charged before selling energy to the grid. Conversely, when the PV electricity generation was not sufficient to meet the overall building energy demand, TES and BESS were discharged. The TES satisfied the building thermal energy demand, while the BESS supplied energy to the auxiliary pump, appliances and lighting in building. During the analysed week, the offline DRL agent did not exceed the peak limit, while the online TL strategy violated the peak limit twice during the last hours of the occupancy period on August 6th. These violations occur due to a mishandling of the thermal-side energy system. Specifically, the chiller was incorrectly activated to provide cooling energy to the environment, even though it did not require energy, as its temperature was within an acceptable range. Consequently, the indoor temperature drops below 25 °C (resulting in a temperature violation), and the total energy demand of the building cannot be met solely by the PV system. Therefore, a 0.52 kWh amount of energy was drawn from the grid, exceeding the prescribed limit and resulting in two peak violations.

In conclusion, during the weekends (i.e., 08/01 and 08/07), the offline DRL controller sold a smaller amount of energy back to the grid compared to the online TL controller to maximise self-consumption, according to the definition of the electricity cost-related term in reward function in target buildings. Notably, on August 1st the offline DRL exploited the energy production from PV to activate the chiller. Considering that the TES is fully charged, the temperature drop is associated with the supply of cooling energy to thermal zones. As a result, the indoor environment is pre-cooled to ensure that the temperature is maintained as close as possible to the temperature acceptability range during the early stages of the occupancy period on Monday, rather than waiting until the nighttime hours near the start of the occupancy period on Monday, as carried out by the online TL controller. This different approach contributes to the difference in energy sold to the

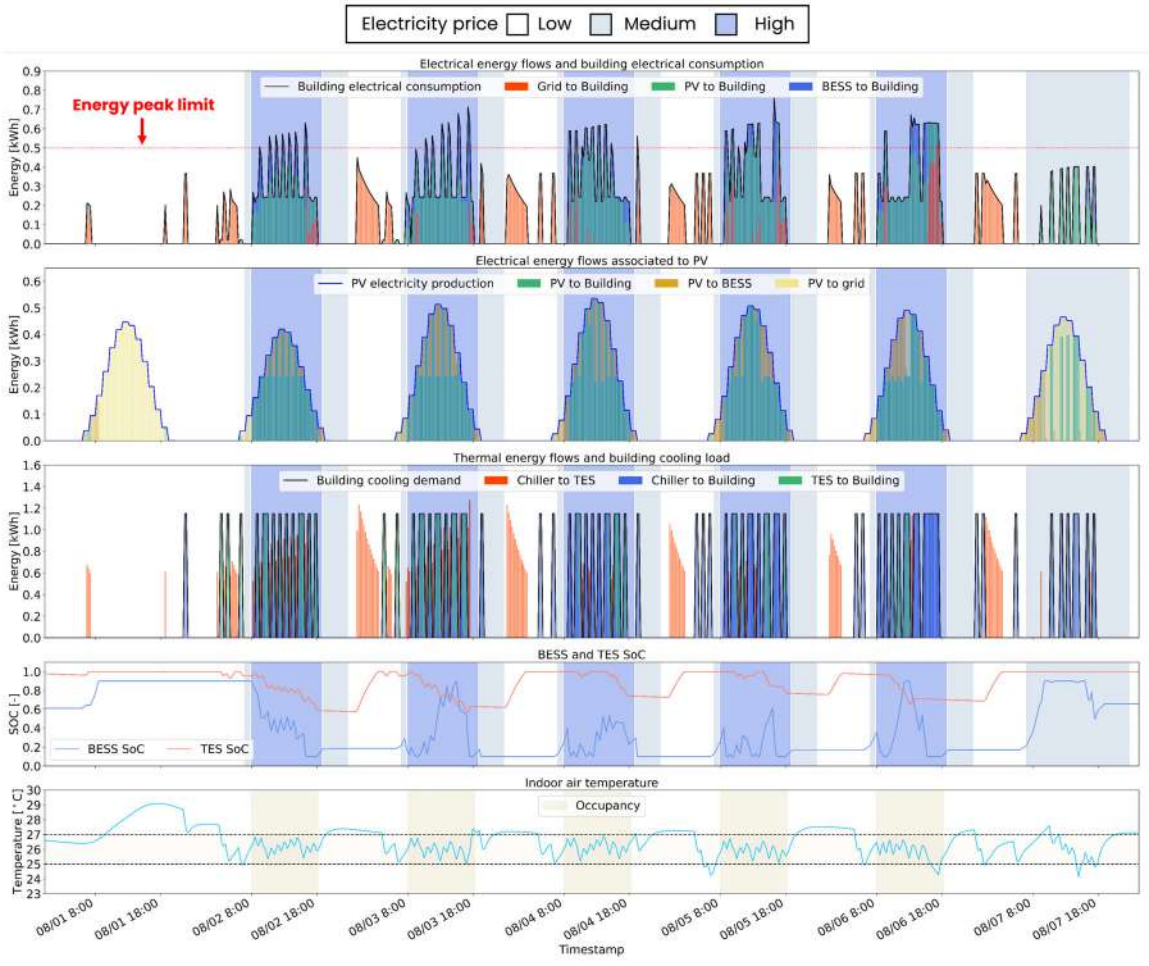

Fig. 11 Overview of electrical/thermal energy flows associated to building and energy system, BESS/TES SOC and indoor air temperature for online TL strategy (after 2 cooling seasons)

grid during the weekend periods between the two analysed strategies.

5 Discussion

In recent years the need for systematic procedures that enable the effective transfer of pretrained DRL controllers to buildings with different characteristics has emerged since real buildings are equipped with different energy systems and should be managed to handle different objectives. In this context, this study introduced a novel online heterogeneous transductive and inductive TL strategy for DRL controllers, combining model slicing, imitation learning, weight-initialisation and fine-tuning techniques.

The weight-initialisation strategy operates effectively only when the number of states and actions is consistent between the source and target agents. However, this ideal scenario does not align with the case study presented in this work. In the case of the target controllers, the state-space includes additional variables related to the operation of PV system and BESS, which are crucial for achieving an optimal control policy. This difference in the state-space between the source and target controllers poses a significant challenge. To overcome this limitation, model slicing was strategically employed to enhance the performance of the online TL process.

The implementation of model slicing in this study showcases a practical solution to address the challenge of divergent state-spaces between source and target controllers. By partitioning the neural networks and selectively preserving knowledge relevant to both buildings, this strategy contributes to the successful deployment of the online TL methodology in a heterogeneous transfer setting. Model slicing not only overcomes the limitations posed by differences in state-space but also enhances the adaptability and efficiency of DRL controllers in real-world building energy management applications.

The primary focus of this study was to explore the implementation of online heterogeneous TL for a DRL controllers pre-trained offline in a source building. The

source DRL controller managed the operation mode of the cooling system, consisting of an electric chiller and a TES, and chose whether or not to provide cooling energy to the building with the objectives of minimising electricity cost and enhancing indoor temperature conditions. This approach was applied in transductive and inductive setting.

In heterogenous transductive TL, target buildings are equipped with a different energy system, since PV and BESS are introduced. In this context, the results obtained for target buildings implementing the same weather conditions as source buildings demonstrate how TL can support the revamping process of an existing energy system. Online TL was found effective in extreme scenarios where target buildings are characterised by different climatic conditions, electricity pricing, occupancy and thermo-physical properties compared to source building. In this context, the results demonstrate the effectiveness of the online TL strategy, since the developed methodology significantly outperforms RBC and online DRL controllers, achieving substantial electricity cost savings and enhancing indoor temperature control while maximising SS and SC in all target building configurations. Despite offline DRL consistently outperforming the other strategies its applicability is limited compared to the online TL controllers since it requires to perform again the offline training process.

Moreover, this study demonstrates that climatic conditions are the primary driver affecting the performance of the online TL strategy, since it performed better when climatic conditions were similar to those of the source building (i.e., Turin or Paris), while the performance gap between online TL and offline DRL increased for buildings located in significantly different climates (as in Helsinki or Palermo).

Furthermore, the developed online TL strategy succeeded in adapting the pre-trained DRL controller to new control objectives in two episodes (corresponding to two cooling seasons) when it is tested for a target building with the same energy system as the other target buildings (i.e., heterogeneous TL), identical boundary conditions as in the source building but including in the objective function the peak shaving in addition to the source controller objectives (i.e., inductive TL).

While offline DRL remains the best solution, even if it is not practically model-free, online TL emerges as a valuable and efficient solution to enhance the scalability of DRL controllers in real buildings while ensuring performance comparable to offline DRL controllers. To quantify the effectiveness of the developed online TL, the definition of metrics and Key Performance Indicators (KPIs) could be beneficial. However, in the context of TL applications for building control, KPIs are not generalisable and should be defined according to each control problem. Therefore, to assess the benefits of online TL in avoiding the development of a building surrogate model for pre-training purposes as in offline DRL controllers, Figure 12 shows a bar chart indicating the number of episodes required for offline DRL to achieve comparable performance to online TL for each target building. The proposed transfer methodology allows for direct deployment of the transferred DRL controller, achieving performance comparable to a DRL agent trained offline for a number of episodes ranging between 15 and 30 episodes as indicated in Figure 12, depending on the characteristics of the target building.

It's worth noting that the results presented in this study are based on simulations, and the validation in real buildings could be included in future works. Conducting real-world

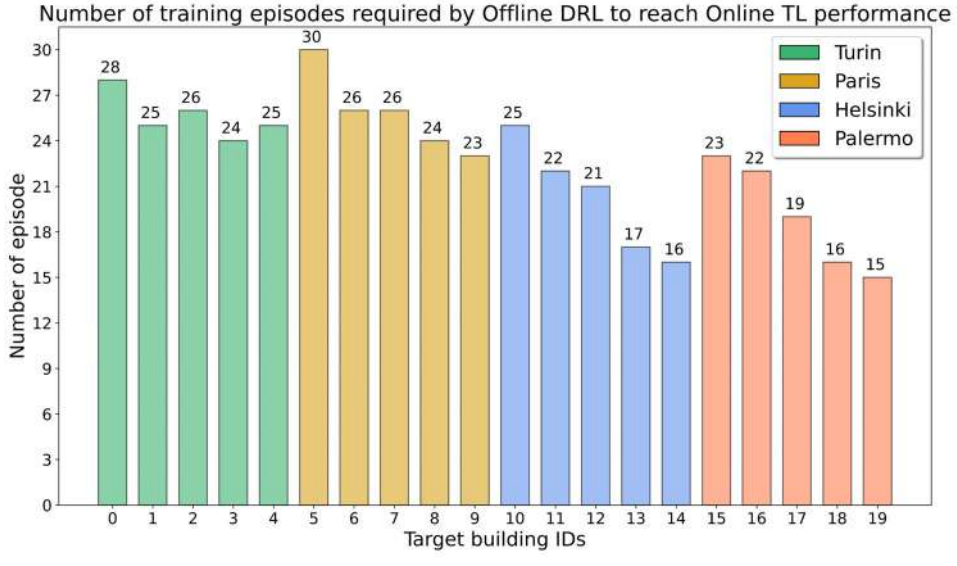

Fig. 12 Number of training episodes required by offline DRL to achieve the same performance level of online TL

validations requires comprehensive data collection from buildings, including sensor data related to energy systems, weather conditions, and building occupancy. Moreover, certain challenges need to be addressed to implement the developed online TL methodology in real buildings, such as ensuring compatibility with existing control systems and the need for continuous data acquisition. To conclude, the sizes and capacities of chillers, TES, PV, and BESS are critical factors influencing the results. In practical applications, the choice of equipment sizes will significantly impact the performance of the control strategy. While acknowledging the importance of the influence of the sizes of energy components on building energy performance in practical contexts, the primary goal of this study was to assess the effectiveness of online TL in scenarios where the energy systems are different between source and target buildings. However, a sensitivity analysis to assess the effects of different sizes and capacities of chillers, TES, PV, and BESS could be carried out in future works to provide valuable insights into the optimal configurations for specific building scenarios.

6 Conclusion

The present paper aimed to evaluate the effectiveness of an online transfer learning strategy applied to a DRL-based control agent in heterogeneous settings. The DRL controller, based on discrete SAC algorithm, was pre-trained on a source office building equipped with a chiller and a thermal cold water storage. The objective of the controller is to reduce the cost of electricity drawn from the grid while enhancing indoor air temperature control by choosing the operating mode of the energy system and the amount of energy to be provided to the building. The hyperparameters of the source DRL control agent were optimised through an automated procedure based on the Python library Optuna.

Several transfer experiments were performed to assess the performance of transferring the pre-trained controller to multiple target buildings. The target buildings have the same geometry as the source building but implemented a PV system and a BESS, managed by a RBC_{el}, in addition to chiller and TES managed by the transferred DRL controller. Two different settings of heterogeneous TL were evaluated: transductive and inductive. In the first scenario, weather conditions, price and occupancy schedules, and thermophysical properties of the building envelope were modified compared to the source building. In the second scenario, the effectiveness of online TL was assessed when the objective function was modified to include peak shaving.

The online TL framework introduced in this study integrates three key techniques: model slicing, imitation learning and weight-initialisation. Imitation learning is used to initialise the memory buffer of the target controller with transitions collected during a one-week RBC implementation. Weight-initialisation is applied to initialise the control policy of the target agent employing pre-trained weights from the DRL source control policy. The model slicing technique is integrated into the weight-initialisation process to allow pre-trained weights from the source controller could be employed to initialise the weights of the neural networks approximating the control policy for the online TL strategy.

The performance of the online TL strategy was compared with that of three different strategies: RBC, offline DRL and online DRL. In heterogeneous transductive setting, the online TL strategy performed on average worse than the offline DRL controller in all analysed target buildings in terms of electricity cost (i.e., 37%), SS (i.e., 24%), SC (i.e., 19%) and indoor temperature control (i.e., the mean value of the daily average temperature violation rate was 40% higher than offline DRL).

Meanwhile, online TL was more effective than RBC and online DRL respectively when evaluating average savings on electricity cost (i.e., −12% and −11%) and mean value of the daily average temperature violation rate (i.e., −10% and −41%). Moreover, online TL allowed to increase SS by 9% and SC by 11% when compared to RBC. In the heterogeneous inductive scenario, the online TL controller achieved a near-optimal control policy capable of limiting peak shaving violations beyond the 2 kW threshold of power absorption from the grid in two cooling seasons. In this scenario, the online TL agent enhanced indoor temperature conditions by 82% compared to offline DRL, but performed worse by 50% in terms of electricity cost and peak shaving (i.e., 12 versus 2 peak limit violations). While the online transfer learning methodology generally performed worse compared to offline DRL, it offers the advantage of enhancing the scalability and generalisability of advanced controllers in buildings since it does not require the definition of a building surrogate model for training purposes as per offline DRL controllers.

Future works are expected to cover the following directions:

- Evaluating the performance of the proposed method in comparison to other advanced control strategies, such as MPC, to offer a more comprehensive assessment of the advantages it offers when deployed.
- Implementing safety guards to ensure safe operating conditions (e.g., reverting to the baseline control or implementing a fail-safe strategy) to ensure that indoor temperature is maintained within the allowed temperature range.
- Enhancing the building simulation by utilising Spawn of EnergyPlus (Wetter et al. 2023) to create a more detailed simulation environment. Spawn allows for the integration

of the energy system modelled in Modelica with the building energy model developed in EnergyPlus. Consequently, the results obtained through the proposed approach are close to those observed in real-world building operations.
- Developing building archetypes to discover when TL is effective between source and target buildings of different types, avoiding negative TL. This process could exploit the potentialities of TL since building managers and operators can pre-determine, given a target building, the best source building from which to transfer a DRL-based control policy.
- Developing an infrastructure to explore the implementation of the online TL methodology in a real-world testbed.

Appendix

A1 Transfer learning fundamentals and applications for reinforcement learning

Transfer Learning is a machine learning approach aimed at leveraging previously acquired knowledge in one task to enhance performance in a different yet related problem (Pinto et al. 2022c). This method involves sharing knowledge at the initial stages of the learning process, which expedites the convergence of machine learning models compared to starting from scratch without any prior knowledge. As stated in Pan and Yang (2010) and Pinto et al. (2022c), TL enhances the learning of a predictive function in the target domain D_T, associated with task T_T, by leveraging the knowledge gained from the source domain D_S with task T_S. To mathematically define TL, it is necessary to consider the concepts of domain and task, as outlined by Pan and Yang (2010). The domain is composed of a feature space X and a marginal distribution probability $P(X)$, while the task encompasses the label space Y and an objective predictive function $f(\cdot)$. This function is learned from the training data, represented as pairs (x_i,y_i), and it is employed to approximate the conditional probability $P(y|x)$ and make predictions for new instances.

Since this paper evaluates the knowledge sharing between DRL-based controllers, it is necessary to establish a correspondence between the domain, label space, and task defined in generic ML problems and the state-space, action-space, and reward function in RL. In the context of RL, several studies such as Taylor and Stone (2009) and Da Silva and Costa (2019) provide insights for potential applications of TL for this control algorithm. Specifically, in RL the input feature space (i.e., domain) corresponds to the state-space, while the label space aligns with the action-space. In general, for ML problems knowledge sharing can occur in scenarios where the source and target domains, tasks, and solutions may be different or similar. In this framework, Pinto et al. (2022c) identified different classifications according to the similarity of tasks (i.e., label classification), features and labels (i.e., space classification), and modalities of knowledge sharing (i.e., solution classification). For clarity and conciseness, this section describes extensively only the elements included in each classification discussed below, while the others are provided useful references in the literature.

Three categories are defined for classifying TL approaches based on task similarity and the availability of labelled data in the source and target domains.
- **Inductive Transfer Learning**, where labelled data is available in both the source and target domains, and the source and target tasks are different. The focus is not on domain differences but rather on leveraging labelled data from the source domain to improve learning in the target domain.
- **Transductive Transfer Learning**, where, the source and target domains are different, but they share the same task. However, labelled data is only available for the source domain. The objective is to employ the labelled data from the source domain to enhance learning in the target domain.
- **Unsupervised Transfer Learning**, where labelled data is not available in either the source or target domains. The domains may be equal or not, and the tasks in the source and target domains are different. The aim is to leverage the shared information or structure between the domains to enhance learning in the target domain.

A further classification is defined according to the differences in source and target features (i.e., spaces) and labels:
- **Homogeneous Transfer Learning**, considering applications where the source and target spaces, as well as labels, are identical. In this case, there are no differences in the feature spaces or labels between the source and target domains.
- **Heterogeneous Transfer Learning**, considering applications where there are differences in the feature spaces and/or labels between the source and target domains. In this case, the feature spaces or labels (or both) vary between the source and target domains.

Additionally, TL is classified based on the knowledge-sharing method employed in solution classification: instance-based, feature representation-based, relation knowledge-based and model parameter-based transfer learning (Pinto et al. 2022c). This work implements model parameter-based TL, focused on sharing certain parameters or their distributions between the source and target tasks, such as model weights. Furthermore, model parameter-based TL includes three sub-classifications based on the methods of model parameter sharing: feature-extraction, weight-initialisation and relational knowledge-based (Pinto et al.

2022)c. In this paper, weight-initialisation is employed since the target model weights are initialised by means of the pre-trained model weights from the source task. This initialisation provides a starting point for the target model to benefit from the knowledge learned in the source domain. After the weights are initialised, an additional fine-tuning process can be performed. During fine-tuning, the target model is further trained using the data specific to the target task to refine its parameters based on the characteristics of the target domain.

An additional classification for TL can be established based on the differences associated with the source of knowledge, its availability, and the required domain knowledge (Da Silva and Costa 2019; Coraci et al. 2023b): intra-agent and inter-agent transfer learning. This paper evaluates the implementation of **Intra-Agent Transfer Learning**, which encompasses transfer methods that do not require explicit communication or direct interaction between agents to access internal knowledge. In this case, the transfer of knowledge occurs within a single agent, specifically the source agent, without the target agent being aware of the future implications of the new training process for the source agent. In Intra-Agent TL, the knowledge acquired by the source agent up to a certain point, regardless of whether the training process has been completed or not, is transferred to the target agent. This knowledge transfer can include learned representations, model parameters, or other internal knowledge that the source agent has accumulated during its training.

To conclude, Da Silva and Costa (2019) and Zhu et al. (2020) indicate three possible settings of knowledge reuse in addition to transfer learning, such as:

- **Imitation Learning (IL)**, where the target control agent learns an optimal strategy for a specific task by observing the behaviour of an expert. The expert, such as a RBC, optimises the same task and serves as a source of knowledge for the target agent. During the IL process, the target agent has access to the transitions generated by the expert. These transitions represent the actions from the expert and the resulting state transitions. The target agent can store these transitions in a buffer for later use.
- **Inverse Reinforcement Learning (IRL)**, where expert demonstrations of a specific task are observed and employed to extract an underlying reward function, used to train a RL agent straightforwardly, aiming to optimise performance in the same task. This approach represents a form of indirect imitation, leveraging the underlying intentions and objectives of the expert instead of direct mapping from states to actions as seen in behavioural cloning.
- **Learning from Demonstration (LfD)**, where the target control agent learns from the demonstrations of the expert by observing their actions and the resulting state transitions. However, unlike in pure IL, the expert controller in LfD can inform the target agent about the chosen set of actions. Additionally, in LfD, the target agent may have access to reward signals associated with the actions selected by the expert.

A2 Details on envelope features of target buildings and hyperparameters of source DRL controller

Table A1 reports the five envelope configurations for the target buildings.

Table A1 Opaque and transparent envelope features for target buildings

Efficiency configuration	U_{OP} [W/(m²·K)]	χ_i [kJ/(m²·K)]	U_{TR}[W/(m²·K)]	Solar factor g
0	0.16	38.9	0.5	0.49
1	0.3	43.5	1.3	0.35
2	0.34	44.2	1.9	0.65
3	0.18	40.0	1	0.5
4	0.45	48.0	2.5	0.35

The envelope efficiency configuration labelled as "0" corresponds to the reference building used as source. The remaining configurations are defined based on the building standards of each specific location. The thermophysical feature values for the reference buildings in Turin (i.e., configuration 1), Paris (i.e., configuration 2), Helsinki (i.e., configuration 3), and Palermo (i.e., configuration 4) were chosen according to Ministry of Economic Development (2015a, 2015b); Bienvenido-Huertas et al. (2019) and Huynh et al. (2021).

Table A2 reports the search domain and the best values of DRL hyperparameters for the source controller.

Table A2 Search domain and best values of DRL hyperparameters for the source controller

Hyperparameter	Search domain	Best value
# Hidden layers	[2, 4]	2
# Neurons per layer	[64, 128, 256]	64
Discount factor γ	[0.9, 0.95, 0.99]	0.99
Actor/critic learning rate μ	[0.0005, 0.001, ..., 0.005]	0.001
Reward electricity cost-term weight factor δ	[1, 2, ... 20]	8
Reward temperature-term weight factor β	[0.01, 0.015, ..., 0.1]	0.045

Acknowledgements

The work of Silvio Brandi and Alfonso Capozzoli is funded by the project NODES which has received funding from the

MUR – M4C2 1.5 of PNRR funded by the European Union – NextGenerationEU (Grant agreement no. ECS00000036).

Funding note: Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. Tianzhen Hong and Alfonso Capozzoli are Editorial Board members of *Building Simulation*.

Author contribution statement

Davide Coraci: conceptualization, methodology, software, investigation, formal analysis, data curation, writing – original draft, visualization; Silvio Brandi: conceptualization, methodology, investigation, writing – review & editing; Tianzhen Hong: methodology, validation, writing – review & editing; Alfonso Capozzoli: conceptualization, methodology, validation, writing – review & editing, supervision.

References

Akiba T, Sano S, Yanase T, et al. (2019). Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Amato A, Bilardo M, Fabrizio E, et al. (2021). Energy evaluation of a PV-based test facility for assessing future self-sufficient buildings. *Energies*, 14: 329.

Anvari-Moghaddam A, Rahimi-Kian A, Mirian MS, et al. (2017). A multi-agent based energy management solution for integrated buildings and microgrid system. *Applied Energy*, 203: 41–56.

ARERA (2022). Arera - andamento del prezzo dell'energia elettrica per il consumatore domestico tipo in maggior tutela. Available at https://www.arera.it/it/dati/eep35.htm. Accessed 23 Aug 2022. (in Italian)

ASHRAE (2021). High performance sequences of operation for HVAC systems. Atlanta, GA, USA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Austin Energy (2023). Electricity Tariff Pilot Programs. Available https://austinenergy.com/ae. Accessed 23 Aug 2022.

Bellman R (1966). Dynamic programming. *Science*, 153: 34–37.

Bienvenido-Huertas D, Oliveira M, Rubio-Bellido C, et al. (2019). A comparative analysis of the international regulation of thermal properties in building envelope. *Sustainability*, 11: 5574.

Brandi S, Fiorentini M, Capozzoli A (2022a). Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management. *Automation in Construction*, 135: 104128.

Brandi S, Gallo A, Capozzoli A (2022b). A predictive and adaptive control strategy to optimize the management of integrated energy systems in buildings. *Energy Reports*, 8: 1550–1567.

Brandi S, Piscitelli MS, Martellacci M, et al. (2020). Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy and Buildings*, 224: 110225.

Brockman G, Cheung V, Pettersson L, et al. (2016). Openai Gym.

Chiang Y-T, Lu C-H, Hsu JY-J (2017). A feature-based knowledge transfer framework for cross-environment activity recognition toward smart home applications. *IEEE Transactions on Human-Machine Systems*, 47: 310–322.

Christodoulou P (2019). Soft actor-critic for discrete action settings. arXiv:1910.07207.

Coraci D, Brandi S, Piscitelli MS, et al. (2021). Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings. *Energies*, 14: 997.

Coraci D, Brandi S, Capozzoli A (2023a). Effective pre-training of a deep reinforcement learning agent by means of long short-term memory models for thermal energy management in buildings. *Energy Conversion and Management*, 291: 117303.

Coraci D, Brandi S, Hong T, et al. (2023b). Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. *Applied Energy*, 333: 120598.

Crawley DB, Lawrie LK, Winkelmann FC, et al. (2001). EnergyPlus: Creating a new-generation building energy simulation program. *Energy and Buildings*, 33: 319–331.

Da Silva FL, Costa AHR (2019). A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research*, 64: 645–703.

Deltetto D, Coraci D, Pinto G, et al. (2021). Exploring the potentialities of deep reinforcement learning for incentive-based demand response in a cluster of small commercial buildings. *Energies*, 14: 2933.

Dey S, Marzullo T, Henze G (2023a). Inverse reinforcement learning control for building energy management. *Energy and Buildings*, 286: 112941.

Dey S, Marzullo T, Zhang X, et al. (2023b). Reinforcement learning building control approach harnessing imitation learning. *Energy and AI*, 14: 100255.

Durisch W, Bitnar B, Mayor J-C, et al. (2007). Efficiency model for photovoltaic modules and demonstration of its application to energy yield estimation. *Solar Energy Materials and Solar Cells*, 91: 79–84.

Elehwany H, Ouf M, Gunay B, et al. (2024). A reinforcement learning approach for thermostat setpoint preference learning. *Building Simulation*, 17: 131–146.

Esrafilian-Najafabadi M, Haghighat F (2023). Transfer learning for occupancy-based HVAC control: A data-driven approach using unsupervised learning of occupancy profiles and deep reinforcement learning. *Energy and Buildings*, 300: 113637.

European Commission (2019). European Green Deal.

Fang X, Gong G, Li G, et al. (2023). Cross temporal-spatial transferability investigation of deep reinforcement learning control strategy in the building HVAC system level. *Energy*, 263: 125679.

Finck C, Beagon P, Clauß J, et al. (2018). Review of applied and tested control possibilities for energy flexibility in buildings—A technical report from IEA EBC Annex 67 Energy Flexible Buildings.

Fulpagare Y, Huang K-R, Liao Y-H, et al. (2022). Optimal energy management for air cooled server fans using deep reinforcement learning control method. *Energy and Buildings*, 277: 112542.

Grubinger T, Chasparis GC, Natschläger T (2017). Generalized online transfer learning for climate control in residential buildings. *Energy and Buildings*, 139: 63–71.

Haarnoja T, Zhou A, Hartikainen K, et al. (2019). Soft actor-critic algorithms and applications. arXiv: 1812.05905.

Holmgren WF, Hansen CW, Mikofski MA (2018). Pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software*, 3: 884.

Huynh A, Dias Barkokebas R, Al-Hussein M, et al. (2021). Energy-efficiency requirements for residential building envelopes in cold-climate regions. *Atmosphere*, 12: 405.

IEA (2019). World Energy Outlook 2019. Available at https://www.iea.org/reports/world-energy-outlook-201. Accessed 14 Sept 2023.

Jacobson MZ, Jadhav V (2018). World estimates of PV optimal tilt angles and ratios of sunlight incident upon tilted and tracked PV panels relative to horizontal panels. *Solar Energy*, 169: 55–66.

Kaya M, Bilge H (2019). Deep metric learning: A survey. *Symmetry*, 11: 1066.

Li H, Chaudhari P, Yang H, et al. (2020). Rethinking the hyperparameters for fine-tuning. arXiv: 2002.11770.

Li A, Xiao F, Fan C, et al. (2021). Development of an ANN-based building energy model for information-poor buildings using transfer learning. *Building Simulation*, 14: 89–101.

Li G, Chen L, Liu J, et al. (2023). Comparative study on deep transfer learning strategies for cross-system and cross-operation-condition building energy systems fault diagnosis. *Energy*, 263: 125943.

Li J, Zhang C, Zhao Y, et al. (2022). Federated learning-based short-term building energy consumption prediction method for solving the data silos problem. *Building Simulation*, 15: 1145–1159.

Lissa P, Schukat M, Keane M, et al. (2021). Transfer learning applied to DRL-Based heat pump control to leverage microgrid energy efficiency. *Smart Energy*, 3: 100044.

Ministry of Economic Development (2015a). Interministerial Decree of 26 June 2015. Appendix A. Available at https://www.mise.gov.it/index.php/it/normativa/decretiinterministeriali/decreto-interministeriale-26-giugno-2015applicazione-delle-metodologie-di-calcolo-delle-prestazionienergetiche-e-definizione-delle-prescrizioni-e-dei-requisiti-minimi-degli-difici?cldee=ZW5lcmdpYS5kZW1hcmNvQGxpYmVyby5pdA%3D%3D&urlid=0?hitcount=0. Accessed 23 Aug 2022.

Ministry of Economic Development (2015b). Interministerial Decree of 26 June 2015. Appendix b. Available at https://www.mise.gov.it/index.php/it/normativa/decretiinterministeriali/decreto-interministeriale-26-giugno-2015applicazione-delle-metodologie-di-calcolo-delle-prestazionienergetiche-e-definizione-delle-prescrizioni-e-dei-requisiti-minimi-degli-edifici?cldee=ZW5lcmdpYS5kZW1hcmNvQGxpYmVyby5pdA%3D%3D&urlid=0?hitcount=0. Accessed 23 Aug 2022.

Mnih V, Kavukcuoglu K, Silver D, et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518: 529–533.

Modelica Association (2000). Modelica® – A unified object-oriented language for physical systems modeling. Tutorial (1.4 ed.). Available at http://www.modelica.org/documents/ModelicaTutorial14.pdf.

Mosaico G, Saviozzi M, Silvestro F, et al. (2019). Simplified state space building energy model and transfer learning based occupancy estimation for HVAC optimal control. In: Proceedings of IEEE 5th International Forum on Research and Technology for Society and Industry (RTSI), Florence, Italy.

Nagy Z, Henze G, Dey S, et al. (2023). Ten questions concerning reinforcement learning for building energy management. *Building and Environment*, 241: 110435.

Nweye K, Sankaranarayanan S, Nagy Z (2023). MERLIN: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities. *Applied Energy*, 346: 121323.

Pan SJ, Yang Q (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22: 1345–1359.

Pinto G, Deltetto D, Capozzoli A (2021). Data-driven district energy management with surrogate models and deep reinforcement learning. *Applied Energy*, 304: 117642.

Pinto G, Kathirgamanathan A, Mangina E, et al. (2022a). Enhancing energy management in grid-interactive buildings: a comparison among cooperative and coordinated architectures. *Applied Energy*, 310: 118497.

Pinto G, Messina R, Li H, et al. (2022b). Sharing is caring: An extensive analysis of parameter-based transfer learning for the prediction of building thermal dynamics. *Energy and Buildings*, 276: 112530.

Pinto G, Wang Z, Roy A, et al. (2022c). Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives. *Advances in Applied Energy*, 5: 100084.

Piscitelli MS, Brandi S, Capozzoli A, et al. (2021). A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. *Building Simulation*, 14: 131–147.

PVSites (2016). European climate zones and bio-climatic design requirements. Available at https://www.pvsites.eu/downloads/category/project-results? page=4. Accessed 23 Aug 2022.

Ruusu R, Cao S, Manrique Delgado B, et al. (2019). Direct quantification of multiple-source energy flexibility in a residential building using a new model predictive high-level controller. *Energy Conversion and Management*, 180: 1109–1128.

Salsbury TI (2005). A survey of control technologies in the building automation industry. *IFAC Proceedings Volumes*, 38: 90–100.

Smith SL, Kindermans PJ, Ying C, et al. (2017). Don't decay the learning rate, increase the batch size. arXiv: 1711.00489.

Sutton RS, Barto AG (2018). Reinforcement Learning: An Introduction, 2 edn. Cambridge, MA, USA: MIT Press.

Taylor ME, Stone P (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10: 1633–1685.

Tsikaloudaki K, Laskos K, Bikas D (2012). On the establishment of climatic zones in Europe with regard to the energy performance of buildings. *Energies*, 5: 32–44.

Vázquez-Canteli JR, Ulyanin S, Kämpf J, et al. (2019). Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. *Sustainable Cities and Society*, 45: 243–257.

Vázquez-Canteli JR, Dey S, Henze G, et al. (2020). CityLearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management. arXiv:2012.10504.

Wang L, Geng X, Ma X, et al. (2019). Ridesharing car detection by transfer learning. *Artificial Intelligence*, 273: 1–18.

Wang D, Zheng W, Wang Z, et al. (2023a). Comparison of reinforcement learning and model predictive control for building energy system optimization. *Applied Thermal Engineering*, 228: 120430.

Wang X, Kang X, An J, et al. (2023b). Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. *Energy and Buildings*, 301: 113696.

Wei Z, Calautit J (2023). Evaluation of model predictive control (MPC) of solar thermal heating system with thermal energy storage for buildings with highly variable occupancy levels. *Building Simulation*, 16: 1915–1931.

Wetter M, Benne K, Tummescheit H, et al. (2023). Spawn: coupling Modelica Buildings Library and EnergyPlus to enable new energy system and control applications. *Journal of Building Performance Simulation*, https://doi.org/10.1080/19401493.2023.2266414.

Xiong Q, Li Z, Cai W, et al. (2023). Model free optimization of building cooling water systems with refined action space. *Building Simulation*, 16: 615–627.

Yang L, Nagy Z, Goffin P, et al. (2015). Reinforcement learning for optimal control of low exergy buildings. *Applied Energy*, 156: 577–586.

Zelinka J, Prágr M, Szadkowski R, et al. (2022). Traversability transfer learning between robots with different cost assessment policies. In: Proceedings of International Conference on Modelling and Simulation for Autonomous Systems.

Zhang Z, Chong A, Pan Y, et al. (2019). Whole building energy model for HVAC optimal control: a practical framework based on deep reinforcement learning. *Energy and Buildings*, 199: 472–490.

Zhang T, Aakash Krishna GS, Afshari M, et al. (2022a). Diversity for transfer in learning-based control of buildings. In: Proceedings of the 13th ACM International Conference on Future Energy Systems.

Zhang Z, Li Y, Wang J, et al. (2022b). ReMoS: Reducing defect inheritance in transfer learning via relevant model slicing. In: Proceedings of IEEE/ACM 44th International Conference on Software Engineering (ICSE), Pittsburgh, PA, USA.

Zhu Z, Lin K, Jain AK, et al. (2020). Transfer learning in deep reinforcement learning: A survey. arXiv: 2009.07888.

Zou Z, Yu X, Ergan S (2020). Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. *Building and Environment*, 168: 106535.