Deep reinforcement learning control architectures for industrial multi-energy systems: from single-agent to hierarchical multi-agents

Energy Conversion and Management 350 (2026) 120963

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Deep reinforcement learning control architectures for industrial multi-energy systems: from single-agent to hierarchical multi-agents

Andrea Franzoso, Gabriele Fambri *, Marco Badami

Department of Energy, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Keywords:
Deep reinforcement learning
Industrial energy decarbonization
Hierarchical control
Energy management system
Industrial multi-energy systems

ABSTRACT

The increasing share of renewable energy sources in industrial, multi-energy systems has introduced significant challenges for the optimal coordination of multiple energy carriers. This work focuses on improving the cost-effective and robust operation of an industrial plant that simultaneously produces electricity, steam, hot water, and chilled water. It explores whether data-driven control strategies based on deep reinforcement learning can enhance the economic and energy performance of the plant, compared to traditional methods, and whether multi-agent structures can further improve robustness. Three control architectures were designed and evaluated – centralized, decentralized, and hierarchical – using a year of actual operational data from the industrial plant. The results show that all of the proposed strategies reduced total operating costs by more than 6% compared to existing rule-based control. The hierarchical configuration achieved the best performance and demonstrated superior robustness to variations in energy prices. These findings highlight the potential of learning-based hierarchical coordination as a practical and resilient framework to manage complex industrial energy systems.

1. Introduction

Industrial multi-energy systems (MES) typically combine electricity, heat, and cooling prosduction to fulfill various process requirements, while reducing costs and emissions [1]. The strategic objectives of global climate policy include the decarbonization of industry; in line with these targets, the share of renewable energy sources (RES) in industrial energy consumption is already increasing and is expected to grow markedly over the next decades [2]. The growing penetration of RES has made the design and operation of industrial energy systems increasingly complex [3]. Recent research shows how RES can be effectively introduced into industrial MES. Analyzing renewable configurations for industrial facilities at the system design level reveals the cost and emissions trade-offs that arise with increased photovoltaic and wind penetration [4]. Operational scheduling for industrial energy systems then demonstrates how peak-load management and coordinated scheduling can improve the use of on-site RES and reduce reliance on the grid [5]. More detailed flexibility modelling shows that industrial systems can adjust their processes to better accommodate RES variability, providing a structured way to support fluctuating generation [6]. The joint scheduling of all energy uses in a cluster further shows that conversion and storage technologies play a central role in smoothing renewable output and enabling cost-efficient operation [7].

In this context, having an Energy Management System (EMS) that can operate efficiently, react to changing conditions, and reliably control and coordinate all energy conversion devices is of paramount importance. EMS control approaches range from rule-based control (RBC) to mathematical optimization models (such as Linear Programming – LP or Mixed-Integer Linear Programming – MILP), or to Artificial Intelligence (AI)-based techniques [8]. Although RBC systems are

Abbreviations: **AB**, after burner; **AI**, artificial intelligence; **CHP**, combined heat and power; **COP**, coefficient of performance; **DDPG**, deep deterministic policy gradient; **DNN**, deep neural networks; **DQN**, deep Q network; **DRL**, deep reinforcement learning; **EMS**, energy management system; **EV**, electric vehicle; **HVAC**, heating, ventilation and air conditioning; **ICE**, internal combustion engine; **KPI**, key performance indicator; **LMPC**, linear model predictive control; **LP**, linear programming; **LSTM**, long short term memory; **MADDPG**, multi-agent DDPG; **MADRL**, multi-agent DRL; **MATD3**, multi-agent TD3; **MES**, multi-energy system; **MILP**, mixed integer linear programming; **MTs**, microturbines; **OU**, ornstein–uhlenbeck noise model; **PER**, prioritized experience replay; **PLC**, programmable logic controller; **PPO**, proximal policy optimization; **PSO**, particle swarm optimization; **PV**, photovoltaic system; **RBC**, rule-based controller; **RES**, renewable energy sources; **RL**, reinforcement learning; **SAC**, soft actor critic; **SADRL**, single-agent DRL; **SG**, steam generator; **TD3**, twin delayed DDPG; **VDN**, value decomposition network; **VPP**, virtual power plant.

* Corresponding author.
E-mail address: gabriele.fambri@polito.it (G. Fambri).

https://doi.org/10.1016/j.enconman.2025.120963
Received 17 October 2025; Received in revised form 1 December 2025; Accepted 14 December 2025
Available online 31 December 2025

simple, they are also inflexible, and their accuracy depends on the developer's skill in defining optimal control rules [9]. On the other hand, mathematical optimization methods can achieve near-optimal results, but they depend on accurate forecasts and simplified models, thereby limiting their scalability and real-time applicability in large, dynamic systems [10]. Meta-heuristic methods (e.g., genetic algorithms, particle swarm, simulated annealing) offer flexibility, but they lack guarantees of global optimality, require tuning, and scale poorly with several variables [10].

Recent research has turned to Deep Reinforcement Learning (DRL) to solve the combined problem of the control and optimization of energy systems [11]: DRL extends traditional Reinforcement Learning (RL) by using Deep Neural Networks (DNN) to handle large or continuous state spaces [12]. In DRL algorithms, agents learn optimal policies by interacting with the environment through trial-and-error, and they are guided by rewards. DRL, which uses DNN to approximate value functions or policies, can scale to high-dimensional state and action spaces, thereby enabling the adaptive, data-driven control of complex real-world scenarios. Algorithms, such as Proximal Policy Optimization (PPO) [13], Deep Q-Networks (DQN) [14], Deep Deterministic Policy Gradient (DDPG) [15], and Soft Actor-Critic (SAC) [16] support efficient real-time decision-making, and they often outperform traditional optimization methods in speed and accuracy [17].

DRL control can be structured in different ways, depending on how the decision-making is distributed. In the centralized single-agent DRL (SADRL) approach, one agent controls all the components and has full visibility of the system. In this setup, the centralized agent can coordinate all components and target a globally optimal strategy; however, its state–action space grows rapidly with system size, making training increasingly difficult and reducing scalability [18]. Decentralized multi-agent DRL (MADRL) configurations divide the control task among independent agents, and this leads to improved flexibility and resilience, but it also often leads to conflicts and suboptimal coordination [18]. A hierarchical multi-agent architecture combines both principles: a higher-level agent provides global guidance, while lower-level agents handle local decisions, and this offers better scalability and coordination for complex systems. This structure preserves scalability and adaptability, while introducing an additional layer of coordination complexity, with the aim of combining the strengths of both centralized and decentralized paradigms [19].

The centralized configuration is the most commonly adopted one, and practical energy management applications are popular at the building level, in particular to control HVAC (Heating, Ventilation and Air Conditioning) [20], water heating [21], and lightning systems [22], or even all of them together in a centralized manner [23] to optimize costs and comfort in both residential and commercial buildings. Research on EVs (Electric Vehicles) charging stations has also proliferated, with the aim of optimizing the charging schedules, minimizing costs, and ensuring grid stability; many works have been developed concerning the scheduling and management of the stochastic nature of this problem using different algorithms [24]. A few studies have addressed economic energy dispatch problems, driven by the increased use of RES and electricity storage in microgrids [25] or in MES that interlink various energy vectors and conversion devices, to achieve more holistic management optimization [26]. Alabi et al. [27] showed the improvement brought about by the utilization of DRL on the management of complex energy systems, compared to simple RBC strategies. Ceusters et al. [28] and Bousnina et al. [29] compared DRL algorithms with Linear Model Predictive Control (LMPC) for MES management purposes in the case of high-RES penetration, and they showed that DRL agents can successfully learn optimal strategies. Furthermore, Ruan et al. [30] compared the performance of DRL (DDPG and Twin Delayed DDPG – TD3) with other optimization methods (MILP and Particle Swarm Optimization – PSO), and they found better performances for DRL than for PSO, and that DRL was faster than both MILP and PSO. However, only a few examples have been reported in literature regarding industrial energy systems. One of these is that of Ghione et al. [31] whose aim was to optimize the hourly dispatch of a real cogeneration unit on the basis of the energy demands of an industrial facility. Lu et al [32] instead considered both demand and production units.

The decentralized MADRL framework has gained significant attention in the context of microgrid and multi-energy systems, due to its conceptual similarity with the game theory and its ability to model distributed decision-making processes. In such applications, each agent usually represents an autonomous energy entity that interacts with others to achieve individual or collective objectives. For instance, MADRL agents often represent individual microgrids or prosumers that interact in a large network [33] or energy conversion devices that operate in a grid [34]. Zhang et al. [35] studied the interactions that occur between multiple renewable, energy-powered, multi-energy hubs by optimizing the operational and environmental costs. Qiu et al. [33] formulated the problem of a double auctioned market for multi-energy microgrids and solved it using an MADRL framework, considering multiple TD3 agents to lower the overall costs of the system. May et al. [36] designed a dynamic pricing system for community-based, peer-to-peer energy markets to improve coordination among prosumers, while reducing the total electricity costs. Similarly, Ye et al. [37] developed an MADRL model to optimize local electricity markets by focusing on energy trading and the provision of flexibility services: each agent represented a prosumer with flexible energy demands. Park and Moon [38] applied a similar approach to optimize the charging and discharging operations of EVs, in which each EV operated as an agent within a smart grid that included photovoltaic (PV) systems, storage, and real-time pricing. These approaches have also been applied to manage multi-zone HVAC systems, in which each agent represents a different zone of a building [39]. In the industrial field, Zhu et al. [40] proposed a decentralized control approach for an industrial park, where each agent managed a specific unit (cogeneration, boilers, storage, batteries) and was trained through a shared reward mechanism to foster cooperation.

Fewer works have been dedicated to the analysis of hierarchical architectures, where agents are organized at different decision levels: a higher-level leader agent makes system-wide decisions, while lower-level follower agents act on both the environment and according to the leader's commands. Such a structure enables scalable yet coordinated control. Wang et al. [41] applied a Stackelberg game–based MADRL framework for Virtual Power Plant (VPP) management purposes, where the VPP agent acted as a leader and the EV manager as a follower. Zhang et al. [42] adopted a similar approach to coordinate a CHP (Combined Heat and Power) plant operator with a heat-and-power aggregator.

Table 1 summarizes the literature that has dealt with the application of DRL to different energy systems. It clearly shows that only a few works have investigated DRL-based control through a hierarchical architecture and, at the same time, only a limited number of studies have addressed the optimization of industrial energy systems, even though industrial energy efficiency remains a topic of high current relevance for both environmental and economic reasons. Moreover, to the best of the authors' knowledge, no previous work has implemented a hierarchical, multi-agent DRL architecture for the operational optimization of an industrial multi-energy system.

This work analyzes a real industrial multi-energy system. This system provides a frozen-food production facility with electricity, steam, hot water, and chilled water. Using one year of hourly operational data from a dedicated measurement campaign, centralized, decentralized and hierarchical multi-agent DRL controllers are compared. These control strategies are evaluated against the company's current rule-based energy management system, and the potential economic savings are quantified.

Table 1
Classification of representative DRL applications for energy management. The studies are grouped according to the adopted final DRL control structure (Single-Agent, Multi-Agent, or Hierarchical) and by the application domain.

Configuration	Ref.	Application domain	DRL algorithm
SADRL	Stoffel et al. 2023 [20],	Building energy management	SAC
SADRL	Somer et al. 2017[21]	Building energy management	Q-learning
SADRL	Park et al. 2019 [22]	Building energy management	Q-learning
SADRL	Ding et al. 2019 [23]	Building energy management	DQN
SADRL	Chen et al. 2018 [43]	Building energy management	Q-learning
SADRL	Brandi et al. 2022 [44]	Building energy management	SAC
SADRL	Zhang et al. 2022 [45]	Building energy management	DDPG
SADRL	Tuchnitz et al. 2021 [46]	Electric mobility & charging	DQN
SADRL	Rossi et al. 2025 [24]	Electric mobility & charging	PPO-clip
SADRL	Mansour et al. 2025 [47]	Electric mobility & charging	SAC, PPO, DDPG, TD3
SADRL	Nakabi et al. 2021 [25]	Microgrid operation	DQN, PPO, A3C
SADRL	Ji et al. 2019 [48]	Microgrid operation	DQN
SADRL	Bao et al. 2023 [49]	Microgrid operation	PERSAC, SAC, DDPG, DQN
SADRL	Zhou et al. 2022 [26]	MES	PER + LSTM SAC
SADRL	Alabi et al. 2023 [27]	MES & CO_2 capture	SAC, TD3
SADRL	Ceusters et a 2021 [28]	MES	PPO, TD3
SADRL	Bousnina et al. 2024 [29]	MES	DDPG
SADRL	Ruan et al. 2023 [30]	MES	DDPG, TD3
SADRL	Franzoso et al. 2025 [50]	MES	DDPG, TD3
SADRL	Ghione et al. 2025 [31]	Industrial MES	SAC, DDPG, DQN
SADRL	Lu et al. 2024 [32]	Industrial MES	DDPG, DQN, PPO and DDPG + DQN
MADRL	Shen et al. 2022 [51]	Building energy management	DQN with VDN
MADRL	Bo et al. 2023 [39]	Building energy management	MA-TD3
MADRL	Park et al. 2022 [38]	Electric mobility & charging	MA-DDPG and COMA
MADRL	Qiu et al. 2023 [33]	Microgrid operation	Based on MATD3
MADRL	Safiri et al. 2023 [34]	Microgrid operation	Actor-Critic method
MADRL	May et al. 2023 [36]	Microgrid operation (markets)	MA-PPO
MADRL	Ye et al. 2023 [37]	Microgrid operation (markets)	MA-SAC
SADRL, MADRL	Zhang et al. 2022 [35]	MES	MADDPG
SADRL, MADRL	Zhu et al. 2022 [40]	Industrial MES	DDPG, SAC (and attention mechanism)
Hier. MADRL	Wang et al. 2022 [41]	EVs and VPP	SAC and TD3
Hier. MADRL	Pei et al. 2025 [52]	Building energy management –	SAC and PPO
Hier. MADRL	Gao et al. 2024 [53]	Electric mobility & charging	Q-mix
SADRL, MADRL, Hier.MADRL	**This work**	**Industrial MES**	**DDPG – SAC**

The main novelties of this work pertain to:

- A real industrial case and data-driven study: application of DRL-based optimization to a real industrial multi-energy system, using roughly one year of measured operational data from a dedicated monitoring campaign.
- A hierarchical MADRL architecture for an industrial MES: implementation and assessment of hierarchical leader–follower, multi-agent DRL controllers for a real industrial MES, which, to the best of our knowledge, has not yet been explored in the literature.
- A systematic comparison of DRL architectures: a head-to-head comparison of single-agent, decentralized, multi-agent and hierarchical, multi-agent, DRL controllers for the same industrial plant.
- A robustness and sensitivity analysis: a comprehensive robustness assessment of the proposed architectures, through sensitivity analyses, considering the electricity price, natural gas price, industrial energy demand, and the installed PV capacity.

2. Materials and methods

This section details the methodological framework that we have adopted in this study, which encompasses both the modeling of the industrial, multi-energy system and the design of the control strategies. First, the configuration and operational characteristics of the plant are introduced. Then, the mathematical formulation of the main energy conversion units is presented. Next, it describes the implemented control approaches, including the reference RBC, the optimization benchmark, and the DRL architectures investigated. Lastly, the setup of the training environment, the data preparation, and the performance evaluation criteria adopted are presented to ensure reproducibility and comparability of the results.

2.1. Energy demands

First, the energy consumption of the industrial facility under analysis and its trends had to be identified. The energy vectors that are involved are electricity, steam, hot water and chilled water. The energy vectors involved are electricity, steam, hot water, and chilled water. The energy system has no flexibility because the production of these vectors is directly linked to and constrained by the corresponding demands. The provided dataset covered the April 2023-March 2024 period and was composed of the hourly energy demand (electricity, steam, hot water, and chilled water), energy production, and fuel consumptions.

As can be seen in Fig. 1, the duration curve graph indicates that electricity consumption (blue line) begins near 6.4 MW: electricity is the

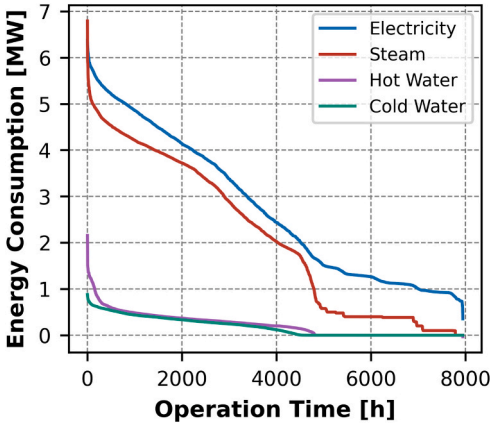

Fig. 1. Duration curves of the energy demands.

energy vector that is required the most, and it shows the dominant consumption with a significant base load. Steam consumption (dark orange line) begins at approximately 6.8 MW but declines more steeply in the second half of the period, ultimately settling at a much lower level. Hot water consumption (purple line) peaks at roughly 2.0 MW and then quickly drops, flattening out at a few hundred kW for much of the duration, thus indicating a moderate base consumption. Finally, chilled water (green line) has the lowest peak-under 1.0 MW and, after a rapid initial decline, remains at a comparatively small demand level.

The weekly patterns and seasonal variations can be seen in the different panels in Fig. 2, which highlights how the energy usage profiles vary throughout both the week and the year. The horizontal axis in each subfigure spans from Sunday to Saturday, and the colored lines represent the average consumption for winter, summer, spring, and fall. The electricity demand in Fig. 2a shows marked peaks on weekdays, thereby reflecting high factory activity levels, whereas weekends exhibit low loads, due to reduced operations. Notably, the highest electricity consumption line corresponds to summer (green line), due to increased cooling needs, while the lowest refers to winter (blue line). The spring (orange line) and fall (red line) curves lie in between, since cooling systems run less intensely in these periods. Steam consumption (Fig. 2b) shows a lower seasonal variability, thus suggesting that the process is independent of the environmental conditions. Instead, Fig. 2c shows that the hot water consumption is much lower than the steam consumption, and it is mostly constant throughout the seasons. A peak in the early hours of Monday is observable, and this is most likely connected to the restarting of the circuits after the weekend when consumption is null. Finally, Fig. 2d displays the chilled water consumption and highlights the diurnal and weekly cycles, as well as distinct seasonal patterns. As previously mentioned, the chilled water consumption considered here is for a specific process, but consumption is clearly higher during the summer.

Fig. 3 reports the monthly average energy costs over the period analyzed. The electricity purchase costs (shown in orange) are consistently the highest of the three considered vectors, and they fluctuate around 200–250 €/MWh, with a slight reduction in the summer months. The natural gas costs (in dark blue) are lower and more stable, ranging between 60 and 80 €/MWh. The electricity selling prices (light green) remain significantly below the purchase costs, and they average between 80 and 140 €/MWh.

2.2. Energy system

As previously stated, this study is based on a real industrial plant. As previously stated, the analyzed factory requires electricity, steam, hot water, and chilled water, all of which are supplied by a combination of on-site devices and external sources, including both renewable and conventional technologies. All the energy production and conversion units of the facility have been modelled in order to faithfully reproduce the internal energy flows. Specifically, the plant resorts to a gas-powered internal combustion engine (ICE), three microturbines (MTs) attached to an afterburner (AB), four steam generators (SG), and a PV system (see Fig. 4). All the units are installed locally and employed on-site, and the system can also buy or sell electricity to the power grid. The available

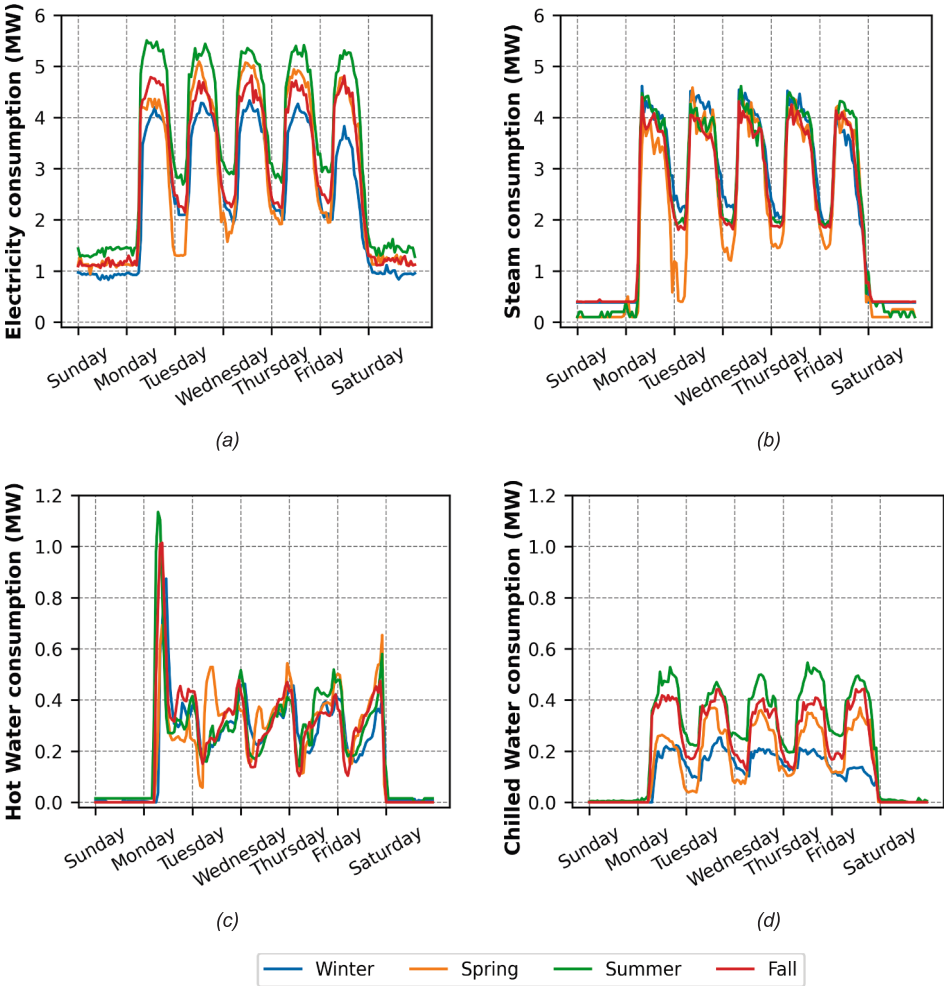

Fig. 2. Profiles of the mean seasonal energy consumption. Note that panels (a) and (b) use a different y-axis scale from panels (c) and (d).

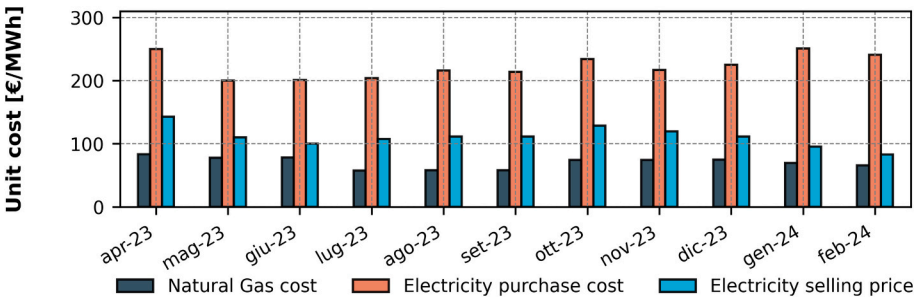

Fig. 3. Monthly trends of the natural gas costs, electricity purchase costs, and electricity selling prices.

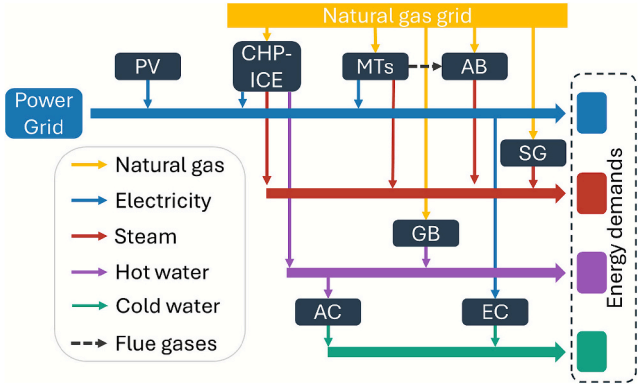

Fig. 4. Schematic representation of the industrial energy system showing the energy flows of the natural gas, electricity, steam, hot water, and chilled water between the generation units, and the corresponding energy demands.

dataset used for the modelling of the energy system includes the fuel consumption and energy output of each generation unit (electricity, steam, hot water, and chilled water).

The installed ICE generates electricity, steam, through the exhaust gas heat recovery system, and hot water from the engine jacket water heat recovery system, which satisfies the hot water demand and feeds the hot side of the absorption chiller. The MTs use natural gas to generate electricity and produce steam through the exhaust gas heat recovery system. The AB burns additional fuel to further heat the exhaust gases, thereby enabling a higher steam generation. Its output scales with the number of active turbines. A consistent amount of electricity is produced by the PV system, which provides renewable electricity to offset a portion of the electricity demand. The factory is connected to the main power grid to ensure a stable power supply; excess electricity generated on site can be sold back to the grid, thereby providing an additional revenue stream. Steam generators are required to meet the steam demand, since the production from the ICE, MTs, and AB is insufficient. Likewise, gas boilers are present for hot water generation purposes. Lastly, chilled water is required for some processes and to increase the exploitation of the ICE, while absorption chillers are employed to produce chilled water instead of using compression chillers, which would increase the electricity demand. The absorption chillers are similar to the compression chillers, but instead of using a prime mover to increase the pressure of the refrigerant vapors, they use a hot source – generally at a temperature of between 80 °C and 100 °C – to produce chilled water at 5–10 °C. The performance of an absorption chiller is generally expressed by considering its Coefficient of Performance (COP), which is calculated as the ratio between the useful thermal output (such as the cooling load) and the supplied heat energy. Table 2 reports the most relevant parameters available for the energy devices considered, while Table 3 summarizes the mathematical models adopted to describe their behaviors. These equations were obtained by fitting real operational data collected during a dedicated measurement campaign at the considered plant. Specifically, linear and piecewise-linear regressions were performed on the measured gas consumption and thermal/electrical outputs of each device to derive simplified, input–output relations that would be suitable to integrate into the DRL environment. Therefore, the numerical coefficients that appear in Table 3 (e.g., slopes, intercepts, and load thresholds) reflect the best-fit parameters derived from the experimental datasets, and they represent the empirical performance of the actual equipment under typical operating conditions. A detailed description of the data processing procedures and of the regression methodology, and validation of these models against the measured profiles are provided in the Supplementary Material.

Table 2
Components and nominal characteristics.

Component	Variable	UoM	Value
ICE-CHP	**Nominal Electric Power**	kW	4300
	Nominal Electrical Efficiency	–	0.44
	Steam Production	kW	1700
	Hot Water Production	kW	2500
	Load Factor Range	–	0.7–1
MT	**Number**	–	3
	Nominal Electric Power	kW	200
	Nominal Electrical Efficiency	–	0.31
	Steam Production	kW	200
	Load Factor Range	–	0.75–1
PV	**Installed capacity**	MW_p	2.2
Steam generator	**Number**	–	4
	Nominal Power	kW	2000
	Load Factor Range	–	0–1
Absorption chiller	**Nominal Power**	kW	900
	COP	–	0.62

Table 3
Mathematical description of the components involved in the energy system. The formulas are expressed in MW.

Device	Quantity	Formula	
ICE	Gas cons.	$\begin{cases} \dot{m}_b \bullet LHV = \alpha \bullet P_{el,nom} \bullet 2.2 + 0.597\alpha \geq 0.7 \\ \dot{m}_b \bullet LHV = 0\alpha < 0.7 \end{cases}$	(1)
	Steam	$Steam = P_{el} \bullet 0.25 + 0.65$	(2)
	Hot water	$HotWater = P_{el} \bullet 0.6$	(3)
MTs	Gas cons.	$\begin{cases} \dot{m}_b \bullet LHV = \alpha \bullet P_{el,nom} \bullet 2.0 + 0.237\alpha \geq 0.75 \\ \dot{m}_b \bullet LHV = 0\alpha < 0.75 \end{cases}$	(4)
	Steam	$Steam = P_{el} \bullet 1$	(5)
After-burner	Steam	$Steam = \beta \bullet N_{MTs,ON} \bullet 0.417$	(6)
	Gas cons.	$\dot{m}_b \bullet LHV = \frac{Steam}{1.1}$	(7)
Steam gen.	Gas cons.	$\dot{m}_b \bullet LHV = \alpha \bullet P_{nom} \bullet 1.1 + 0.038$	(8)
Gas boilers	Gas cons.	$\dot{m}_b \bullet LHV = \frac{\alpha \bullet P_{nom}}{0.9}$	(9)
Absorption chillers	Chilled water	$Chilledwater = HotWater \bullet 0.62$	(10)

*where $\dot{m}_b$ is the gas consumption, LHV is the Lower Heating Value, $P_{el,nom}$ is the rated electric output, P_{el} is the electric output, α is the part load factor, and β is a binary control variable.

2.3. Control methods

Different control architectures, based on DRL, including single-agent (centralized) approaches and multi-agent architectures – either fully decentralized or organized hierarchically, have been explored in this work. The baseline is represented by the rule-based controller that is currently in use, and it was structured around working days and operational shifts. Currently, the energy system is managed according to a rigid and simple rule-based strategy that is based on work shifts.

- When available, the turbines are set ON at full capacity;
- During weekdays, the ICE is maintained at full capacity;
- During nights on weekdays, the ICE is kept stable at 70% of its nominal power.
- During weekends and festivities, the ICE is turned off;
- The afterburner is only turned on during weekdays, when the heat demand is higher;
- Steam generators are only required during weekdays.

This control strategy does not take into account the electricity generated by the PV system, which leads to surplus generation and a sub-optimal performance of the system. This is because the EMS is operated manually rather than automatically, thus necessitating the use of a simple strategy.

All the control strategies were benchmarked against a reference solution, derived from an MILP model [54], to enable a comparative assessment, in terms of performance and limitations. This step ensures that the DRL approaches can be evaluated against a reference that is close to the mathematically optimal solution. The MILP formulation relies on perfect knowledge, whereas DRL methods – and real-world applications in general – cannot achieve such accuracy. Therefore, the MILP solution should not be considered a feasible control strategy, but rather an upper bound of all the possible strategies. The simulation environment, as described in Section 2.2 and summarized in Table 2 and Table 3, is identical for all the tested control configurations in order to ensure that any comparison of the methods relies on consistent operational assumptions and system constraints.

2.3.1. Deep reinforcement learning

DRL builds on a Markov Decision Process that is defined by states, actions, rewards, and transitions. Using DNNs, DRL approximates the state-action value function (Q-function) or directly determines optimal actions (policy), and this enables it to handle high-dimensional inputs and complex systems. The state (s_i) represents the current configuration of the environment, and it is approximated through a set of punctual observations, which provide discrete pieces of information from which the agent determines its action (a_i). The environment responds with a new state (s_{i+1}) and a reward (r_i). The agent then iteratively refines its policy through trial-and-error and using reward feedback.

In this work, three different structures have been evaluated, and the obtained results have been compared using centralized, decentralized, and hierarchical configurations of the agents (Fig. 5) whereby:

- Centralized control provides global optimization by considering all the system components simultaneously, simplifying coordination and decision-making, and enabling real-time monitoring of the entire system. However, this type of control suffers from some critical drawbacks, including vulnerability to a single point of failure, limited scalability (i.e. the ability of the architecture to handle growing system dimensionality without an exponential growth of the state-action space), higher communication overheads, and reduced adaptability to local disturbances.
- Decentralized control is more resilient, as the subsystems can still operate independently if one of them fails. It scales well for large and distributed systems and reduces communication overheads. However, it may lead to a suboptimal global performance, due to conflicting local decisions.
- Hierarchical control establishes a hierarchy between different control agents in such a way that, through coordination, they can overcome the suboptimal global performance of decentralized control systems and improve system management.

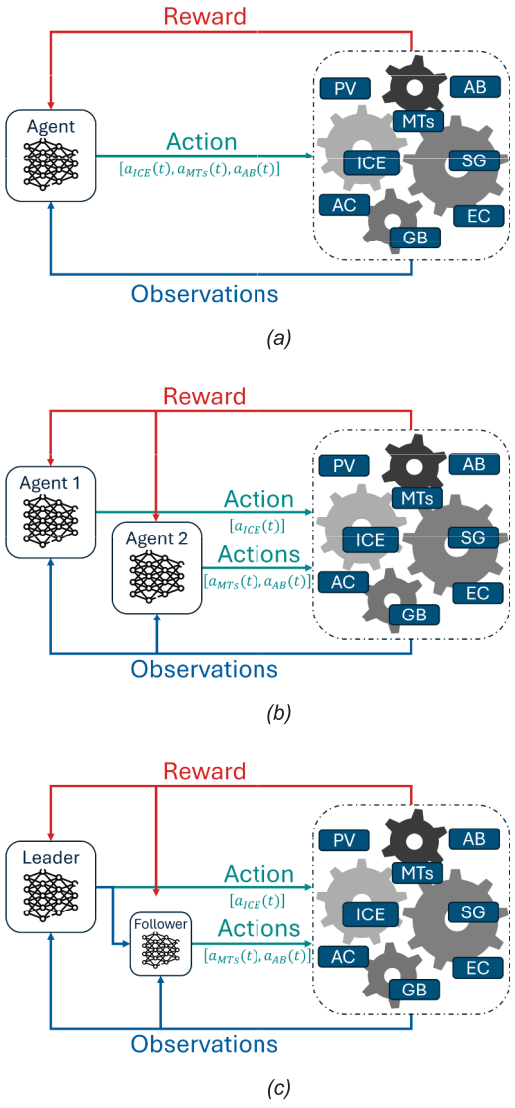

Fig. 5. Scheme of the DRL optimization loops for the (a) centralized, (b) decentralized, and (c) hierarchical configurations.

A centralized control system was studied by developing and training an agent to control the ICE, three MTs, and an afterburner. In the second and third cases, a decentralized control scheme was employed in which the cogeneration units were divided between two agents, with one agent responsible for the CHP-ICE unit and the other agent responsible for the MTs and the AB. The energy demands were satisfied in all the considered cases as the deficits were balanced by the power grid, the SG, gas boilers (hot water) and by compression chillers.

In the cooperative hierarchical control setting, the interaction between agents was represented through a leader–follower configuration: one agent determines its strategy first, taking on the role of leader, while the other optimizes its response based on the leader's observable decision, acting as the follower. Within this framework, the leader has no direct knowledge of the follower's actions and therefore learns, during training, to anticipate the follower's behavior. The leader, by adopting a

cooperative hierarchical structure for the considered problem, is selected from the group exerting the greatest influence on the global reward, namely the CHP-ICE group. The follower, instead, is responsible for controlling the group composed of the MTs and AB.

The optimal solution is achieved only if the follower does not move away from equilibrium by selecting a non-optimal action; any deviation would reduce the follower's performance and, at the same time, negatively affect the leader. Suboptimal behaviors are nevertheless expected, since DRL relies on a trial-and-error learning process; the goal is therefore to approach the optimal solution as closely as possible. Table 4 summarizes the methods that we used, their requirements, and the other characteristics, such as scalability, real-time capability, coordination complexity, and adaptability.

Observations, actions, and rewards

The selection of observations is critical to accurately represent the state of a system. The environment should be characterized by the minimum number of observations necessary to minimize noise and redundancy. This careful selection can positively affect both the training duration and the quality of the resulting models, and it can ultimately enhance their overall performance. The agents observe the environment at each time step, *t,* considering the following variables:

- *Net load*(t), which is the difference between the electricity demand and the PV production;
- *Steam demand*(t), which is the demand for hot steam;
- *HW demand*(t), which is the demand for hot water;
- *CW demand*(t), which is the demand for chilled water;
- *electricity cost*(t), which is the buying cost of electricity;
- *electricity price*(t), which is the selling price of electricity;
- *Gas cost*(t), which is the buying cost of gas;
- *Hour*(t), which is the hour of the day;
- *Day Of Week*(t), which is the day of the week;
- $a_{ICE}(t)$, which is the action undertaken by the CHP-ICE agent.

These quantities can be used in different combinations, depending on the case, as shown in Table 5. The vectors of the observations used were obtained by iterating over different configurations, and the best combinations were then chosen. A follower agent in the hierarchical MADRL configuration receives a reduced set of direct observations, compared to an agent in the flat, non-hierarchical setting. Indeed, such an agent does not observe certain system-level variables, including energy demands for vectors outwith its production capability (e.g., hot and chilled water) and energy costs. Additionally, the follower observes the action taken by the leader agent.

The controllable devices included in the action space are the following:

- CHP-ICE, controlled by the action $a_{CHP}(t)$, which determines its ON/OFF status and load factor;
- The three MTs, each operated through the actions $[a_{MT1}(t), a_{MT2}(t), a_{MT3}(t)]$ which determine their ON/OFF status and load factor;
- The AB, controlled by the action $a_{AB}(t)$ which sets its activation status (according also to the operating conditions of the MTs).

As previously mentioned, in this case, the action space is not entirely continuous, and the approach adopted to handle this aspect is described in Section 2.3.2.

Lastly, given the cooperative nature of the task, all the configurations share the same reward formulation, which reflects the operating costs of the energy system; this aspect is further detailed in Section 4. The energy costs (C) are related to the purchasing of fuel and electricity, while its revenues (R) are related to the selling of electricity to the grid.

$$r(t) = R(t) - (C_{NG}(t) + C_{el}(t) + C_{maint}(t) \tag{11}$$

where:

- $R(t)$ is the total revenue at timestep t;
- $C_{NG}(t)$ is the total cost of the natural gas consumed at timestep t;
- $C_{el}(t)$ is the total cost of the electricity bought at timestep t;
- $C_{maint}(t)$ is the total cost of the maintenance associated with the utilization of the technology (ICE or MTs) at timestep t;

The revenues are only due to the sale of electricity to the grid:

$$R(t) = c_{el,selling}(t) \bullet E_{eltogrid}(t) \tag{12}$$

where $c_{el,selling}$ is the unit cost of electricity [€/MWh].

The total gas costs are calculated as follows:

$$C_{NG}(t) = c_{NG}(t) \bullet \dot{m}_{NG}(t) \bullet LHV_{NG} \tag{13}$$

where c_{NG} is the unit cost of natural gas [€/MWh], and $\dot{m}_{NG}$ is the total gas consumption.

The costs of the electricity bought from the grid are calculated as follows:

$$C_{el}(t) = c_{el,buying}(t) \bullet E_{el,fromgrid}(t) \tag{14}$$

where $c_{el,buying}$ is the unit cost of electricity [€/MWh].

The maintenance costs for the ICE and MTs are therefore calculated as:

$$C_{maint,GE} = c_{maint,GE} \bullet E_{el,GE} \tag{15}$$

Table 4
Comparison of the methods adopted in this work.

Method	Principle	Model Requirements	Scalability	Real-Time Capability	Coordination Complexity	Adaptability
Rule-Based Control (RBC)	Predefined if-then rules	None (rule logic only)	High (but inflexible)	High (instant execution)	None	Very Low (manual updates)
MILP Optimization	Mathematical programming (global optimization under constraints)	Full linearized system model	Low	Strongly dependent on the time horizon and the number of constraints	Centralized solver handles everything	Low (fixed structure)
SADRL	Deep RL with one agent controlling all the devices	Moderate (historical data + hyperparameter tuning)	Low (single agent must scale ccording to the size of the problem)	High (once trained)	Centralized agent handles everything	High (learns from data)
Independent MADRL	Separate DRL agents for each device group	Moderate (historical data + hyperparameter tuning)	High (agents can be added/removed)	High (once trained)	Medium (no coordination, possible, conflicts)	High (local adaptation)
Hierarchical MADRL	Hierarchical DRL with a leader–follower architecture	Moderate (data + hyperparameter tuning)	Medium-High (better coordination than an independent MADRL)	High (once trained)	High (inter-agent coordination via the hierarchy)	High (structured adaptation)

Table 5
Summary of observations, actions, and rewards for the different DRL configurations.

Configuration	Devices	Observations	Action(s)	Reward
SADRL	CHP – ICE, MTs and AB	*Net load*(t), *Steam demand*(t), *HW demand*(t), *CW demand*(t), *electricity cost*(t), *electricity price*(t), *Gas cost*(t), *Hour*(t), *Day Of Week*(t)	$a_{ICE}(t)$, $a_{MT1}(t)$, $a_{MT2}(t)$, $a_{MT3}(t)$, $a_{AB}(t)$	Operating costs
Independent MADRL	CHP – ICE	*Net load*(t), *Steam demand*(t), *HW demand*(t), *CW demand*(t), *electricity cost*(t), *electricity price*(t), *Gas cost*(t), *Hour*(t), *Day Of Week*(t)	$a_{ICE}(t)$	Operating costs
	MTs and AB	*Net load*(t), *Steam demand*(t), *electricity cost*(t), *electricity price*(t), *Gas cost*(t), *Hour*(t), *Day Of Week*(t)	$a_{MT1}(t)$, $a_{MT2}(t)$, $a_{MT3}(t)$, $a_{AB}(t)$	Operating costs
Hierarchical MADRL	CHP – ICE	*Net load*(t), *Steam demand*(t), *HW demand*(t), *CW demand*(t), *electricity cost*(t), *electricity price*(t), *Gas cost*(t), *Hour*(t), *Day Of Week*(t)	$a_{ICE}(t)$	Operating costs
	MTs and AB	*Net load*(t), *Steam demand*(t), *Hour*(t), *Day Of Week*(t), $a_{ICE}(t)$	$a_{MT1}(t)$, $a_{MT2}(t)$, $a_{MT3}(t)$, $a_{AB}(t)$	Operating costs

$$C_{maint,MT} = c_{maint,MT} \bullet E_{el,MT} \tag{16}$$

where $c_{maint,ICE}$ and $c_{maint,MT}$ are the unitary costs for the maintenance of the ICE and MTs, and $E_{el,CHP}$ and $E_{el,MT}$ are the electricity produced by the ICE and MTs.

The total maintenance costs can easily be calculated as:

$$C_{maint} = C_{maint,MT} + C_{maint,ICE} \tag{17}$$

2.3.2. Algorithm implementation and training setup

Two state-of-the-art continuous control algorithms have been tested to evaluate the control performance of the proposed DRL architectures: DDPG and SAC. Both are off-policy, actor–critic algorithms that were designed for continuous control tasks: DDPG employs deterministic policies, whereby Q-learning is combined with deterministic policy gradients [15], while SAC extends this framework by introducing stochastic policies and entropy maximization to encourage exploration [16]. Both approaches use replay buffers to stabilize learning and target networks to smooth value updates, and they mainly differ in how they balance exploration and exploitation. Since the objective of this work has not been to compare DRL algorithms per se, but rather to investigate different DRL control architectures (single-agent, independent multi-agent, and hierarchical multi-agent), the choice of the algorithms has not been the core focus of the analysis. However, for completeness, both algorithms were implemented and tested under identical conditions.

Consistently with the findings of previous studies (see [31,55]), DDPG achieved a slightly better performance, in terms of cost minimization and training efficiency, for the examined industrial multi-energy system. Therefore, for clarity and conciseness, the results presented in the main paper refer to the DDPG-based controllers. The complete results obtained with SAC are reported in the Supplementary Material. As detailed therein, the SAC-based controllers lead to the same qualitative conclusions, regarding the relative performance of the different control architectures, thereby reinforcing the robustness and generality of the findings reported in this work.

Since the main results discussed in this work refer to the DDPG-based controllers, this subsection briefly outlines the mathematical formulation of the DDPG algorithm as implemented in our study. The aim has not been to provide an exhaustive derivation, but rather to summarize the key elements of the actor–critic structure, the updated rules, and the exploration mechanism used in the training process.

Specifically, the actor network ($\mu(s_i|\theta^{\mu})$, where μ is the policy, s_i is the set of environment observations at timestep i, and θ^{μ} represents the parameters of the actor network) learns a deterministic policy that directly maps the observed states onto the actions. The critic network ($Q(s_i, a_i|\theta^{Q})$, where s_i is the set of the environment observations at timestep i, a_i is the action selected at timestep i, and θ^{Q} represents the parameters of the critic network), estimates the value of the actions that have been taken by approximating the Q-function. In order to improve stability, DDPG employs experience replay, where past interactions are stored in a buffer and sampled randomly in mini-batches (of size N) for training purposes, while reducing the correlation between updates. Additionally, DDPG uses target networks ($Q'(s, a|\theta^{Q'})$ and $\mu'(s|\theta^{\mu'})$), which are gradually updated toward the learned networks to mitigate oscillations and divergence during training.

The critic network is updated by minimizing the loss function, L, across all the sampled experiences:

$$L = \frac{1}{N}\sum_{i}\left(y_i - Q(s_i, a_i|\theta^{Q})\right)^2 \tag{18}$$

where y_i is the sum of the experience reward, r_i, and the discounted (γ, discount factor) future reward. The future reward is estimated using the target actor ($\mu'(s|\theta^{\mu'})$) and the target critic ($Q'(s, a|\theta^{Q'})$) evaluated at the next state (s_{i+1}), and it is defined as follows:

$$y_i = r_i + \gamma Q'\left(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}\right) \tag{19}$$

After each updating of the critic network, the actor network is also updated, using the following sampled policy gradient to maximize the expected, discounted, long-term, cumulative reward.

$$\nabla_{\theta^{\mu}} J \approx \frac{1}{N}\sum_{i}\nabla_a Q(s, a|\theta^{Q})|_{s=s_i, a=\mu(s_i)}\nabla_{\theta^{\mu}}\mu(s|\theta^{\mu})|_{s=s_i} \tag{20}$$

The target networks are updated periodically and smoothly, that is, after a certain number of steps (referred to as the target update frequency) and applying a smoothing factor (referred to as target smooth factor τ).

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \tag{21}$$

$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \tag{22}$$

The Ornstein–Uhlenbeck noise model (OU_{noise}) is used to encourage exploration, so that, during training, the action selected at each timestep is:

$$a_i = \mu(s|\theta^{\mu}) + OU_{noise} \tag{23}$$

Since the DRL algorithms adopted to investigate the different configurations only manage continuous action spaces, the non-differentiable or discrete action spaces represent a challenge. The chosen approach uses what is effectively a step function: as shown in Fig. 6, the continuous action produced by an agent within the normalized range $[-1, 1]$ is translated into the corresponding operational mode and load factor of each device. The resulting load factor spans from 0.6 to 1.0, where values below the minimum operating threshold – 0.7 for the ICE (blue) and 0.75 for the MTs (orange) – correspond to an OFF state. A simpler rule is applied for AB (green): negative action values lead to an OFF state, whereas positive ones activate the device. This deterministic mapping was introduced as a practical way of embedding minimum-load constraints and ON/OFF decisions into a continuous-action DRL framework. For the system under analysis, it has proved to be an effective and practically viable way of incorporating discrete and non-differentiable behaviors into a continuous-control algorithm, as shown by the close agreement between the DRL solutions and the MILP

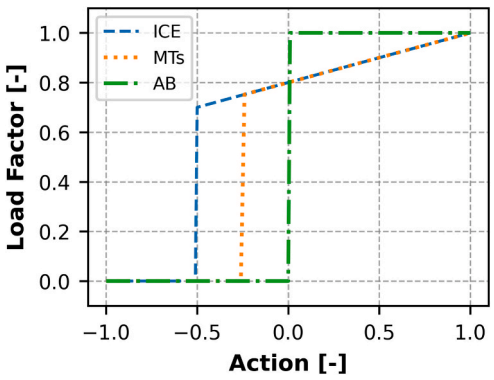

Fig. 6. Post-processing of the actions to obtain the load factor.

benchmark (see Section 3). However, it is worth noting that assessing whether this methodology performs equally well in other settings would require dedicated studies, which are beyond the scope of this work.

The DNN architecture was iteratively tuned but kept consistent across all the algorithms and control structures to guarantee comparability. Both the actor and critic networks were implemented as fully connected, feedforward, neural networks with two hidden layers of 350 nodes each, and the configuration was selected after iterative testing. The hyperparameters were optimized through iterative testing, and the best-performing sets are shown in Table 6: the aim of the hyperparameter tuning is to balance exploration, exploitation, and training stability. The tuning process was initially set up following the general guidelines provided in [15] and then gradually refined to enhance the performance in the SADRL configuration. According to the findings of [29], comparable approaches generally result in similar hyperparameter configurations; hence, the MADRL tuning process was initialized from the optimal configuration obtained for the SADRL framework. It was observed that the hierarchical MADRL achieved optimal results for the same configuration as SADRL, whereas the independent MADRL required slight adjustments. The actor and critic networks were updated using the ADAM optimizer and the learning rates, which determine the step size used to update the weights of the networks after each gradient descent iteration, and they therefore play a huge role in balancing the convergence speed and training stability. Other important hyperparameters for DDPG that required tuning includes the discount factor, which reflects the importance of future rewards, relative to immediate ones (0.95); the size of the experience buffer (1e6), which ensures a diverse set of past interactions for training; the minibatch size (128), which determines the number of samples used at each update; and the target update frequency (1000 and 1200), as well as the target smooth factor (1e-3), which regulates how quickly the target networks follow the main ones and increase stability. The exploration issue was tackled using the OU noise model, which produces correlated and bounded variations in the actions, with a mean reversion constant of 0.15 and a standard deviation of 0.3.

The dataset described in Section 2.1 was partitioned into training and testing subsets to successfully train the agent and to maintain the temporal coherence of the time series. Specifically, the training set was made to contain the even weeks, while the test set was made to include the odd weeks. During training, the agents were permitted to explore the environment freely, provided they remained within the feasible region. While doing so, each agent actively controlled the specified technologies and sought the optimal solution from an economic perspective. Each episode was composed of a sequence of 10 consecutive days, and each step corresponded to one operational hour; after each training episode, the environment was completely reset and the agent was then exposed to a different part of the training set. The training of the agents in the multi-agent configuration was simultaneous, and the interaction of the agents generated a high number of samples, thereby indirectly increasing the exploration of the environment by the agents.

Overfitting was ruled out, according to the results presented in Table 7, as the performance levels remain consistent across the agents. The observed variations across the datasets can be attributed to differences in the energy demands and energy costs, which in turn influence the resulting global reward.

Table 6
The hyperparameters used for DDPG training.

Hyperparameter	SADRL	Ind. MADRL	Hier. MADRL
Critic learn rate	1e-3	0.75e-3	1e-3
Actor learn rate	0.75e-3	1e-3	0.75e-3
Experience buffer size	1e6	1e6	1e6
Minibatch size	128	128	128
Target update frequency	1200	1000	1200
Target Smooth Factor	1e-3	1e-3	1e-3
Discount Factor	0.95	0.95	0.95
OU Mean Reversion Constant	0.15	0.15	0.15
OU Standard Deviation	0.3	0.3	0.3

3. Results

The performance of the different control strategies presented in Section 2.3 is here evaluated and compared. The analysis focuses on key economic and energy-related indicators, with particular attention to operational costs and the utilization profiles of the main technologies. The RBC currently implemented in the plant and the DRL-based strategies – SADRL, independent MADRL, and cooperative hierarchical MADRL – are tested and compared with the MILP optimization model that serves as a benchmark to evaluate the optimality of the DRL approaches.

3.1. Comparative performance of the control strategies: economic and energy analyses

Both economic and energy indicators have been considered to comprehensively evaluate the performance of the different control strategies. Although the total operating cost remains the primary metric – as it is the objective function that is minimized by both the MILP and DRL models – energy Key Performance Indicators (KPIs) provide a deeper insight into how each strategy manages the utilization of the key technologies in the system. The results obtained with the different DRL architectures have been compared with those achieved by the current RBC control (i.e., the currently implemented, rule-based control) and with the outcomes of the MILP-based control, which has been considered the upper bound of the performance of the plant.

As mentioned in Section 2.3.1, the operational costs account for the fuel costs of each device, the maintenance costs, and the electricity costs and revenues. In terms of energy KPIs, both the consumption and production of the main energy vectors were considered in the analysis. Tables 8 and 9 provide detailed reports on the costs and external energy exchanges, while Fig. 7 highlights differences in technology utilization across the strategies (RBC in blue, SADRL in orange, Independent MADRL in green, and Hierarchical MADRL in red), compared to the MILP benchmark (black dashed line). The differences are shown in terms of (a) costs, (b) natural gas consumption, (c) electricity production or electricity exchanged with the grid, and (d) steam production.

The RBC strategy, which has been considered as the baseline for

Table 7
The performance of the agent on the different sets (average step reward).

Average reward [€]	SADRL	Ind. MADRL	Hier. MADRL
Training set	546.23	550.16	545.34
Test set	540.15	543.37	539.18

Table 8
Comparison of the economic KPIs for the different analyzed strategies (RBC, MILP, SADRL, Independent MADRL, and Hierarchical MADRL).

Quantity	UoM	RBC (baseline)	MILP (benchmark)	SADRL	Ind. MADRL	Hier. MADRL
Electricity costs	M€	0.394	0.318	0.421	0.539	0.381
Fuel costs	M€	4.47	3.82	3.74	3.73	3.81
Maintenance costs	M€	0.393	0.312	0.298	0.291	0.310
− Electricity revenues	M€	0.635	0.171	0.144	0.217	0.200
Operating costs	**M€**	**4.63**	**4.29**	**4.32**	**4.34**	**4.31**

Table 9
Comparison of the energy KPIs for the different analyzed strategies (RBC, MILP, SADRL, Independent MADRL, and Hierarchical MADRL).

Quantity	UoM	RBC (baseline)	MILP (benchmark)	SADRL	Ind. MADRL	Hier. MADRL
Electricity bought	GWh	1.79	1.45	1.92	2.41	1.72
Electricity sold	GWh	5.80	1.55	1.29	1.96	1.81
ICE Gas Cons.	GWh	42.9	39.7	38.4	41.2	40.0
MT and AB Gas Cons.	GWh	18.7	10.8	9.38	6.27	10.4
Steam Gen. Gas Cons.	GWh	2.58	4.95	6.11	6.74	4.91
Total Gas Cons.	GWh	64.1	55.4	53.9	54.2	55.3

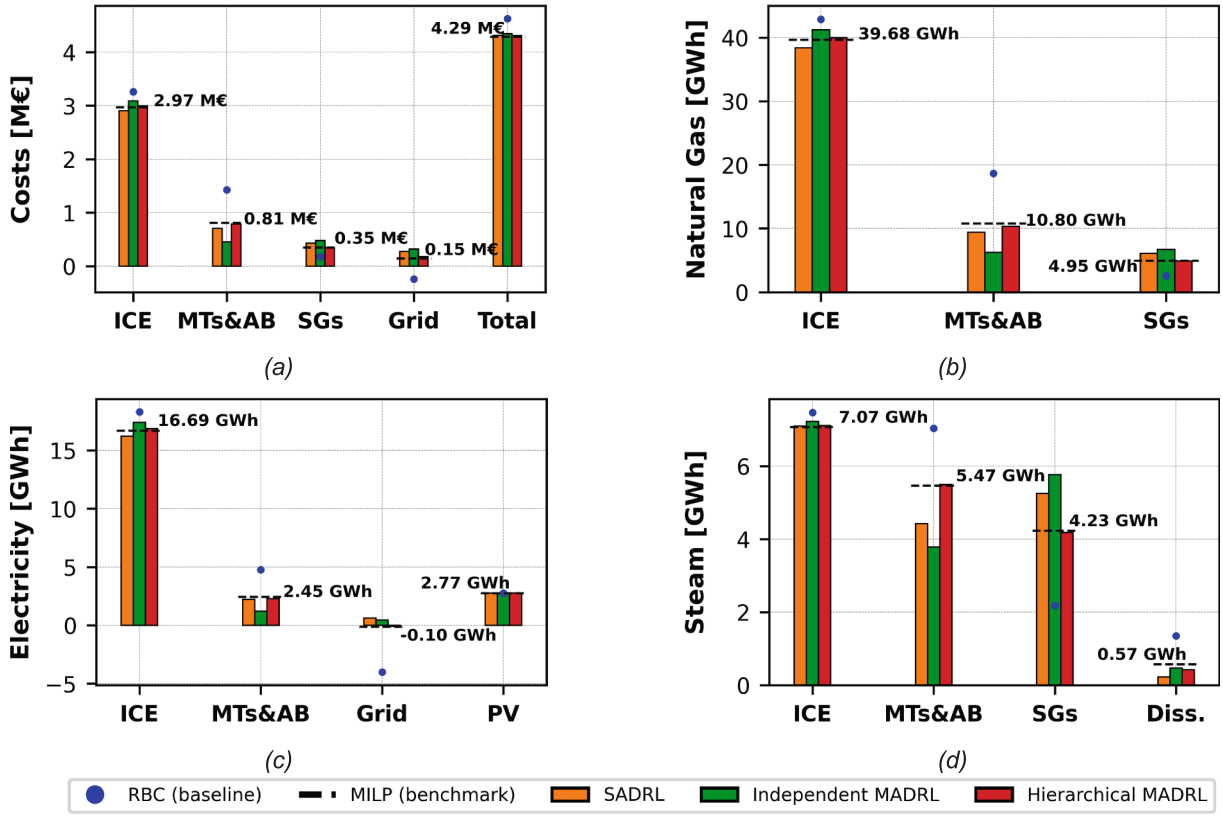

Fig. 7. Comparison of the costs, energy production, and fuel consumption for the control strategies. (a) Operating costs by technology; (b) Natural gas consumption by device; (c) Electricity balance (generation, grid exchanges, PV); (d) Steam production from different sources.

improvements, is the worst-performing one (4.63 M€), due to its lack of flexibility. It exhibits the highest fuel and maintenance costs, it frequently overproduces electricity, which is sold to the grid at low prices (Fig. 7c, blue dots), and it fails to offset the increased production costs. Moreover, it results in the largest amount of dissipated steam (Fig. 7d), thereby further highlighting the inefficiencies of this control strategy. Altogether, these factors lead to the highest overall operating costs of all the analyzed configurations.

Cooperative hierarchical-MADRL demonstrates the best balance of the DRL methods, that is, it achieves a total operating cost of 4.31 M€ (−6.9% compared to the actual RBC control used as a baseline), which is nearly equivalent to that of MILP (4.29 M€, −7.3%) and significantly lower than the RBC, as it effectively reduces fuel and maintenance costs while balancing grid imports. When comparing the technology utilization images, it is possible to assert that the hierarchical MADRL solution has very similar utilization patterns to those of MILP optimization, with deviations mostly for the bought (+18.6%) and sold (+16.4%) electricity, and for the MT and AB gas consumption (−3.8%).

The SADRL configuration also shows a great reduction in costs, when compared with RBC (4.32 M€, −6.7%), and it is almost equivalent to the hierarchical MADRL configuration; however, from a technology utilization point of view, it is possible to notice that the SADRL strategy tends to use the ICE, MTs and AB less than the MILP optimal solution and the hierarchical MADRL, although it involves buying more electricity (+32.4%) from the grid and producing more steam through the steam generators (+23.4%).

The worst performing DRL configuration is that of the independent MADRL (4.34 M€, −6.3%), due to the lack of coordination of the agents in the system: the differences from MILP in the utilization of ICE (+3.8%) and MTs (−41.9%) are sizeable, and this leads to higher utilization of the steam generators. The bought and sold electricity also differs, thus signaling a poor usage of the MTs.

It is worth noting that the results obtained with the DDPG algorithm are very close to the MILP benchmark, thus confirming that the approach used to manage the discrete or non-differentiable actions does not compromise the stability or the convergence of the training process.

The electricity and steam flows of the plant are shown in Fig. 8a–b for the current RBC solution and the optimal solution obtained with the MILP, respectively. The main difference is that RBC does not adjust the output of the ICE or MTs, and this leads to a significant overproduction both during renewable generation peaks and at night when the electricity demand decreases. As reported in Table 9, the electricity sold to the grid in the RBC case is more than 2.5 times higher than in the optimal case. In addition, the ICE is switched off during weekends, while the MTs constantly remain on; these differences are sufficient to cover the plant's energy needs, but at the cost of increasing the amount of electricity purchased from the grid. It should also be noted that the extensive use of the ICE on weekdays enables the production of a large amount of cogenerated steam, thus reducing the use of SGs. However, even though the utilization of SG decreases, the production of steam is still not optimized and, as shown in Fig. 7d, the overproduction of steam is greater for the RBC case than for the MILP case.

The solutions in the DRL configurations (Fig. 8c–e) closely reproduce the optimal behavior identified by the MILP on weekdays, thus confirming the agents' ability to correctly adapt to variations in demand and renewable generation, with only minor differences in the management of the MTs. However, more pronounced differences emerge for weekends: in particular, SADRL and the independent MADRL architecture adopt strategies that deviate from the optimal one, while the hierarchical configuration remains close to the MILP reference one. This deviation is mainly due to the operating characteristics of the system under low-demand conditions. Indeed, the energy requirements are reduced during weekends, and the presence of technical constraints – such as the minimum output of cogeneration units – leads to situations in which switching on a unit inevitably results in producing more energy than is required. In such cases, the difference between keeping the unit off (purchasing the unavailable energy from the grid) and keeping it on (with an excess fed into the grid) is economically very small (see the total operating costs of the three solutions reported in Table 8), thus making it harder for DRL to distinguish between the optimal strategy and a suboptimal one.

3.2. Impact of cost structure perturbations on the control performance

A stress test was conducted to assess the robustness and adaptability of the DRL-based control strategies by independently varying the cost of natural gas and electricity. This sensitivity analysis examined scenarios that deviated from the usual relationship between energy vectors, such as those resulting from market volatility, increased use of renewable energy sources, or changes in regulations.

Two parameters were systematically adjusted:

- Gas Cost Multiplier: the baseline natural gas price was scaled from 0.6 × to 1.4 × .
- Electricity Cost Multiplier: the electricity purchase price was varied from 0.2 × to 1.8×, while maintaining a constant ratio between the buying and selling prices to preserve market symmetry (the electricity selling prices were on average equal to 50% of the electricity buying cost).

The obtained scenarios can be grouped into four distinct cases (Fig. 9) on the basis of the relative cost of electricity and natural gas, and they reveal different optimal operating regimes and strategic responses of the energy system.

- *Zone A:* this is the most populated region regarding the tested scenarios, and it represents the current and likely short- and medium-term market conditions. Although the design of the energy system was completed prior to the introduction of the PV system, it was originally optimized on the basis of these economic scenarios. From a control strategy standpoint, it is more convenient, on average, to follow the net electrical demand while treating the associated heat as a secondary product to be exploited when available. The DRL agents were trained using cost data from the central section (green tile). Therefore, Zone A can be considered the main benchmark to interpret the results. The agents' behavior and robustness should primarily be assessed within this region, as it reflects the operational environment for which the real system was conceived and designed.
- *Zone B:* the electricity prices in this region are high enough and the gas prices are low enough to justify prioritizing steam production. The resulting optimal strategy is to follow the thermal load. Although the system under analysis was originally designed to follow the electrical demand, such a heat-driven operation is commonly adopted in many industrial settings as a control strategy.
- *Zones C and D:* these regions correspond to extreme price conditions under which the current design of the energy system would be economically inappropriate. In Zone C, the cost of electricity generated by the CHP units exceeds the grid purchase price, even when full heat and steam recovery is considered. In such a situation, a rational operator would import electricity and meet the thermal demand with heat-only technologies, since the operation of ICEs and MTs would be economically unjustified. Conversely, in Zone D, the electricity price becomes so favorable, relative to the gas cost, that producing electricity solely to export it to the grid is profitable, regardless of the plant's thermal demand. However, these scenarios should be regarded as stress tests rather than realistic operating conditions, since such drastic economic shifts fall well outside the training conditions and would require not only an adjustment of the control policy but also a redesigning of the energy system itself. Accordingly, any performance degradation observed in zones C and D should be interpreted in this light, while the main robustness assessment should be focused on the more realistic cost variations explored in zones A and B.

The relative deviation of the operating costs was computed for each price combination to compare all the DRL control strategies with the MILP benchmark. It should be noted that the DRL agents (SADRL, independent MADRL, and hierarchical MADRL) were evaluated without any retraining with the new energy costs. Fig. 10 summarizes these differences across all the control architectures, with each heatmap representing a distinct DRL variant. The performance is segmented into the above-described operating zones – A, B, C and D – and is highlighted in the plots. The obtained results show that:

- *Zone A:* Despite a broad variation in energy costs across this region, all the DRL agents demonstrate a high degree of robustness, with relative errors generally below 1.5% and often under 1% in the core of the zone. Only in one borderline scenario – closer in nature to Zone C – are error levels observed in the 5–8% range. This confirms that the DRL agents have successfully learned to operate the system optimally under conditions that are similar to those of the historical dataset. When considering the DRL architectures, it can be observed that the hierarchical MADRL configuration (Fig. 10c) appears slightly more robust than the SADRL one (Fig. 10a), while the independent MADRL configuration (Fig. 10b) shows a marginally lower performance across all the scenarios within Zone A.
- *Zone B* presents a regime in which the optimal dispatch logic shifts slightly. Despite this, the MADRL configurations continue to perform

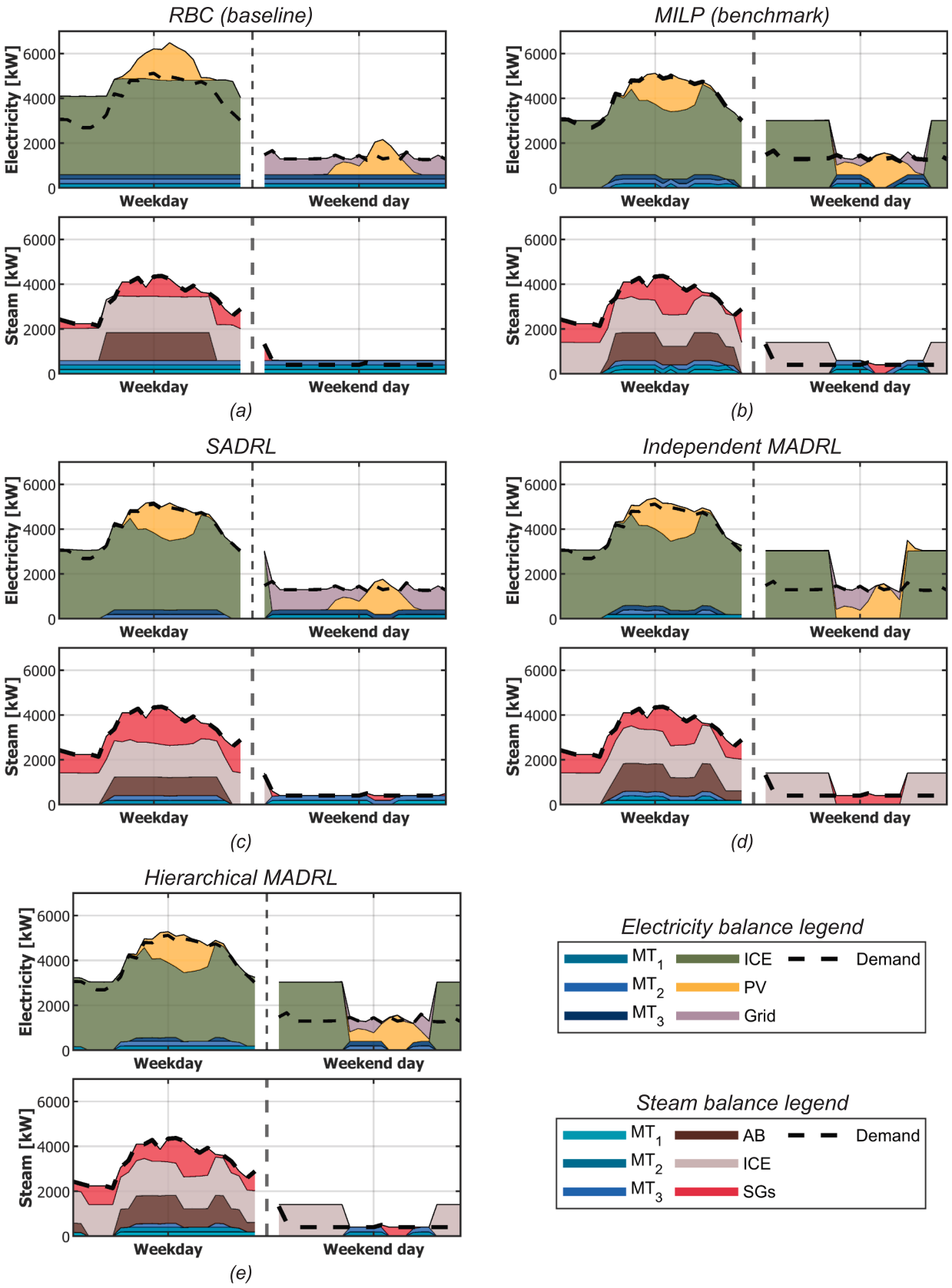

Fig. 8. Comparison of the technology utilization profiles (electricity and steam) over two sample days obtained with: (a) RBC (baseline), (b) MILP (benchmark), (c) SADRL, (d) Independent MADRL, and (e) Hierarchical MADRL.

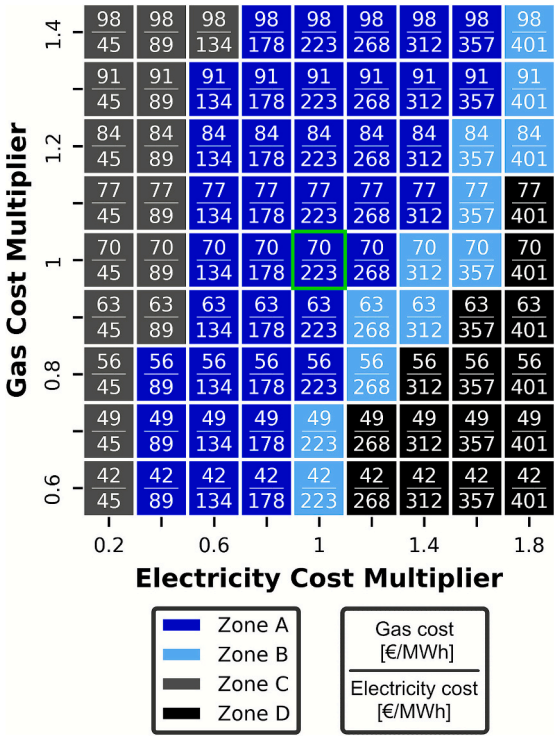

Fig. 9. Division of the cost parameter space into optimal operational zones; each zone (A, B, C or D) represents a distinct control strategy that minimizes the operating costs for different combinations of electricity and gas costs. The values reported in the figure are the annual mean gas and electricity costs.

reasonably well – especially the hierarchical one – showing errors below 3–5% in most tiles, whereas SADRL shows increasing deviations (~10% in the worst tiles). This suggests that multi-agent approaches provide improved generalization by decoupling the control of different subsystems and specializing their responses, even when the reward is shared between agents.

- *Zones C and D* exhibit extreme and structurally distinct operating conditions, compared to the training data. As expected, deviations from the benchmark increase markedly, exceed 30% in Zone C and reach up to 90% in Zone D. Such results are not unexpected, as these configurations correspond to cost structures that are very different from the conditions for which the control strategies were developed. As previously mentioned, they would imply such different market situations for which the current system design would no longer be appropriate and, for this reason, the reduced performance of the trained agents in these regions is of limited significance.

3.3. Impact of demand and renewable energy generation perturbations on the control performance

In a similar manner, the total energy demand and renewable electricity production were disturbed to evaluate the robustness of each control configuration under altered operating conditions. Two parameters were independently adjusted:

- The energy demand multiplier: the main energy demands (electricity and steam) were scaled from 0.5 × to 1.5 × .
- PV capacity multiplier: the installed PV capacity, and consequently the renewable electricity production, were varied from 0 × to 2 × .

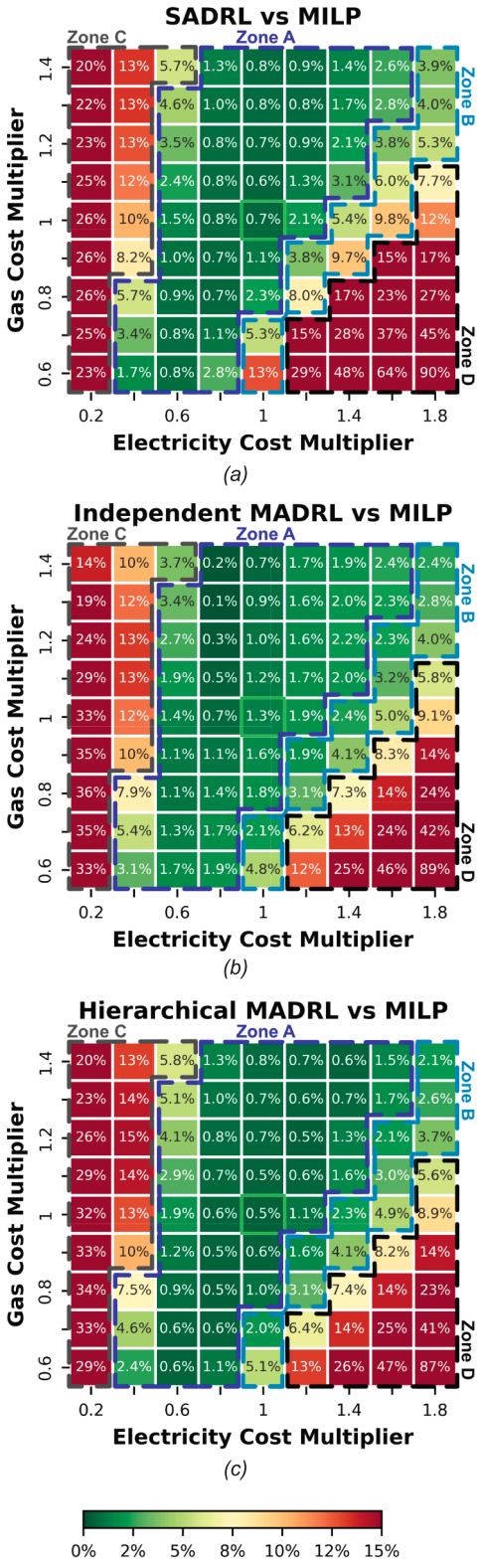

Fig. 10. Relative economic performance of (a) SADRL, (b) the independent MADRL and (c) the hierarchical MADRL, with respect to MILP, under different cost scenarios.

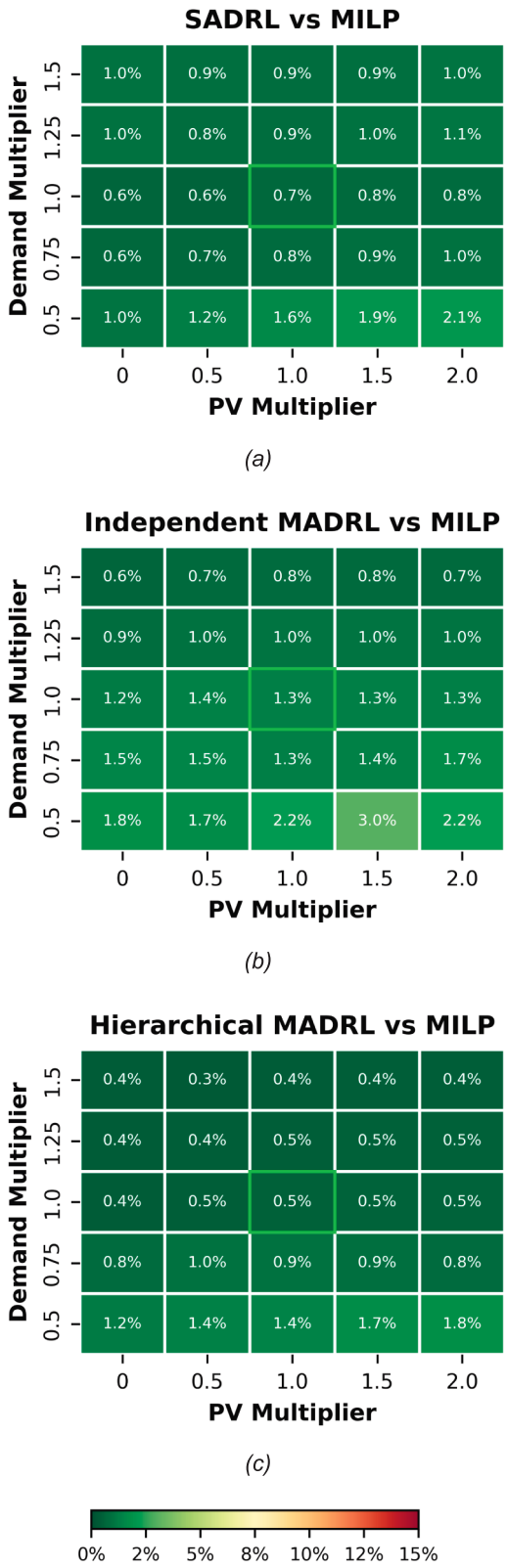

Fig. 11. Relative economic performance of (a) SADRL, (b) the independent MADRL, and (c) the hierarchical MADRL with respect to the MILP for different energy demands/renewable production scenarios.

Although no significant short-term reduction in the energy demand or installed PV capacity was expected, this analysis examined the behavior of the control strategies under a wide range of operating conditions to test the robustness of the proposed control algorithms. As can be seen from Fig. 11, the performance remains stable as these two parameters vary, and it remains close to the mathematical optimum in all the analyzed scenarios. The largest deviations are observed for cases with a very low energy demand and very high PV production. It is worth noting that the system under analysis was designed to meet the current energy demand of the considered industrial plant; therefore, when the demand is reduced by 50%, the installed capacity becomes significantly oversized. This, in turn, makes the control task more challenging: the minimum load constraints of the cogeneration units induce the agent to select slightly suboptimal operating strategies. Nevertheless, even in these cases, the deviation from the optimal operating cost (MILP benchmark) never exceeds 3%, thus demonstrating the high level of robustness of all of the three considered DRL architectures (see Fig. 11). Overall, the hierarchical MADRL configuration again shows the most stable and robust performance across most of the tested domain, with the SADRL only outperforming it in the low-demand/low-PV region.

4. Discussion

This section discusses the key aspects of the proposed control architectures. First, the credit assignment problem is examined for different configurations; the limitations of the simulated environment and the sim-to-real gap are then analyzed, with regards to a real-world implementation.

4.1. The credit assignment problem

Coordinating multiple, interconnected devices is an inherently complex task. The performance of the entire system depends on the combined effect of all the devices. Indeed, the operating point of one device can influence the operation of another, or two different devices can produce similar effects. Under such conditions, a learning-based controller cannot isolate the impact of a single action: the obtained reward signal reflects the aggregate result of all the actions taken by the agent (or agents). Consequently, it becomes difficult for a learning-based controller to understand which specific decisions actually led to the achieved performance. This difficulty is commonly known as the credit assignment problem. The credit assignment problem can be present in both single-agent and multi-agent configurations, but becomes even more severe in the latter: the overall reward depends on the simultaneous actions of multiple agents, and this makes it difficult to attribute the share of credit or responsibility for the obtained result to each agent. When all the agents share the same reward function – for example, the total operational cost of the system – each receives the same signal, regardless of the effectiveness of their individual actions. This can hinder behavioral specialization and reduce the efficiency of cooperative learning. A possible solution to this problem is to assign individual rewards to each agent, thus encouraging the learning of specialized behaviors and enabling a clearer attribution of credit. However, this approach is not applicable in the case under consideration: all the devices in the energy system ICE, MTs, SGs) pursue the same goal, namely, to meet the overall energy demands of the plant while minimizing the total costs. In this context, there are no distinct subgoals or partial metrics that can be assigned to each agent without introducing distortions: one unit of steam or electricity produced by either the ICE or by the MTs has the same operational value for the system, and defining separate rewards would therefore lead to a misalignment with the global objective and could compromise cooperation among the agents. However, the adoption of a hierarchical structure makes it possible to partially reduce the credit assignment problem, while maintaining a shared reward: since the follower agent observes the leader's action, it can react to its decision, while the leader learns to anticipate the action of the follower agent during training. Therefore, the problem is decomposed into two segments: the first is linked to the action of the main ICE and its impact on the common reward, while the second is related to the impact of the MTs and AB, depending on the choice of the first agent.

4.2. Sim-to-real gap

Although the proposed control architectures were trained and validated in a simulated environment, this environment was built upon real operational data, thereby mitigating the sim-to-real gap. However, other challenges could arise in real deployment, particularly regarding control system integration, online adaptation to changing operational conditions, and the ensuring of operational safety.

The present study does not explicitly account for random component failures, which could occur under real operation conditions: this simplification was necessary to isolate the effect of the control architecture itself. It would be interesting to assess the robustness of the different architectures under fault conditions: to accomplish this, the models could be trained by systematically varying the efficiency of the different components within an acceptable range, as a decrease in efficiency could effectively simulate a fault or multiple degradation levels. The development of control strategies capable of maintaining a stable performance, despite efficiency losses or component wear, would be highly valuable in industrial contexts. However, this analysis should ideally be complemented by real operational data, collected under fault conditions, to identify which measurable parameters should be monitored to trigger corrective actions by the agents. Apart from ensuring a satisfactory performance under non-rated conditions, the control system should also be able to detect and signal the occurrence of such faults to enable timely interventions. Although this aspect is beyond the scope of the present work, it will be addressed in future studies.

The temporal resolution of the dataset was set at one hour, which limits the representation of short-term variations in the demand and in the equipment responses that can occur under real operation conditions. Although such a resolution is common in industrial energy management studies, and it is sufficient to capture the main operational trends, higher-frequency fluctuations – for instance due to load transients or rapid renewable generation changes – are inevitably smoothed out. Consequently, the trained policies primarily reflect average system dynamics rather than minute-scale adjustments.

A practical pathway toward real-world implementation could involve a progressive integration of the DRL-based controller within the existing industrial system: rather than a direct replacement of rule-based logic, the proposed architecture could initially operate in an advisory mode, that is, suggesting optimal setpoints to human operators or PLCs (Programmable Logic Controllers). Any potential discrepancies or unforeseen behaviors identified under real operating conditions could be addressed, during a gradual implementation phase, through soft retraining of the agent using transfer learning techniques. In such an approach, the pre-trained policy would serve as a strong baseline, while limited additional training on real plant data would allow the controller to adapt to specific operational nuances, without compromising previously learned behaviors. Ultimately, this kind of incremental deployment could be key to ensuring reliability and operator acceptance of learning-based control systems in industrial environments.

5. Conclusions

This work has involved developing and testing a data-driven EMS, in a simulated environment, for an existing multi-energy industrial plant, which simultaneously satisfies the electricity, steam, hot water, and chilled water demands. The system includes a gas-powered ICE, three MTs with an AB, SGs, a PV system, and absorption chillers: each subsystem can operate autonomously and adaptively, which is crucial in complex environments with multiple energy technologies.

This work has applied and compared different DRL-based control architectures to assess their effectiveness within the context of a real industrial multi-energy system. Specifically, we implemented three control configurations – a centralized SADRL, a fully decentralized independent MADRL, and a cooperative hierarchical MADRL. The performances of the agents were compared with the currently adopted RBC strategy and with a benchmark optimal solution, obtained via MILP, to verify the strategies encountered during training. We found that:

- The DRL strategies were affected by the differences in the small cost-effectiveness of the various devices, thus making it challenging for agents to discern what actions contributed the most to the reward. As a result, certain inefficiencies persisted – especially in the use of MTs, which often played a secondary role in the optimal strategies, due to their smaller scale.
- Nonetheless, all the DRL-based control strategies showed a substantial reduction in the operating costs, compared to the control strategy that is currently adopted. The cooperative hierarchical MADRL achieved the lowest total cost (M€ 4.31), which was very close to the optimal MILP solution (M€ 4.29), and this led to a 6.9% reduction in the total operation costs, compared to the actual RBC strategy. This was followed very closely by SADRL (M€ 4.32, −6.7%). The independent MADRL strategy achieved a reduction of 6.3%, and this shows how a lack of coordination slightly undermined the global optimization. These results confirm the value of DRL for cost-effective operation, and they show that cooperative and centralized approaches can closely replicate the optimal benchmark.
- In terms of energy-related KPIs, the main differences between the control strategies lie in the utilization of the MTs and AB. The hierarchical MADRL configuration achieved a technology utilization profile that closely mirrored the MILP optimal solution. Although the SADRL approach approximated the MILP strategy economically, it also showed some deviations concerning the use of MTs and the AB. The independent MADRL instead suffered from a lack of coordination, and this led to an overproduction of electricity and an underutilization of the MTs, which, in turn, resulted in a drop in the overall performance.

To assess the robustness of the DRL-based strategies, a stress test was conducted by independently manipulating the energy prices (electricity and natural gas), energy demand, and PV production. Such studies are useful for testing the robustness of an implemented architecture and identifying areas that require retraining:

- As the energy costs varied, all the DRL configurations showed performances close to the MILP benchmark, with deviations remaining within the acceptable limits. Noticeable performance degradation only occurred under extreme and highly unlikely cost conditions, such as simultaneous low gas prices and very high electricity prices. The DRL agents diverged more significantly from the MILP reference in these extreme scenarios; however, these discrepancies were linked to market situations that were unrealistic considering the current realities, and they mean that the existing system layout would require redesigning. Among the tested approaches, the hierarchical MADRL configuration demonstrated a higher level of robustness to cost variations than SADRL, particularly in the scenarios that departed slightly from the training distribution.
- As the energy demand and photovoltaic production varied, all three solutions achieved a working condition that was very close to the mathematical instant found with the MILP solution. This highlights the good robustness of the DRL approaches for the optimization of industrial MES systems.

Overall, the results indicate that both DRL approaches offer robust and economically viable solutions across a wide range of operating conditions, and that they outperform simple, rule-based strategies. Our results also show that MADRL – especially when interaction between agents is enabled – is able to perform as well as, or even slightly better, than its centralized counterpart. This study has also shown that a cooperative hierarchical configuration is more stable for scenarios that deviate from the training conditions, and it thus represents a more robust strategy.

Declaration pertaining to the use of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the English and readability of the text. After using this tool, the authors reviewed and edited the content as necessary, and they take full responsibility for the content of the published article.

CRediT authorship contribution statement

Andrea Franzoso: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation. **Gabriele Fambri:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Conceptualization. **Marco Badami:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The authors gratefully acknowledge Trigenia S.r.l. for providing the operational data of the industrial plant used in this study.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enconman.2025.120963.

Data availability

The data that has been used is confidential.

References

[1] Andiappan V. State-of-the-art review of mathematical optimisation approaches for synthesis of energy systems. Process Integr Optim Sustain Oct. 2017;1(3):165–88. https://doi.org/10.1007/S41660-017-0013-2/TABLES/2.
[2] Gajdzik B, Nagaj R, Wolniak R, Bałaga D, Žuromskaitė B, Grebski WW. Renewable energy share in european industry: analysis and extrapolation of trends in EU countries. Energies 2024;17:2476. https://doi.org/10.3390/EN17112476.
[3] Kaygusuz K. Energy efficiency and renewable energy sources for industrial sector. Energy Serv Fundament Financ 2021:213–38. https://doi.org/10.1016/B978-0-12-820592-1.00009-9.
[4] Frieden F, Leker J, von Delft S. A multi-objective analysis of grid-connected local renewable energy systems for industrial SMEs. J Energy Storage Sep. 2024;98:113033. https://doi.org/10.1016/J.EST.2024.113033.
[5] Huang A, Bi Q, Dai L. Integrated economic and environmental optimization for industrial consumers: a dual-objective approach with multi-carrier energy systems and fuzzy decision-making. Energy Jun. 2025;324:135787. https://doi.org/10.1016/J.ENERGY.2025.135787.
[6] Hui H, Bao M, Ding Y, Meinrenken CJ. Incorporating multi-energy industrial parks into power system operations: a high-dimensional flexible region method. IEEE Trans Smart Grid 2025;16(1):463–77. https://doi.org/10.1109/TSG.2024.3426997.
[7] Zuijderwijk IR, Torres JLR, Palensky P. Optimization strategy for flexible operation of integrated multi-energy industrial clusters. In: Proceedings - 2025 IEEE 7th Global Power, Energy and Communication Conference, GPECOM; 2025; 2025. p. 735–40. https://doi.org/10.1109/GPECOM65896.2025.11062020.
[8] Sievers J, Blank T. A systematic literature review on data-driven residential and industrial energy management systems. Energies 2023;16:1688. https://doi.org/10.3390/EN16041688.
[9] Machlev R, Zargari N, Chowdhury NR, Belikov J, Levron Y. A review of optimal control methods for energy storage systems - energy trading, energy balancing and electric vehicles. J Energy Storage Dec. 2020;32:101787. https://doi.org/10.1016/J.EST.2020.101787.
[10] Mandal PK. A review of classical methods and nature-inspired algorithms (NIAs) for optimization problems. Results Control Optim Dec. 2023;13:100315. https://doi.org/10.1016/J.RICO.2023.100315.
[11] Vamvakas D, Michailidis P, Korkas C, Kosmatopoulos E. Review and evaluation of reinforcement learning frameworks on smart grid applications. Energies 2023;16:5326. https://doi.org/10.3390/EN16145326.
[12] Wang X, et al. Deep reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst Apr. 2024;35(4):5064–78. https://doi.org/10.1109/TNNLS.2022.3207346.
[13] Schulman J, Wolski F, Dhariwal P, Radford A, Openai OK. Proximal policy optimization algorithms. Accessed: Aug. 29, 2025. [Online]. Available: https://arxiv.org/pdf/1707.06347.
[14] Mnih V et al. Playing atari with deep reinforcement learning," Dec. 2013, Accessed: Feb. 06, 2025. [Online]. Available: https://arxiv.org/abs/1312.5602v1.
[15] Lillicrap TP et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971; 2015.
[16] Haarnoja T et al. Soft actor-critic algorithms and applications; 2018, Accessed: Feb. 06, 2025. [Online]. Available: https://arxiv.org/abs/1812.05905v2.
[17] Perera ATD, Kamalaruban P. Applications of reinforcement learning in energy systems. Renew Sustain Energy Rev Mar. 2021;137:110618. https://doi.org/10.1016/J.RSER.2020.110618.
[18] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. Artificial Intell Rev 2021;55:2. https://doi.org/10.1007/S10462-021-09996-W.
[19] Lee S, Seon J, Hwang B, Kim S, Sun Y, Kim J. Recent trends and issues of energy management systems using machine learning. Energies 2024;17:624. https://doi.org/10.3390/EN17030624.
[20] Stoffel P, Maier L, Kümpel A, Schreiber T, Müller D. Evaluation of advanced control strategies for building energy systems. Energ Build Feb. 2023;280:112709. https://doi.org/10.1016/J.ENBUILD.2022.112709.
[21] De Somer O, Soares A, Vanthournout K, Spiessens F, Kuijpers T, Vossen K. Using reinforcement learning for demand response of domestic hot water buffers: a real-life demonstration. In: 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT-Europe 2017 - Proceedings, vol. 2018-January; 2017. p. 1–7. https://doi.org/10.1109/ISGTEUROPE.2017.8260152.
[22] Park JY, Dougherty T, Fritz H, Nagy Z. LightLearn: an adaptive and occupant centered controller for lighting based on reinforcement learning. Build Environ Jan. 2019;147:397–414. https://doi.org/10.1016/J.BUILDENV.2018.10.028.
[23] Ding X, Du W, Cerpa A. OCTOPUS: Deep reinforcement learning for holistic smart building control. In: BuildSys 2019 - Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. vol. 19; 2019. p. 326–35. https://doi.org/10.1145/3360322.3360857.
[24] Rossi F, Diaz-Londono C, Li Y, Zou C, Gruosso G. Smart electric vehicle charging algorithm to reduce the impact on power grids: a reinforcement learning based methodology. IEEE Open J Veh Technol 2025;6:1072–84. https://doi.org/10.1109/OJVT.2025.3559237.
[25] Nakabi TA, Toivanen P. Deep reinforcement learning for energy management in a microgrid with flexible demand. Sustain Energy Grids Networks Mar. 2021;25:100413. https://doi.org/10.1016/J.SEGAN.2020.100413.
[26] Zhou Y, Ma Z, Zhang J, Zou S. Data-driven stochastic energy management of multi energy system using deep reinforcement learning. Energy 2022;261:125187. https://doi.org/10.1016/j.energy.2022.125187.
[27] Alabi TM, Lawrence NP, Lu L, Yang Z, Bhushan Gopaluni R. Automated deep reinforcement learning for real-time scheduling strategy of multi-energy system integrated with post-carbon and direct-air carbon captured system. Appl Energy 2023;333:120633. https://doi.org/10.1016/J.APENERGY.2022.120633.
[28] Ceusters G, et al. Model-predictive control and reinforcement learning in multi-energy system case studies. Appl Energy Dec. 2021;303:117634. https://doi.org/10.1016/J.APENERGY.2021.117634.
[29] Bousnina D, Guerassimoff G. Optimal energy management in smart energy systems: a deep reinforcement learning approach and a digital twin case-study. Smart Energy Nov. 2024;16:100163. https://doi.org/10.1016/J.SEGY.2024.100163.
[30] Ruan Y, Liang Z, Qian F, Meng H, Gao Y. Operation strategy optimization of combined cooling, heating, and power systems with energy storage and renewable energy based on deep reinforcement learning. J Build Eng 2023;65:105682. https://doi.org/10.1016/j.jobe.2022.105682.
[31] Ghione G, Randazzo V, Pasero E, Badami M. Optimal cogeneration scheduling: a comparison of genetic and POMDP-based deep reinforcement learning approaches. IEEE Access 2025;13:128562–81. https://doi.org/10.1109/ACCESS.2025.3590255.
[32] Lu R, Jiang Z, Yang T, Chen Y, Wang D, Peng X. A novel hybrid-action-based deep reinforcement learning for industrial energy management. IEEE Trans Industr Inform 2024;20(10):12461–75. https://doi.org/10.1109/TII.2024.3424529.
[33] Qiu D, Chen T, Strbac G, Bu S. Coordination for multienergy microgrids using multiagent reinforcement learning. IEEE Trans Industr Inform Apr. 2023;19(4):5689–700. https://doi.org/10.1109/TII.2022.3168319.
[34] Safiri S, Nikoofard A, Khosravy M, Senjyu T. Multi-agent distributed reinforcement learning algorithm for free-model economic-environmental power and CHP dispatch problems. IEEE Trans Power Syst Sep. 2023;38(5):4489–500. https://doi.org/10.1109/TPWRS.2022.3217905.
[35] Zhang G, et al. A multi-agent deep reinforcement learning approach enabled distributed energy management schedule for the coordinate control of multi-energy hub with gas, electricity, and freshwater. Energy Convers Manag 2022;255:115340. https://doi.org/10.1016/j.enconman.2022.115340.
[36] May R, Huang P. A multi-agent reinforcement learning approach for investigating and optimising peer-to-peer prosumer energy markets. Appl Energy Mar. 2023;334:120705. https://doi.org/10.1016/J.APENERGY.2023.120705.
[37] Ye Y, Papadaskalopoulos D, Yuan Q, Tang Y, Strbac G. Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services

provision in local electricity markets. IEEE Trans Smart Grid Mar. 2023;14(2): 1541–54. https://doi.org/10.1109/TSG.2022.3149266.
[38] Park K, Moon I. Multi-agent deep reinforcement learning approach for EV charging scheduling in a smart grid. Appl Energy Dec. 2022;328:120111. https://doi.org/10.1016/J.APENERGY.2022.120111.
[39] Bo W, Wang X, Jing G, Xu H, Jia L. Comfort and energy management of multi-zone HVAC system based on multi-agent deep reinforcement learning. In: Proceedings of the 18th IEEE Conference on Industrial Electronics and Applications, ICIEA 2023; 2023. p. 1641–6. https://doi.org/10.1109/ICIEA58696.2023.10241916.
[40] Zhu D, Yang B, Liu Y, Wang Z, Ma K, Guan X. Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park. Appl Energy Apr. 2022;311:118636. https://doi.org/10.1016/J.APENERGY.2022.118636.
[41] Wang J, Guo C, Yu C, Liang Y. Virtual power plant containing electric vehicles scheduling strategies based on deep reinforcement learning. Electr Pow Syst Res Apr. 2022;205:107714. https://doi.org/10.1016/J.EPSR.2021.107714.
[42] Zhang X, Yan G, Zhou M, Tang C, Qu R. Energy management method for virtual power plant based on double-layer deep reinforcement learning game. In: 2024 6th International Conference on Energy, Power and Grid (ICEPG); 2024. p. 650–5. https://doi.org/10.1109/ICEPG63230.2024.10775617.
[43] Chen Y, Norford LK, Samuelson HW, Malkawi A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energ Buildings Jun. 2018;169:195–205. https://doi.org/10.1016/J.ENBUILD.2018.03.051.
[44] Brandi S, Fiorentini M, Capozzoli A. Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management. Autom Constr Mar. 2022;135. https://doi.org/10.1016/j.autcon.2022.104128.
[45] Zhang Y, Zhao Q. Energy saving algorithm of HVAC system based on deep reinforcement learning with modelica model. Chinese Control Conference, CCC, vol. 2022-July 2022:5277–82. https://doi.org/10.23919/CCC55666.2022.9901641.
[46] Tuchnitz F, Ebell N, Schlund J, Pruckner M. Development and evaluation of a smart charging strategy for an electric vehicle fleet based on reinforcement learning. Appl Energy Mar. 2021;285:116382. https://doi.org/10.1016/J.APENERGY.2020.116382.
[47] Mansour SH, Azzam SM, Hasanien HM, Tostado-Véliz M, Alkuhayli A, Jurado F. Deep reinforcement learning-based plug-in electric vehicle charging/discharging scheduling in a home energy management system. Energy Feb. 2025;316:134420. https://doi.org/10.1016/J.ENERGY.2025.134420.
[48] Ji Y, Wang J, Xu J, Fang X, Zhang H. Real-time energy management of a microgrid using deep reinforcement learning. Energies 2019;12:2291. https://doi.org/10.3390/EN12122291.
[49] Bao G, Xu R. A data-driven energy management strategy based on deep reinforcement learning for microgrid systems. Cognit Comput Mar. 2023;15(2): 739–50. https://doi.org/10.1007/s12559-022-10106-3.
[50] Franzoso A, Fambri G, Badami M. Deep reinforcement learning as a tool for the analysis and optimization of energy flows in multi-energy systems. Energy Convers Manag Oct. 2025;341:120095. https://doi.org/10.1016/J.ENCONMAN.2025.120095.
[51] Shen R, et al. Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. Appl Energy Apr. 2022;312: 118724. https://doi.org/10.1016/j.apenergy.2022.118724.
[52] Pei Y, Yao Y, Zhao J, Hao J, Ding F, Wang J. Multi-agent hierarchical deep reinforcement learning for HVAC control with flexible DERs. IEEE Trans Smart Grid 2025. https://doi.org/10.1109/TSG.2025.3598082.
[53] Gao H, Zhang G, Xing Q, Yang L. Electric vehicle charging guidance strategy based on hierarchical multi-agent deep reinforcement learning. 2024 IEEE Transportation Electrification Conference and Expo Asia-Pacific, ITEC Asia-Pacific 2024;2024:518–23. https://doi.org/10.1109/ITECASIA-PACIFIC63159.2024.10738543.
[54] Gurobi Optimization LLC, "Gurobi Optimizer Reference Manual," 2024. [Online]. Available: https://www.gurobi.com.
[55] Wang D, Zheng W, Wang Z, Wang Y, Pang X, Wang W. Comparison of reinforcement learning and model predictive control for building energy system optimization. Appl Therm Eng Jun. 2023;228:120430. https://doi.org/10.1016/J.APPLTHERMALENG.2023.120430.