POLITECNICO DI TORINO
Repository ISTITUZIONALE

Deep reinforcement learning as a tool for the analysis and optimization of energy flows in multi-energy systems

Original
Deep reinforcement learning as a tool for the analysis and optimization of energy flows in multi-energy systems / Franzoso, A., Fambri, G., Badami, M.. - In: ENERGY CONVERSION AND MANAGEMENT. - ISSN 0196-8904. - 341:(2025). [10.1016/j.enconman.2025.120095]

Availability:
This version is available at: 11583/3001325 since: 2025-06-27T08:40:39Z

Publisher:
Elsevier

Published
DOI:10.1016/j.enconman.2025.120095

Terms of use:

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

Publisher copyright

(Article begins on next page)

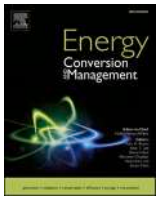
25 June 2026

Energy Conversion and Management 341 (2025) 120095

Contents lists available at ScienceDirect

Energy Conversion and Management

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



Research Paper

Deep reinforcement learning as a tool for the analysis and optimization of energy flows in multi-energy systems

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
Multi-energy system
Energy storage
Energy conversion
Control strategies
Energy management system

ABSTRACT

Deep Reinforcement Learning algorithms not only facilitate the development of optimized control strategies but also serve as powerful tools to explore complex problems and uncover non-obvious control solutions. This paper investigates the application of Deep Reinforcement Learning to optimize a Multi-Energy System in the presence of high Renewable Energy Source penetration. Key energy conversion technologies, such as Combined Heat and Power, Battery Energy Storage Systems, Heat Pumps, and Power-to-Gas, enable bidirectional energy exchanges across different networks, thereby fostering operational synergies. Since these interconnections create interdependencies in which energy flows within one sector significantly affect those in another, the complexity of optimization increases. The aim of this study has been to demonstrate the benefits of a method that can be used to interpret strategies implemented by a Deep Reinforcement Learning algorithm, thereby ultimately increasing the possibility of making optimal decisions. This approach has led to the creation of an optimized rule-based mechanism which has been used to analyze the Multi-Energy System, identify the most advantageous technologies (heat pumps, electric batteries and power-togas, respectively), and highlight the importance of implementing an optimized strategy to achieve effective energy management. Such an optimized strategy led to a reduction in natural gas consumption of about 15%, a decrease in CO_2 emissions of 18%, and a reduction in fuel and electricity costs of 17%.

1. Introduction

Climate change remains one of the most pressing challenges of our time, with its pervasive impacts being felt across the globe [1]. Europe, having recognized the urgency of the situation, is at the forefront in addressing environmental degradation through comprehensive policies and initiatives [2]. The European Union has set ambitious targets to reduce greenhouse gas emissions and has the aim of obtaining a net-zero carbon economy by 2050, a goal that underscores the importance of transitioning to Renewable Energy Sources (RES) [3]. Solar and wind power are key to the energy transition, as they are among the most promising and widely available clean energy sources [4], but their variability and non-dispatchability create challenges for the stability and reliability of energy systems. To address these issues, the Multi-Energy System (MES) approach has emerged as a comprehensive strategy that integrates electricity, heating, cooling, and gas networks, and leverages on synergies across sectors [5]. Unlike traditional methods that often focus on a single energy vector, the MES approach advocates a holistic view that enables the exploitation of synergies between sectors [6]. Technologies such as heat pumps, storage units, electrolyzers, and high efficiency cogeneration, become important in these kinds of systems, and their success depends on the policies that are adopted to control and integrate the various components [7].

The optimal scheduling of MES has traditionally relied on methods that range from rule-based approaches to complex optimization techniques [8]. While rule-based methods offer transparency and reproducibility, they often lack the adaptability required for dynamic environments [8]. On the other hand, mathematical optimization techniques (e.g., LP, MILP, NLP) and *meta*-heuristic methods (e.g., GA, PSO, SA) can yield precise solutions under structured constraints, but they tend to become computationally prohibitive as the complexity of the system or the number of time steps increase, especially when accurate forecasts are impractical. Recent developments in Artificial Intelligence (AI), and Deep Reinforcement Learning (DRL) in particular, offer promising alternatives as they enable real-time decision-making in high-dimensional state and action spaces, without the need for predefined forecast horizons [9].

DRL techniques have evolved significantly over the years, with

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

https://doi.org/10.1016/j.enconman.2025.120095
Received 23 December 2024; Received in revised form 22 May 2025; Accepted 16 June 2025
Available online 26 June 2025
0196-8904/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

continuous advancements in algorithms that have pushed the state-of-the-art in the context of control and optimization [10]. These improvements have led to an enhancement of their stability, sample efficiency, and applicability to complex decision-making problems [11]. Such techniques as Deep Q-Networks (DQN) [12], Proximal Policy Optimization (PPO) [13], Deep Deterministic Policy Gradient (DDPG) [14], and Soft Actor Critic (SAC) [15] enable decision-making in real time and have been shown to outperform traditional optimization and control methods, in terms of both speed and accuracy. The importance of these techniques has been accentuated by the need for the optimization and automation of decision-making processes in energy systems, which are increasingly required to be more adaptive and efficient. Indeed, the review by Perera et al. [9] emphasized the potential of DRL to manage complex energy flows, and it reported that DRL methods have led to significant improvements in energy efficiency and to cost reductions for various applications, such as energy dispatching, building energy management, and renewable energy integration.

Vamvakas et al. [16] conducted a comprehensive review of the recent applications of DRL to optimize energy systems, focusing on RES integration, building energy management systems, and electric vehicle charging stations. The authors concluded that DRL frameworks offer significant benefits, and that they have demonstrated a superior adaptability to dynamic environments and the complex behavior of systems than traditional control approaches.

DRL algorithms have yielded good performances in building energy management systems: from house heating [17], HVAC control (for a single building using DQN [18] and a cluster of coordinated buildings using SAC [19]) to lighting control [20] and the holistic control of all the manageable building elements [21]. Some limited examples of real-world tests and applications exist for HVAC systems [22] that show lower temperature variations from optimality, compared to comparable controllers. DRL methods have also been employed to manage electric vehicle charging stations with the aim of optimizing the charging schedules, peak shaving, and cost reductions. Dorokhova et al. [23] conducted a comparison of DQN and DDPG with other methods, such as rule-based control and MPC, while Zhang et al. [24] proposed LSTM-DDPG to improve data pattern extraction.

Some literature examples can be found concerning the dispatching of energy devices. Abedi et al. [25] developed a real-time control model for Battery Energy Storage Systems (BESS) using an RL approach that focused on optimizing storage management in residential settings using a simple Q-learning algorithm that dynamically adjusted according to the energy demand and pricing. Zhou et al. [26] introduced a specific DRL approach for the operation of Combined Heat and Power (CHP) systems that was aimed at reducing the solution time and at avoiding the need for complicated linearization processes. Guo et al. [27] proposed a real-time, dynamic, optimal energy management model for microgrids that they achieved by comparing DRL algorithms, such as PPO, DQN and DDPG, to control a microturbine and an energy storage system, and they found that it produced a significant reduction in the operation cost and computational burden, compared to such traditional methods as stochastic programming. Ruan et al. [28] used DDPG and Twin Delayed DDPG (TD3) algorithms to optimize the operation of combined cooling, heating, and power systems integrated with renewable energy and energy storage at the building level, and they found that DRL methods outperformed PSO by a significant margin and showed performances close to MILP, as it reduced the computation time during online operations; the TD3 agent improved the performance of the system, compared to DDPG. Zhou et al. [29] developed an improved DRL method that they used to manage a system composed of PV panel wind turbines, BESS, air conditioners, a CHP linked to the electric grid and to the district heating system, and Gas Boilers (GB). The authors used an improved Soft Actor Critic (SAC) algorithm with LSTM networks to obtain an efficient temporal feature extraction. Ceusters et al. [30] compared TD3 and PPO with Linear Model Predictive Control (LMPC) for MES management purposes, and they included such devices as wind turbines, solar PV, CHP units, and battery storage. TD3 showed promise for dynamic and uncertain scenarios and showed a comparable performance with LMPC. Bousnina et al. [31] compared the performance of DDPG and MPC in a digital twin environment of an MES and reported similar results.

Despite the flexibility and scalability of DRL, these algorithms have a common drawback with other optimized control methods: it is not possible to directly identify or extract the underlying optimization logic that the algorithm has discovered. In this respect, DRL works like a "black box" optimization technique. This limitation poses a challenge in terms of interpretability and validation. Therefore, the development of methods to extract, interpret and explain the strategies learned by DRL algorithms is of great value.

- From a research perspective, the ability to extract the optimized control strategies identified by the algorithm enables a deeper understanding of the internal dynamics of the analyzed energy system. It enables the identification of synergies between components and the discovery of relationships that are not obvious a priori. In this respect, such a methodological approach serves as a useful tool for exploring and understanding complex multi-energy systems.
- From an application perspective, the ability to explore the logic discovered by the DRL "black box" optimizer enables users to better understand the solutions found by the algorithm, thus increasing confidence in the correctness of the resulting control strategy [32]. At the same time, not all controllers have sufficient computational resources to directly implement DRL algorithms. The ability to extract rule-based control strategies (with lower computational requirements) therefore makes it possible to apply DRL-derived strategies to resource-constrained controllers.

To the best of the authors' knowledge, only Razzano et al. [33] have explicitly addressed the extraction of rules and the interpretability of solutions generated by DRL, and they applied this methodology to manage an HVAC system in an office building with limited environmental parameters. Leveraging on a rule extraction method from a DRL policy, this approach effectively exploits hidden system dynamics identified during training, and it outperforms the reference benchmark and reduces the complexity associated with the implementation of the algorithm on a physical controller, with respect to DRL. In contrast, the present work applies DRL-based optimization techniques to a more complex, large scale and interconnected MES. The analyzed scenario integrates:

- Renewable energy generation technologies (Wind Turbines – WT, Photovoltaic panels – PV)
- Major energy networks (electricity, District Heating – DH, and gas)
- Energy conversion and storage technologies to facilitate interactions between the different energy sectors (Combined Heat and Power – CHP, Battery Energy Storage Systems – BESS, Heat Pumps – HP, and Power-to-Gas – P2G)

The application of the DRL algorithms uncovered hidden synergies among the system components and thus provided a clearer and more detailed understanding of the optimal operation of the MES. The main contributions of this work are twofold:

- A methodology for DRL interpretability: we propose an approach in which DRL is used as an exploratory tool to analyze the different behaviors of complex systems. This methodology transforms DRL from a black-box optimizer into an interpretable tool by translating its learned policies into intuitive rules. By analyzing the adaptive behaviors and state-action mappings of the DRL agent, it is possible to understand and validate decision-making processes. Additionally, this study highlights the conditions under which DRL achieves optimal solutions and identifies its limitations.

- Analysis of MES optimal strategies: the MES was further analyzed considering the optimal strategies identified by the DRL model. This allowed a clear control strategy to be defined, which, through a sensitivity analysis, helped determine what technologies have the greatest impact on the overall performance of the considered MES.

The paper is structured as follows: Section 2 describes the configuration and mathematical models of the MES, along with the foundational elements of DRL and its specific application to the problem at hand; Section 3 presents and compares the simulation results obtained using both DRL and the derived rule-based strategies, as well as the results of a sensitivity analysis developed using the two rule-based strategies; Section 4 concludes the study by summarizing the main findings and discussing the implications for real-world energy control systems.

2. Methods

This section presents the case study in detail, especially concerning the chosen scenario and the control strategies adopted, focusing on the characteristics of DRL, and specifically on the chosen algorithm (i.e. TD3).

2.1. Description of the scenario

The MES structure and energy scenario were taken from [6] and are detailed in this section. Fig. 1 illustrates the analyzed system, where electricity is sourced from RES (WT and PV) and CHP systems, and batteries are used for storage. Heat is supplied by CHP, HPs, GB, while waste heat is supplied by P2G methanation. The natural gas grid operates as an open system for unrestricted NG transactions. The model simulates a full year in a quasi-steady state but neglects the dynamics and spatial distances of the components. A key aspect of this system is the integration of centralized HPs in district heating, which enhances efficiency and reduces emissions. P2G generates SNG (Synthetic Natural Gas) through a PEM electrolyzer, which produces H_2 and O_2. The obtained H_2 reacts with CO_2 in a methanation reactor to form methane, thereby recovering high-grade heat [34,35]. The SNG can be used in the system itself or sold to the grid, with H_2 being stored in a buffer before methanation.

The presented MES was assumed to be located in northern Italy and to serve a heterogeneous demand consisting of only residential, commercial and industrial users. The district heating grid was instead assumed to serve residential and commercial users. The electricity demand was estimated from National Grid Operator data [36], while the DH demand was derived from the empirical relationship presented in [37]. The nominal capacities of the CHP were designed to be able to cover the maximum electric and heat load peaks. The selected scenario was characterized by a large amount of RES and several technologies capable of flexibly absorbing any random overproductions of renewable energy, namely HP, BESS and P2G. Table 1, Table 2, and Table 3 report the main input data and assumptions pertaining to the selected scenario.

The cost of natural gas was taken from [38] and a carbon tax of 100 €/tonCO_2 was considered for the consumption of fossil fuels. The generated SNG was not considered to have received incentives, and its value was thus assumed to match that of natural gas. Since the MES could buy electricity from the grid, it was considered as a price-taker that did not influence the overall electricity market; the cost of electricity bought from the grid depends on the PUN (*Prezzo Unico Nazionale*, in English Single National Price) which is defined on an hourly basis [38]. As far as the CO_2 emissions of the electricity bought from the grid are concerned, an emission factor of 51.6 g_{CO2eq}/kWh was considered to be coherent with the forecasts for 2040 for the Italian energy system [39]. The cost parameters are summarized in Table 4.

No constraints, such as a minimum capacity factor, were introduced to evaluate how important each technology was for the system, according to the identified optimal strategy. Moreover, the grid connection was also supposed to be sufficiently large to satisfy the electrical energy balance and that GB could balance the thermal demand. This way, the DRL optimizer was free to activate flexible technologies in order to maximize its objective function. However, in a real system, plants obviously have to operate for a certain number of hours to be cost-effective. If the identified optimal solution leads to a low-capacity factor for a given technology, this suggests that this specific technology is unlikely to become a key asset for the analyzed scenario. On the other hand, a technology activated with high priority by the agent is likely to become an important asset for the scenario.

Fig. 2a illustrates the load duration curve of the energy demands. The electricity demand is quite constant throughout the year, whereas the district demand is characterized by higher peaks and seasonal dependence. Electricity demand includes the consumption of passive users, i.e. the electricity consumption of energy storage and conversion technologies (i.e. BESS, HP and P2G) are not considered, and this consumption profile has been called “base load”, while the total electricity consumption has been called “total load”. The load duration curves of the PV and wind generation are reported in Fig. 2b.

A critical parameter in simulations of energy systems is the choice of temporal resolution. Shorter time steps enable more accurate dispatching decisions by capturing the system dynamics in more detail. However, smaller time steps significantly increase the complexity of the calculations. In the context of optimizing energy flows in multi-energy

Table 1
Summary of the electric and heat demands.

Quantity	Total [TWh]	Peak [MW]	Mean [MW]
Electricity demand	3.00	620	342
Heat demand	2.65	1380	302

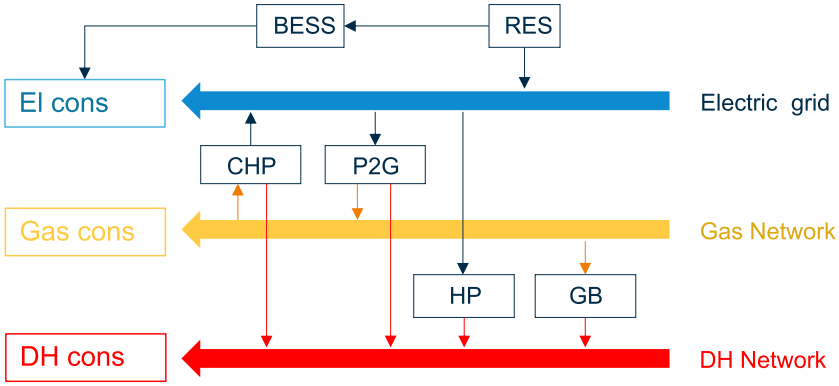


Fig. 1. Scheme of the energy system in which the involved technologies and the interactions between networks are highlighted.

Table 2
Installed capacity of RES technology.

Technology	Nominal capacity [MW_{el}]	Producibility [TWh/y]	Full Load Hours
WT	1000	2.6	2583
PV	750	0.92	1226

Table 3
Installed capacity of the energy storage and conversion technologies.

Technology	Nominal capacity [MW_{el}]	Unit of measurement
HP	170	MW_{el}
BESS	1000	MWh_{el}
P2G	200	MW_{el}
CHP	620	MW_{el}

Table 4
Economic parameters.

Parameter	Value	Ref
Purchase cost of NG	40 €/MWh	[38]
Selling price of SNG	40 €/MWh	Assumption
Carbon tax	100 €/tonCO_2	Assumption
Purchase cost of elec.	Variable (Based on PUN)	[38]
Grid carbon intensity	51.6 g_{CO2eq}/kWh	[39]

scenarios, such as the one analyzed in this study, a time discretization of one hour is usually chosen [40]. This resolution represents a balanced compromise: it ensures sufficient accuracy in representing the interactions between the different subsystems without causing an excessive computational burden. At this resolution, fluctuations in renewable generation and energy demand are effectively captured and it is acceptable to neglect the fast internal dynamics of all individual components except the P2G whose methanation unit exhibits dynamic behavior with time constants exceeding one hour. However, assuming that there is a hydrogen buffer separating the operation of the methanation unit from that of the electrolyzer, this hypothesis is also confirmed for this component (see Section S1.4 of the Supplementary Material for more details). Multi-energy systems are inherently heterogeneous and include subsystems such as the electricity grid, BESS, district heating network, heat pumps and P2G plants. These subsystems differ significantly in their response times. For example, accurately capturing the dynamics of the electricity grid would require a much finer temporal resolution than that required for modeling the operation of a heat pump. If the goal is to evaluate the dynamic responses of all individual components, a viable approach could be co-simulation [41,42]. With this method, each subsystem can be simulated in its own tailored environment, using time step best suited to its dynamic behavior. However, such an approach is beyond the scope of this study. In line with the prevailing scientific literature, a uniform one-hour time step was chosen to focus on energy flow optimization rather than detailed dynamic modeling.

The conversion efficiency values of the different technologies considered in this study are summarized in Table 5. The mathematical model of each energy conversion and storage technology is provided in the Supplementary Material in Section S1.

2.2. Rule-based control algorithm

The scenario described in the previous section was initially addressed using the fixed rule-based control system presented in [6]. In this system, energy conversion and storage technologies are employed to satisfy the energy demands of the scenario and optimize the use of RES. Multiple technologies can be used for the same purpose; for example, heat can be provided by HP, CHP or GB units. The dispatching priority is based on the energy conversion efficiency of each technology, which means that those with the highest efficiency are prioritized. This efficiency-based dispatching logic was used to manage the surplus RES electricity, to meet the electricity demand, and to address the heat demand of the scenario. Fig. 3 shows the flowchart of the control logic that is applied. Specifically, when renewable electricity generation exceeds the demand of the system, the surplus energy is utilized by energy conversion and storage technologies, in descending order of efficiency. The most efficient technology, HP, is used first to cover the heat demand. If any surplus electricity remains after being converted by heat pumps, it is stored in the BESS systems. Once the BESS systems reach full capacity, any remaining surplus electricity is directed to P2G plants. Only after all these options have been exhausted is any residual surplus electricity curtailed.

The electricity demand is first met through RES generation. If RES is insufficient to cover the demand of the scenario, the energy stored in the BESS is used. Any remaining unsatisfied demand is supplied by CHP plants. Therefore, the primary role of CHP plants is to balance the electricity system, while heat is treated as a secondary product.

Table 5
Efficiency of the implemented components.

Technology	Efficiency	Ref
CHP	$\eta^{el}_{chp} = 0.5$; $\eta^{th}_{chp} = 0.4$	[43]
P2G	$\eta^{SNG}_{P2G} = 0.6$; $\eta^{th}_{chp} = 0.12$	[44,45]
BESS	$\eta^{dis}_{BESS} = \eta^{char}_{BESS} = 0.92$; $C_{rate} = 0.25$	[6]
HP	$COP_{HP} = 2.7$	[46,47]
GB	$\eta^{th}_{GB} = 0.90$	[6]

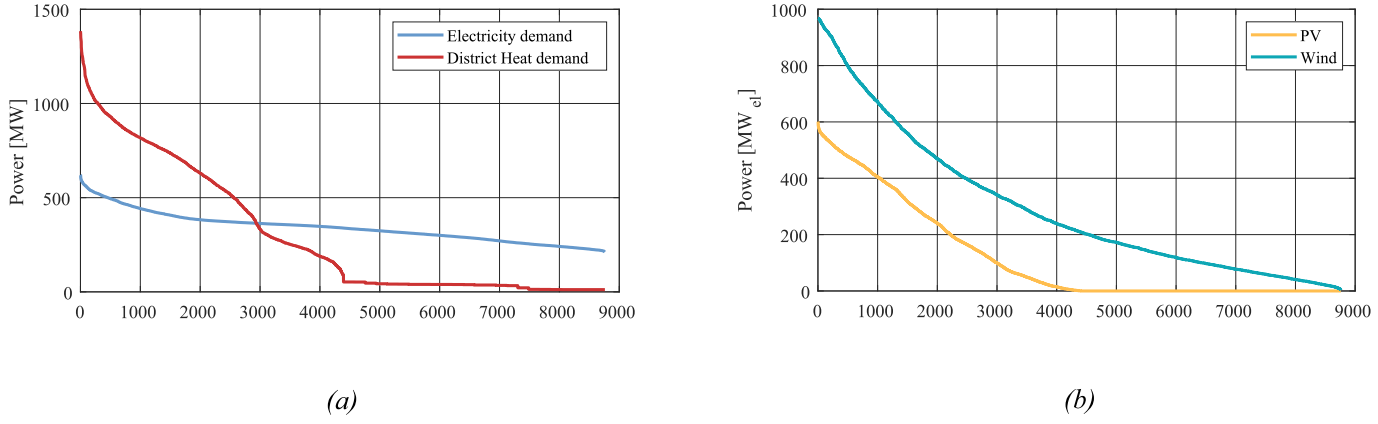


Fig. 2. (a) Load duration curves of the energy demands (electricity and district heating), and (b) Duration curve for WT and PV generation.

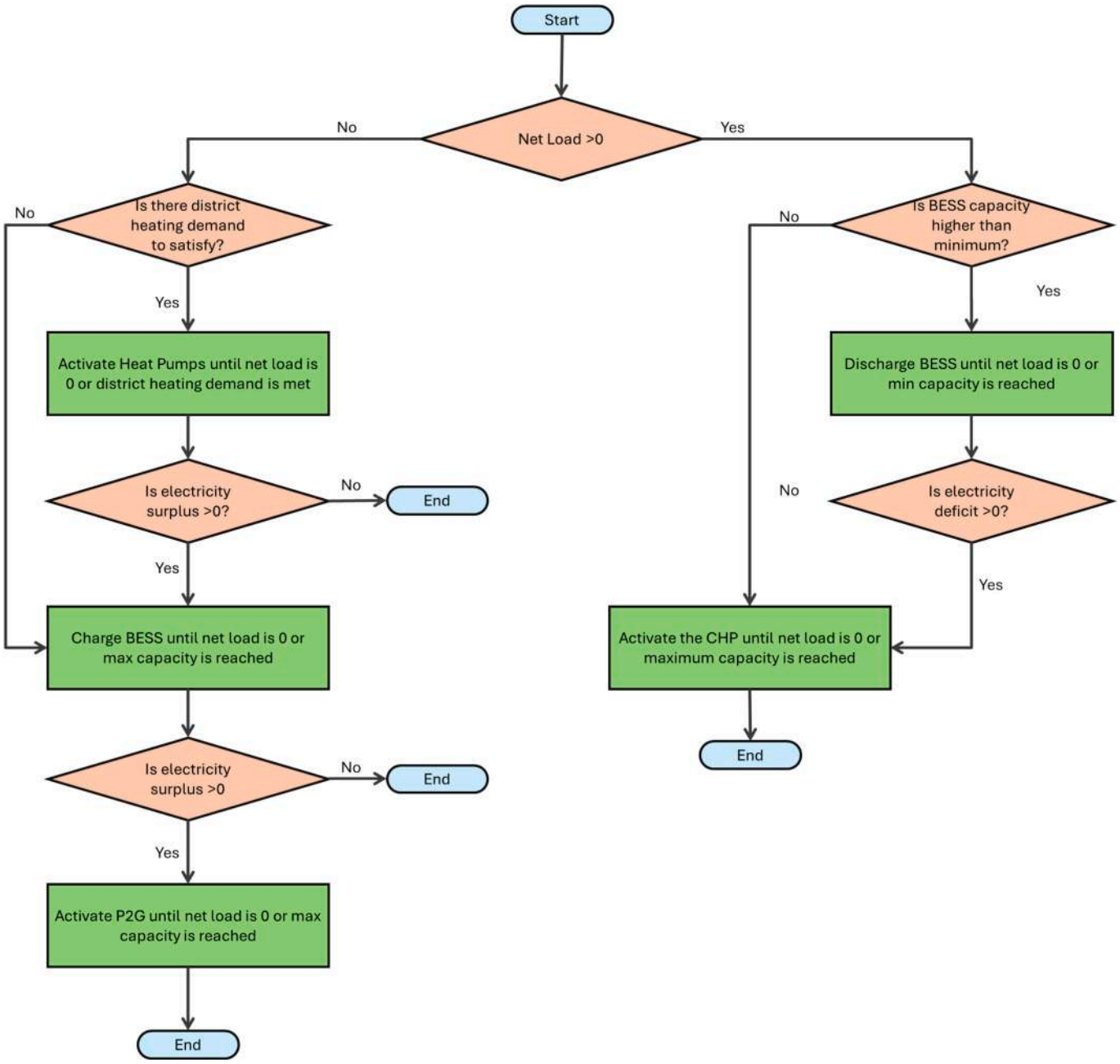


Fig. 3. Flowchart of the rule-based strategy highlighting both the dispatch priority for the electricity and district heating demand.

Heat supplied to district heating users can be generated by HPs, CHP plants, GBs, or P2G units. Priority is given to HPs, which convert renewable electricity into heat. If the heat generated by HPs is insufficient to meet the demand, the heat recovered from P2G plants, produced during the methanation phase, is utilized. Any remaining heat demand is satisfied by resorting to cogenerated heat from CHP and GB plants.

2.3. Deep reinforcement learning algorithms

Several different choices can be made, in the context of developing an optimized new rule-based model, on the optimization method used. Classic optimization methods, such as MILP or *meta*-heuristic methods can be used. However, MILP requires the formulation of constraints and such objective functions as linear equations and inequalities, and linearity can limit the ability of the model to capture the inherent complexity of exploring large and non-linear solution spaces, thus making MILP less suitable when the problem structure cannot be approximated clearly by means of linearization. On the other hand, DRL solutions do not require any linear approximation. This makes them advantageous, since they can be applied to both problems that can be approximated well linearly and to those for which such an approximation is not possible. In addition, MILP methods may require significant computational resources to solve optimization problems, while DRL models, once trained, can provide solutions faster [28]. Meta-heuristic methods are often effective for the navigation of poorly understood or highly complex solution spaces, but their performance tends to deteriorate as the number of optimization variables increases, a challenge that becomes significant when it is compounded by multiple control variables over numerous time steps. DRL offers a promising alternative as it iteratively interacts with the environment to learn dynamic control policies, and it was inherently designed to manage high-dimensional state and action spaces, thus making it suitable for complex, large-

scale problems. A DRL agent can be trained directly within a simulated representation of the physical system, thereby enhancing both its learning process and the control design process. However, DRL methods share some limitations with *meta*-heuristic methods, including the challenge of hyperparameter tuning and the lack of guarantees pertaining to achieving a global optimum, although they have shown better performances [28]. For this reason, a post-processing evaluation of the black-box solution found by the DRL agent is necessary to resolve any possible inaccuracies (see Section 3.2).

2.3.1. Definition of the Markov decision process

DRL problems are modeled as Markov Decision Processes (MDP) and are defined by three main elements, which are:

- S is a finite set of states of the environment;
- A is a finite set of actions;
- $R_a(S,S')$ represents the reward received after A transitions the environment from S to S';

The S state contains information that can be used to uniquely represent the configuration of the environment at time t. An agent chooses an action, A, which is performed, and the environment then reacts by returning the new state, S', and the reward, R. The agent learns what discounted rewards result from the executed actions through a trial-and-error learning process: the trial phase involves selecting an action to execute within the environment, while the error phase corresponds to the received feedback reward, which is subsequently used to update the policy. The policy function, which is the core component of the optimization, maps the relationship between the states and actions within a probability distribution, thereby ultimately guiding the decision-making process. Specifically, policy π represents a conditional probability distribution that defines the likelihood of each possible action, given a particular state. Fig. 4 shows the application of this framework to this particular case study.

The agent observes the environment at each time step t, through the following quantities:

$$S_t = \begin{bmatrix} Net\ load(t),\ Heat\ demand(t), \\ electricity\ price(t),\ SOC(t),\ Hour(t) \end{bmatrix} \tag{1}$$

The S_t state consists of two parts: one is the endogenous part, which is a consequence of the previous action (SOC), while the other is the exogeneous part, which is global information (energy demands, renewable production, and electricity price), which is not impacted by the agent's actions. Since the trends of energy demand, renewable production, and electricity price, to a certain extent, follow a daily trend, the hour of the day was also used as an observation.

The agent regulates the power level of the CHP (a_{CHP}), the HP heat production (a_{HP}), the BESS charge and discharge rate (a_{BESS}), and the SNG production (a_{P2G}) at each time step, t. All the actions were normalized over the [-1, 1] range and then scaled back to an appropriate value to improve the stability of the neural networks.

$$A_t = [a_{CHP}(t),\ a_{HP}(t),\ a_{BESS}(t),\ a_{P2G}(t)] \tag{2}$$

As far as the BESS, whose actions are subjected to the constraints specified in Section S1.3, is concerned, the agent's a_{BESS} output is limited to p_{BESS}.

The reward is the environmental feedback to the agent, and it is the objective function that has to be maximized, with its design playing a critical role in the agent's training: in this case, the main goal is to satisfy both the heat and electricity demand, while minimizing the operating costs and, at the same time, running within the bounds. Therefore, the reward structure includes a real-valued element $r_1(t)$, which corresponds to the true objective function, and two additional terms, $r_2(t)$ and $r_3(t)$, designed as soft constraints of the system to guide the agent's behavior effectively while ensuring it acts within the feasible action space.

$$R_t = r_1(t) + r_2(t) + r_3(t) \tag{3}$$

The first part, $r_1(t)$, is indicative of the economic performance of the system, and it accounts for the cost of electricity and natural gas. Since the aim of DRL is to maximize the reward, costs are represented as negative values. Because of the involvement of neural networks, it was necessary to reduce the magnitude of the first term in order to keep the reward values within an acceptable range for the neural networks.

$$r_1(t) = -\frac{|C_{NG}(t) + C_{el}(t)|}{\delta_1} \tag{4}$$

where $C_{NG}(t)$ refers to the total gas costs and $C_{el}(t)$ to the total electricity costs.

The total gas costs are calculated as follows:

$$C_{NG}(t) = c_{NG} \cdot \dot{m}_{NG}(t) \cdot LHV_{NG} + CO_2^{em}(t) \cdot Carbon\ tax \tag{5}$$

where c_{NG} is the unit cost of natural gas [€/MWh], $\dot{m}_{NG}(t)$ is the net amount of natural gas bought (if positive) or sold (if negative), and CO_2^{em} refers to the total gas emissions (see Section 2.1). Therefore, the term $\dot{m}_{NG}(t)$ also accounts for SNG production.

The cost of the electricity bought from the grid is calculated as follows:

$$C_{el}(t) = c_{el}(t) \cdot P_{grid}(t) \tag{6}$$

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

$r_2(t)$ is a penalty term that is introduced to address two issues: the agent's lack of inherent knowledge about battery constraints (e.g.,

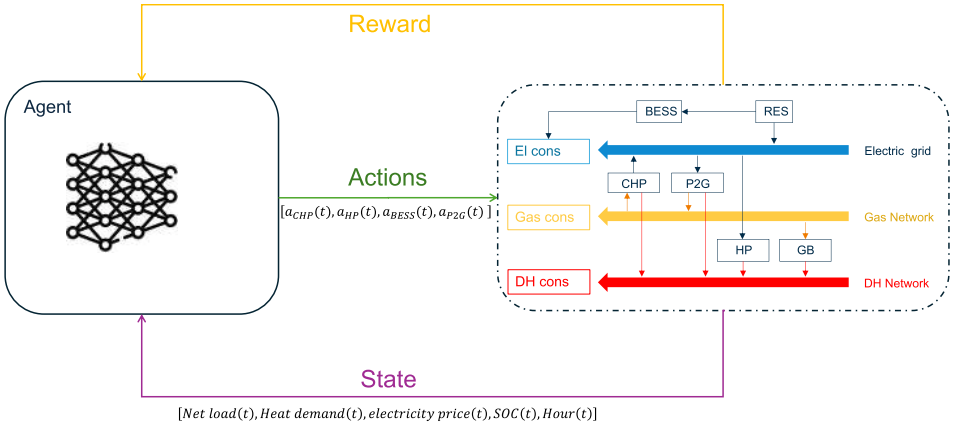


Fig. 4. Diagram of MES optimization with DRL.

charging/discharging limits and capacity) and the temporal sparsity of battery-related rewards. It discourages non-executable actions and misaligned decisions, and it guides the agent toward feasible and efficient battery management practices. Similar solutions were tested in previous research [28,31] to penalize actions that caused the SOC to move outside the permissible range, and this work has used the same approach:

$$r_2(t) = -|\delta_2 \cdot (a_{BESS}(t) - p_{BESS}(t))| \quad (7)$$

where δ_2 is the weight attributed to the penalty, a_{BESS} is the action ordered by the agent, and p_{BESS} is the action that is actually executed in the environment.

Similarly, a small penalty, $r_3(t)$, was also introduced to prevent excess electricity dissipation in the HP, but this is not directly penalized in the $r_1(t)$ definition:

$$r_3(t) = -\left| \delta_3 \cdot \frac{\left(\dot{Q}_{HP}(t) - \dot{Q}_{dem}(t)\right)}{COP_{HP}} \right| \quad if \quad \dot{Q}_{HP} > \dot{Q}_{dem} \quad (8)$$

where δ_3 is the weight given to the penalty, $\dot{Q}_{HP}$ is the total heat produced by the HP, and $\dot{Q}_{dem}$ is the total heat demand.

The δ_2 and δ_3 terms were obtained iteratively to prevent the agent from developing suboptimal policies due to excessive bias and an overemphasis on penalty avoidance. Table 6 summarizes the values that were obtained.

2.3.2. Comparison and validation of deep reinforcement learning algorithms

In this study, two state-of-the-art DRL algorithms, that is, DDPG and TD3, were implemented and compared. Both methods follow an actor-critic framework and operate in a continuous action space, with TD3 introducing key improvements that mitigate the overestimation of Q-values observed for DDPG. Details of their implementation, including the network architecture, training processes, and hyperparameter selection are provided in the Supplementary Material section (Section S2.1). The training methodology involved splitting the dataset into distinct training and testing subsets and then employing a shuffled episodic strategy to enhance generalization and limit overfitting. The training progression and evaluation metrics showed that the two algorithms achieved comparable performances, with TD3 demonstrating, as expected, improved stability. Since the two algorithms performed in a very similar manner after training, only the results obtained using the TD3 agent are presented in the next section. A comprehensive discussion of the training setup, reward progression, and performance analysis can be found in the Supplementary Material section (Section S2.2).

The trained model was tested separately on the training and test sets to assess any potential underfitting or overfitting. Underfitting occurs when a model is too simple to capture data patterns, and this leads to poor performance of both sets. Overfitting, on the other hand, results in excellent training performance, but also poor test results, due to an excessive noise fitting.

Thus, in order to rule out these issues, each agent was trained twice, and the training and test sets were switched for comparison purposes. Table 7 presents the economic rewards obtained with both training configurations. They show minimal differences, thus confirming the absence of overfitting. The consistency of the results obtained when the sets were swapped indicates that the training period involved slightly higher costs than the test period, but this variation is unrelated to the overfitting or underfitting phenomenon. Additional information can be found in the Supplementary Material section (Section S2.3).

Table 6
Penalty parameters δ_2 and δ_3.

Parameter	Value	Unit of measurement
δ_1	1000	/
δ_2	0.25	/
δ_3	50	€/MWh

Table 7
Fuel and electricity costs obtained by the DRL agent over different portions of the dataset.

Fuel and electricity costs [M€]	Original splits	Inverted splits
Training set (even weeks)	66.9	67.7
Test set (odd weeks)	63.3	63.5
Full year	130.0	131.0

2.4. Analysis, interpretation and rule-extraction

DRL algorithms are not inherently interpretable, as they lack explicit mechanisms for weighting input features. Therefore, well-known interpretability techniques, such as Feature Importance Analysis (FIA) [48], cannot be directly applied to them. To overcome this limitation, surrogate models are used to approximate and analyze the decision process of the DRL agent [49]. In this work, an enhanced version of decision trees, Random Forests (RF), was used as a surrogate for the DRL algorithm. The RF model was trained using the observations (see Section 2.4.2) as features and the relevant technologies' actions (see Section 2.4.2) as outputs. The performance of the surrogate model was verified by evaluating the R2 value, which exceeded 0.98, demonstrating that the RF model accurately captures the DRL policy [48]. An FIA was then conducted with the RF surrogate models to correlate the importance of each observation with action selection. State-action correlation plots were also used to highlight the interactions between the energy conversion technologies and their respective activation orders, providing a complementary perspective to the FIA results. These plots allow a direct visualization of how specific system states, such as renewable generation levels, storage state-of-charge, or thermal demand, affect the control decisions taken by the DRL policy. By analyzing these correlations, it is possible to detect operational patterns, discover potential inconsistencies or identify unexpected behaviors in the learned strategy. This combined approach, leveraging both FIA and state-action correlation analyzes, enhances the interpretability of the DRL policy by offering both global (feature importance) and local (state-action relationships) insights. The findings from these analyzes served as the basis for the formulation of an optimized rule-based strategy designed to emulate the DRL policy while being transparent and easy to implement in the practical operation of the energy system (see Fig. 5).

3. Results and discussion

This section presents a comparative analysis between the DRL-based strategy and the simple rule-based strategy. The focus is on the energy and economic KPIs, the FIA scores and the most relevant state-action correlation plots. Further details can be found in the Supplementary Material. Building on this comparison, the development of an optimized rule-based strategy derived from both the original rule-based approach and the DRL policy is described, followed by a performance comparison with the DRL-trained strategy. The FIA results and state-action correlations were analyzed to investigate and interpret any discrepancies. Finally, a sensitivity analysis of the capacity factors of the different energy conversion and storage technologies was performed using the newly developed deterministic strategy.

Although only the results obtained using the TD3 algorithm are reported in this section, a simulation using the DDPG algorithm was also performed and the results are reported in the Supplementary Material. Since linear optimization methods are used in the literature (where applicable) [29,30,31] as a standard for benchmarking DRL solutions, in the Supplementary Material, the results obtained using a traditional

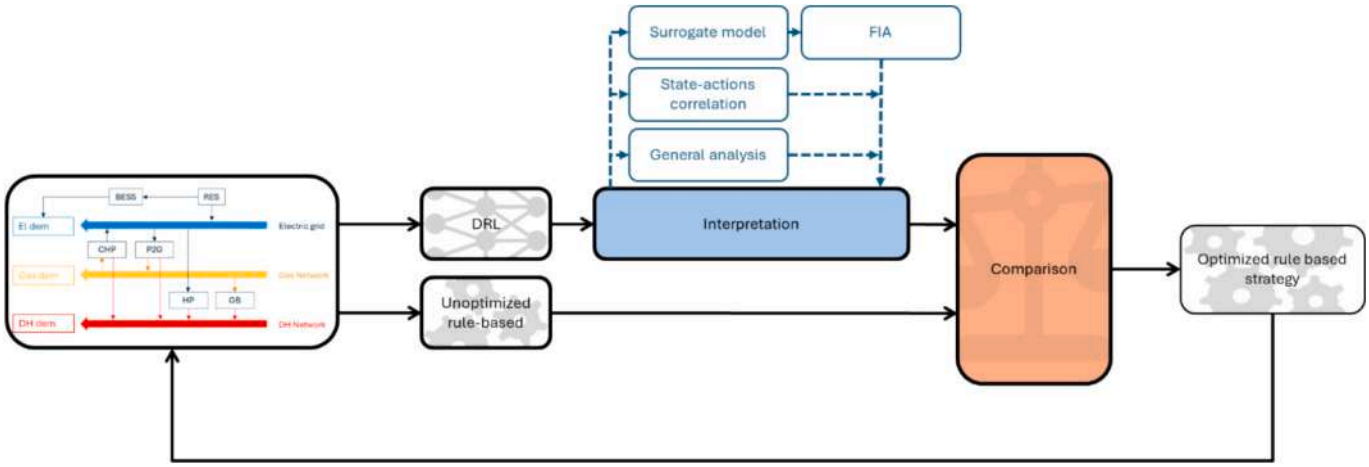


Fig. 5. Methodological flowchart.

MILP optimization [50] were reported. This step is taken to ensure that the DRL approaches achieve a solution that is close to the mathematically optimal one. The DRL results were very similar to those of the MILP optimization, confirming the reliability of the implemented algorithms. MILP optimization is expected to outperform the DRL approach because MILP works under the assumption of perfect foresight of all future demand and production profiles and system conditions over the entire optimization horizon. Such perfect knowledge cannot be achieved in real applications. Therefore, the MILP solution represents a theoretical upper limit. In contrast, the DRL agent makes its decisions based solely on current observations and past data, without access to future information, which limits its ability to perfectly predict upcoming variations.

3.1. Comparative analysis and findings

The DRL strategy was compared with the best strategy found in [6] in which a rigid dispatch and supply priority was considered. The chosen dispatch order exploited storage/conversion energy technologies based on decreasing efficiency: HP, BESS and, finally, P2G. Table 8 reports some of the key parameters and compares the performance of the DRL agent with the rule-based mechanism; the differences were found to be sizeable. The DRL model maintained the gas consumption of the CHP at around 1.8 TWh, which is higher than the 1.6 TWh observed for the rule-based approach. However, the GB consumption was reduced by 60%, and this led to a significant decrease in CO_2 emissions for the DRL strategy. The emissions were 426 kt, compared to the 524 kt observed for the rule-based system. Additionally, the economic analysis demonstrated that the DRL agent incurred fuel and electricity costs of approximately €130 million, which was considerably lower than the cost of €156.4 million associated with the rule-based method.

Table 8
Comparison of the results over the entire year obtained using the DRL agent compared with the rule-based approach adopted in [6].

Quantity	UoM	Rule-based	DRL Agent	Rel. Diff. [%] (Rule-Based)
Gas Consumption (CHP)	**[TWh]**	1.28	1.82	42.2%
Gas Consumption (Boiler)	**[TWh]**	1.61	0.63	−60.9%
Gas Consumption (Total)	**[TWh]**	2.88	2.44	−15.3%
SNG production	**[TWh]**	0.28	0.34	21.4%
Heat from HP	**[TWh]**	0.94	1.58	68.1%
BESS (absorbed)	**[TWh]**	0.12	0.06	−50.8%
Electricity from the grid	**[TWh]**	0	0.04	–
Renewable curtailment	**[TWh]**	0.41	0.39	−4.9%
CO_2 emissions	**[kton]**	524	426	−18.5%
Fuel and electricity costs	**[M€]**	156.4	130.0	−16.9%

Fig. 6, Fig. 7 , and Fig. 8 show the energy flows (electricity generation, electricity demand, and district heating, respectively) of a characteristic week taken from the test set, and they compare the rule-based strategy with the one identified by the DRL agent. Some key differences can be observed concerning the technologies involved:

- Fig. 6 shows different utilization patterns for CHP (blue area). In the case of rule-based operation, the electricity production of the CHP never exceeds the base load. In other words, the electricity produced by the CHP is never used to power storage or energy conversion technologies. It can be noted that the electricity produced by the CHP increases in the DRL strategy.
- Fig. 7 shows different utilization patterns for HP, which is used more in the DRL strategy (light green) and in the BESS charging phase (purple);
- Fig. 8 confirms the previous observations about HP (light blue) and CHP (brown), and it shows their impact on the district heating sector, proving that both are used more in the DRL strategy.

Even though it is possible to draw some hypothesis at this stage as to why the DRL strategy showed much better results, a more detailed analysis of this comparison was conducted using methods that are generally adopted for the analysis of Machine Learning models, but which have here been adapted to DRL.

As discussed in Section 2.4, DRL models are black-box systems, and they thus lack inherent interpretability. To analyze the DRL strategy, FIA was conducted to quantify the impact of features (i.e., observations). The FIA of both the rule-based strategy and the DRL agent are reported in Fig. 9, but, for the sake of completeness, additional information is presented in the Supplementary Material section (Section S4). The FIA graphs show the dependency of the state of an observation on the action of one (or more) controlled devices. It is worth noting that these graphs show both direct and indirect dependencies. In the case of the rule-based algorithm, the operation of the BESS is not directly controlled by the heat demand. However, a higher heat demand leads to increased utilization of the HPs, which, in turn, absorb more energy from renewable sources. As a result, the amount of excess renewable energy that could be stored in the BESS is reduced, thus creating an indirect dependency of the heat demand on BESS utilization.

When the FIA graphs of the rule-based solution and the solution found by DRL are compared, similarities and differences can be observed. In both cases, the net load has a marked influence on CHP, BESS, and P2G. In the DRL case, the dependency of the net load on the heat pumps largely decreases, while the dependency on the heat demand

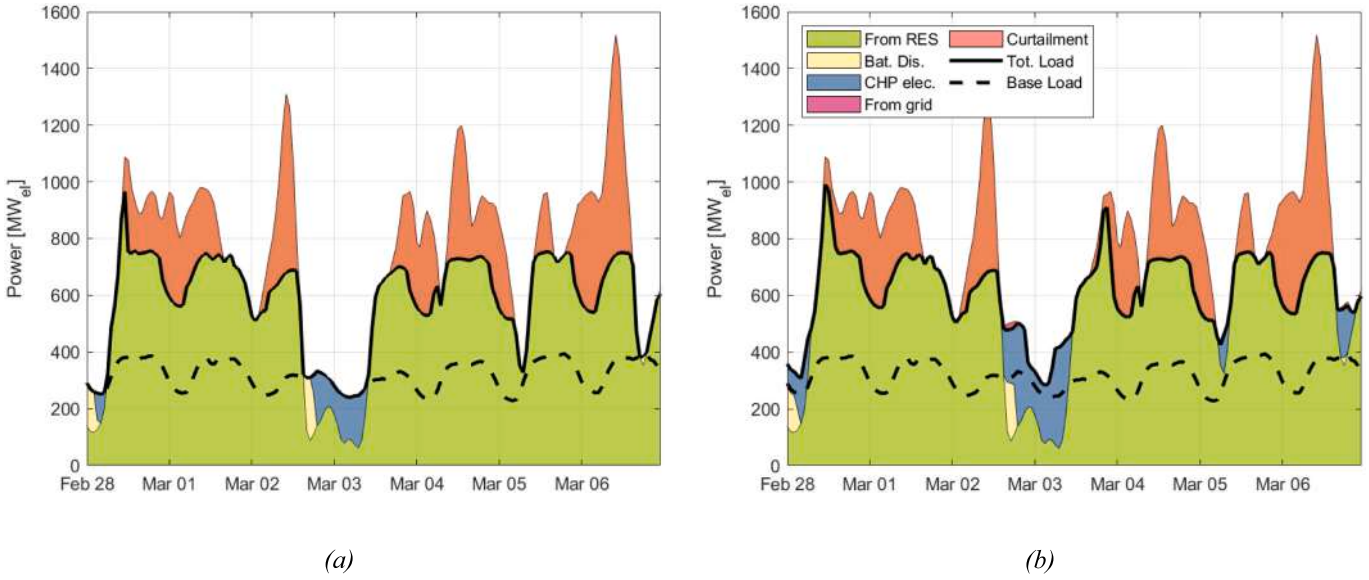


Fig. 6. Comparison of power management strategies of (a) the Rule-Based algorithm, and (b) the DRL algorithm for the period from February 28th to March 6th, showing the energy contributions of the CHP, RES and the grid, battery operation, and the load demands.

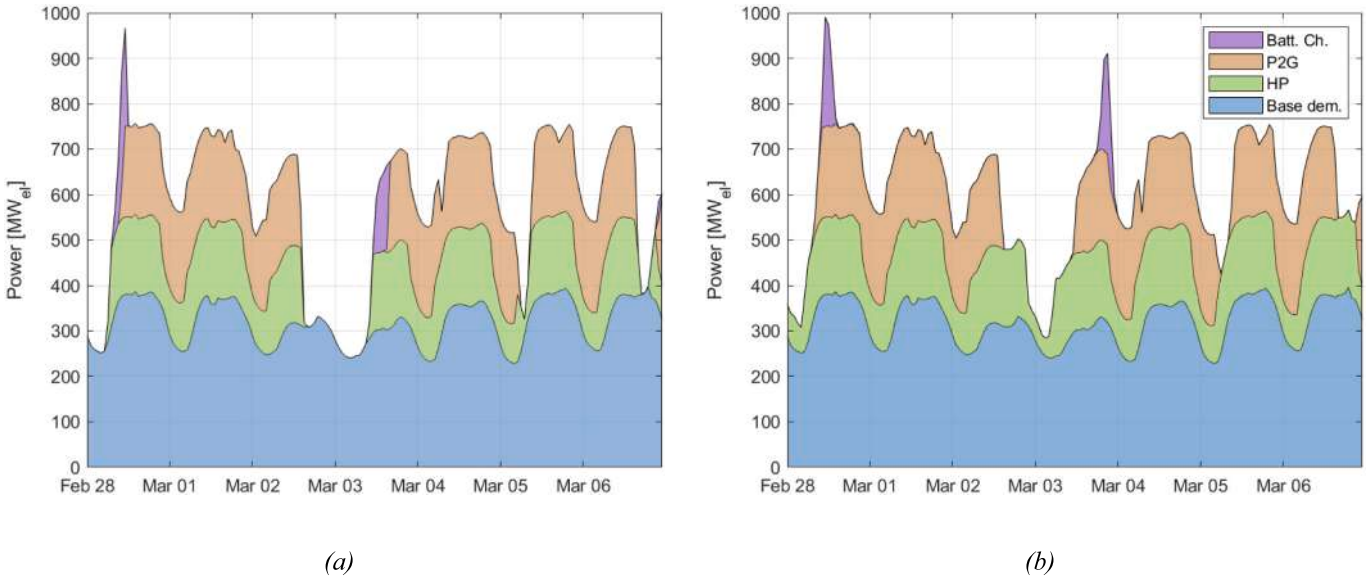


Fig. 7. Comparison of the electric consumption of (a) the Rule-Based algorithm, and (b) the DRL algorithm for the period from February 28th to March 6th, showing the contributions of the HP, BESS and P2G systems to the baseline energy demand.

increases; a similar, albeit less pronounced trend, can be seen for CHP. In the rule-based case, the various technologies were controlled on the basis of the net load value. This discrepancy shows that the heat demand plays a greater role in the optimal control of the system. It can also be seen that P2G depends on the BESS state of charge in the rule-based case. This is because P2G only works when BESS is completely full, due to the dispatching priority that is implemented in the rule-based case. This effect is not observed in the DRL, which indicates that the strategy followed by the DRL does not involve the same priority order as BESS and P2G. Finally, it should be noted that DRL also shows a slight dependence on electricity costs, especially for battery use. However, this dependence is negligible compared to that on the net load: because of the high RES overproduction in this scenario, it is more convenient to use BESS to exploit the RES overproduction than to follow electricity cost variations.

In order to further understand the relationship between the actions and observations, we analyzed the obtained state-action plots, in particular concerning those components, such as CHP and HP (Fig. 10) or BESS and P2G (Fig. 11), which showed very different behaviors from those of the rule-based strategy.

Fig. 10b-d shows that the DRL agent allows cooperation to be achieved between CHP and HP, since the two technologies are often operated at the same time. This effect can also be noted by comparing Fig. 6a, Fig. 7a, and Fig. 8a with Fig. 6b, Fig. 7b, and Fig. 8b, which represent the energy flows for a characteristic week of the test set: in fact, it can be seen that the difference between the two approaches mainly concerns the management of the district heating system: in Fig. 6b, the total generated electricity (RES in green, CHP plants in blue, and BESS in yellow) was higher than the demand during the night of 28th February. This is because the extra amount of produced electricity was used by the HP plants (see Fig. 7b in light green) to generate heat for the DH (Fig. 8b in light blue). In the case controlled by rule-based rules, whenever the CHP did not produce enough heat, the GBs were activated to compensate

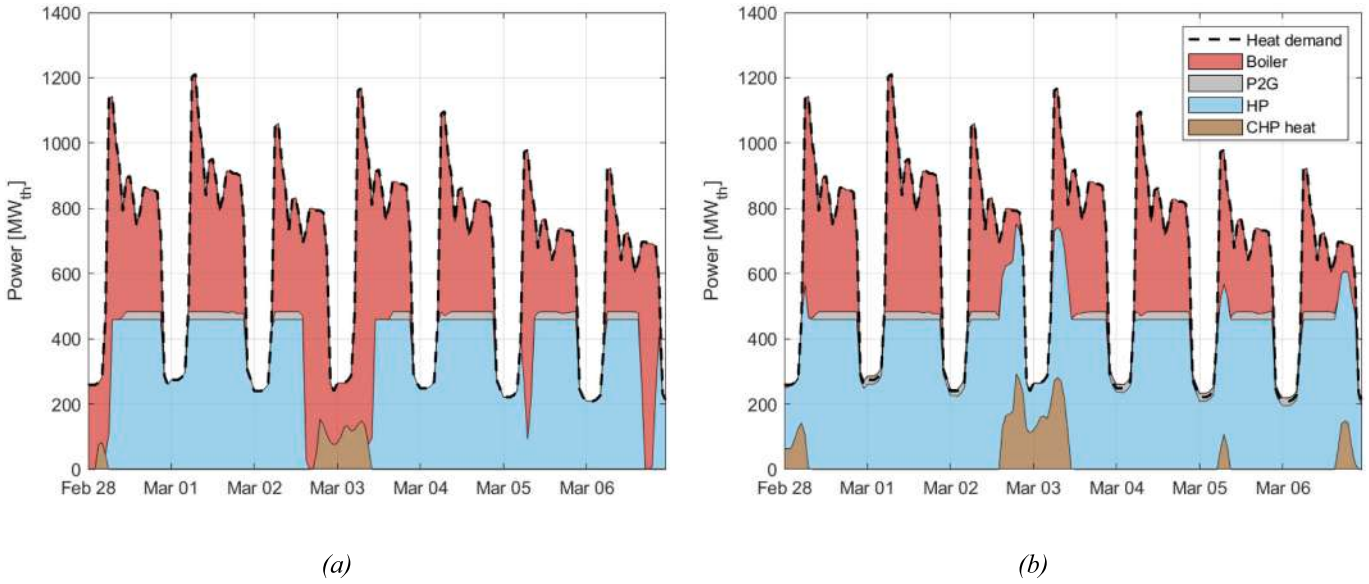


Fig. 8. Comparison of the heat management of (a) the Rule-Based algorithm, and (b) the DRL algorithm for the period from February 28th to March 6th, showing the contributions from the boilers, HP, P2G, and CHP systems to the overall heat demand.

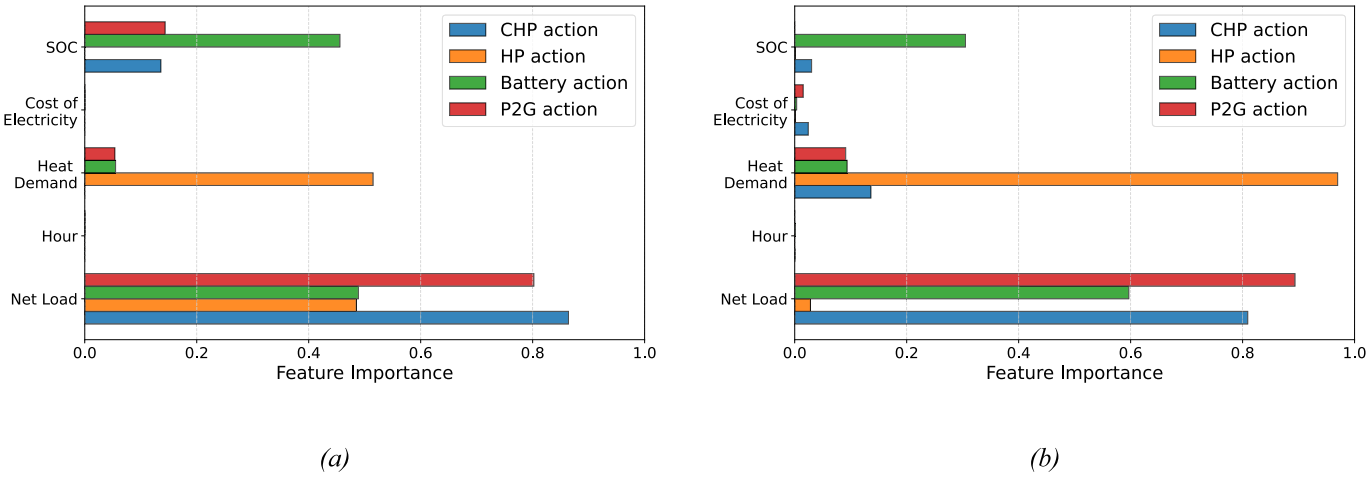


Fig. 9. The relative FIA results for (a) the rule-based strategy, and for (b)the implemented DRL agent strategy.

for the shortage. If we consider 1 MWh to have been produced by the GBs, the cost is €40. However, in the case of DRL, the 1 MWh of thermal energy would be covered by increasing the CHP load, and this would allow the HP to be activated. In this case, the CHP would produce 0.23 MWh of heat and 0.29 MWh of electricity that the HP would convert into 0.77 MWh of heat, for a total cost of NG of about €34, therefore producing savings of about 49% per unit of heat with respect to the utilization of gas boilers, as used in the rule-based mechanism.

The BESS and P2G utilizations show somewhat different trends, as a higher priority is assigned to BESS charging than to P2G methane production for the rule-based mechanism. This can be observed in Fig. 11a–b: the DRL agent only charged the BESS when the P2G was already powered with the maximum power. A difference in the use of these two technologies can also be seen in Fig. 7: a different placement of the purple area (which represents the energy absorbed by the BESS) for Fig. 7a and Fig. 7b. In the case of DRL control, the P2G was operated before BESS was charged; in this way the electrical energy stored in the BESS was reduced and the production of SNG was increased (see Table 8). However, considering the efficiencies and costs of energy carriers, storing energy in the BESS has proved to be more advantageous than producing SNG. Indeed, 0.85 MWh of electricity can be generated for every 1 MWh of energy stored in the BESS, and this leads to a reduction of the electricity production costs of approximately €100, depending on the scenario. If that 1 MWh were used in the P2G process, it would produce 0.6 MWh of SNG (for a value of €36). Therefore, in this case, the DRL algorithm fails to find the optimal solution. This suboptimal effect can be attributed to the agent's focus on maximizing the discounted reward during training, which may result in a shortsighted approach that fails to fully consider the value of the energy stored in the BESS. This stored energy can only be discharged several hours after the charging phase, and the agent may therefore struggle to find a connection between the action of having charged the battery at a certain moment and the benefit of using the stored energy at a future moment [28].

3.2. *Optimized rule-based algorithm*

The DRL agent implemented more efficient operational strategies than those used by the rule-based control algorithm presented in [6]. The superior performance of the DRL algorithms was due to the fact that

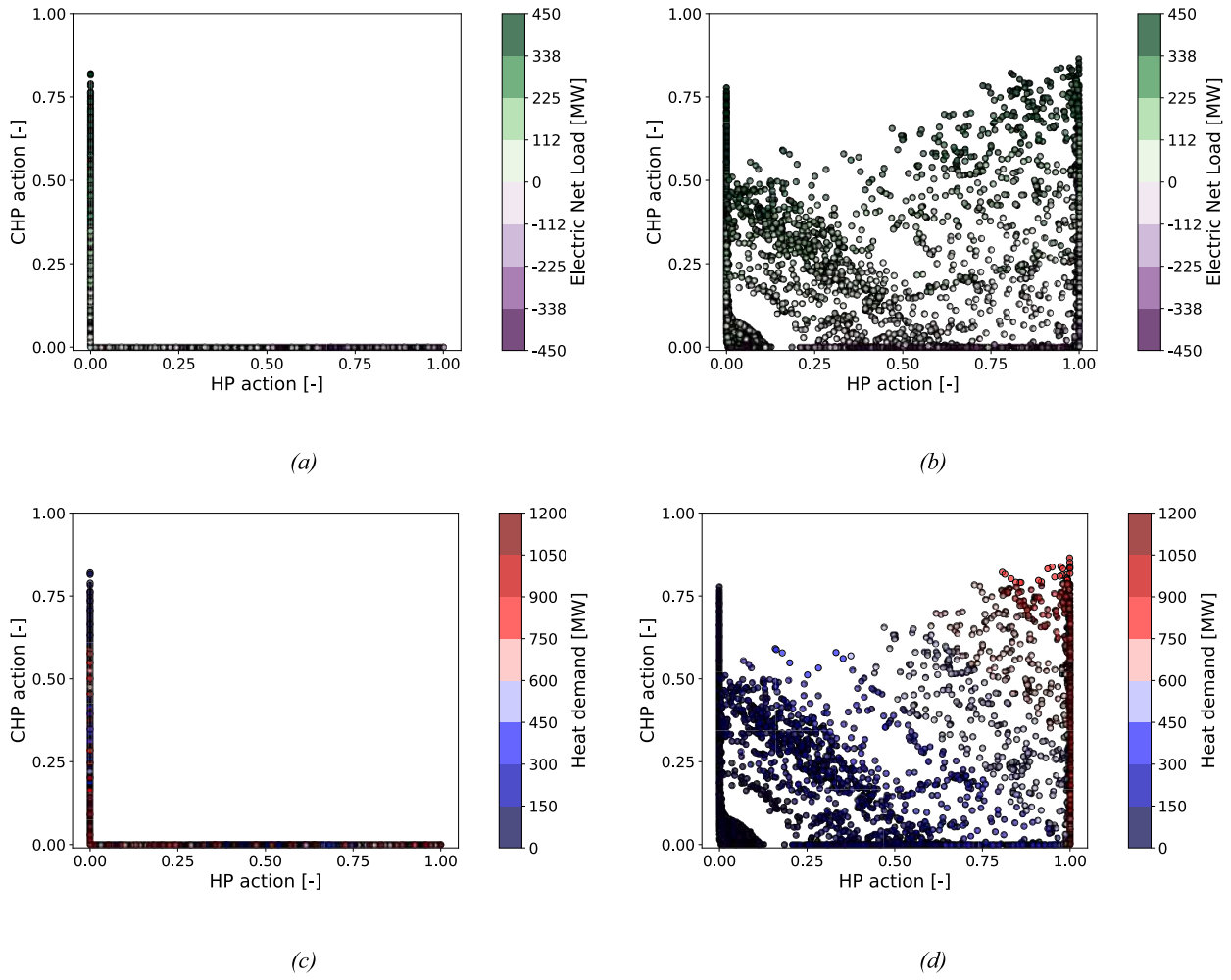


Fig. 10. State-action correlation with focus on the CHP-HP interaction: (a, c) correspond to the rule-based strategy, while (b, d) depict the DRL strategy. The colors represent the net load (a, b) and the heat demand (c, d).

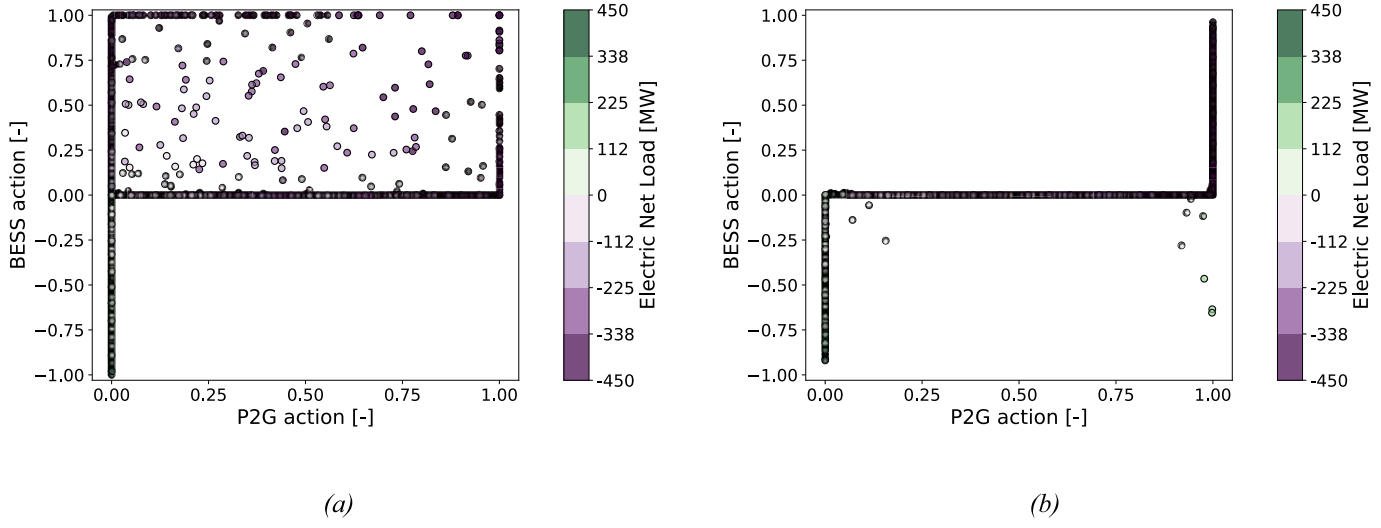


Fig. 11. State-action correlation with focus on the P2G and BESS interaction with respect to the net load, with (a) representing the rule-based strategy, and (b) the DRL strategy.

these optimization algorithms exploited synergies between the components of the energy system that the original unoptimized rule-based algorithm had not taken into account. The evidence uncovered by the DRL algorithms, and which was highlighted through the FIA on the RFR surrogate model, enabled a better understanding of the analyzed system. The following rules were added to the control logic and are shown in

Fig. 12: when the CHP was operating and the heat demand exceeded its heat output, the CHP generated sufficient electricity to augment the heat production of the HPs until either all the heat demands were met, or the HPs reached their maximum capacity. In this paper, the optimized rule-based algorithm was developed for different reasons: first, to prove the accuracy of the explanation of the solution identified by the DRL agents, as described in Section 3.1, second, to benchmark the agent' performance in identifying any suboptimal or unexpected behavior and, third, to evaluate the importance of the different technologies installed in the energy system described in Section 2.1.

As can be seen in Table 9, the results of the optimized rule-based algorithm differ slightly from those of the DRL algorithm. Indeed, the total operating costs resulting from the DRL learned policy are 1.5 higher; this difference can in part be imputed to the different priority in the dispatching order of the additional loads and in part to the deterministic nature of the rule-based algorithm, which allows a more precise control to be achieved for certain scenarios. CO_2 emissions are slightly lower (−0.2%), mainly because of the higher utilization of the P2G technology. Curtailment is 2.5% lower when the DRL agent is considered, due to the different dispatching orders of HP, BESS and P2G.

An RFR surrogate model was built to observe whether the new implemented rules accurately reflected the DRL policy, and FIA was developed (Fig. 13). FIA showed similar trends to those depicted in Fig. 9b, when considering the relationship between HP and the heat demand; however, the dispatching order of BESS and P2G was derived from the rule-based model, and FIA reflects this aspect.

The effects of implementing the new rules regarding the CHP and HP interaction can also be confirmed by comparing Fig. 14a–b with Fig. 10b–d, it can be observed that the two plots are almost overlapping. However, one of the limitations of DRL algorithms is their inability to constantly achieve perfect control, particularly in highly dynamic systems, and the policy employed by the DRL agent may not achieve the same level of consistency and accuracy as a rule-based system. Differences between the new rule-based strategy and the DRL one can also be noted by analyzing the operation of P2G and BESS. As mentioned in Section 3.1, DRL fails to find the best control of BESS. This suboptimal solution has been corrected in the new optimized algorithm. If we compare Fig. 15 with Fig. 11b, we can observe a shift in the

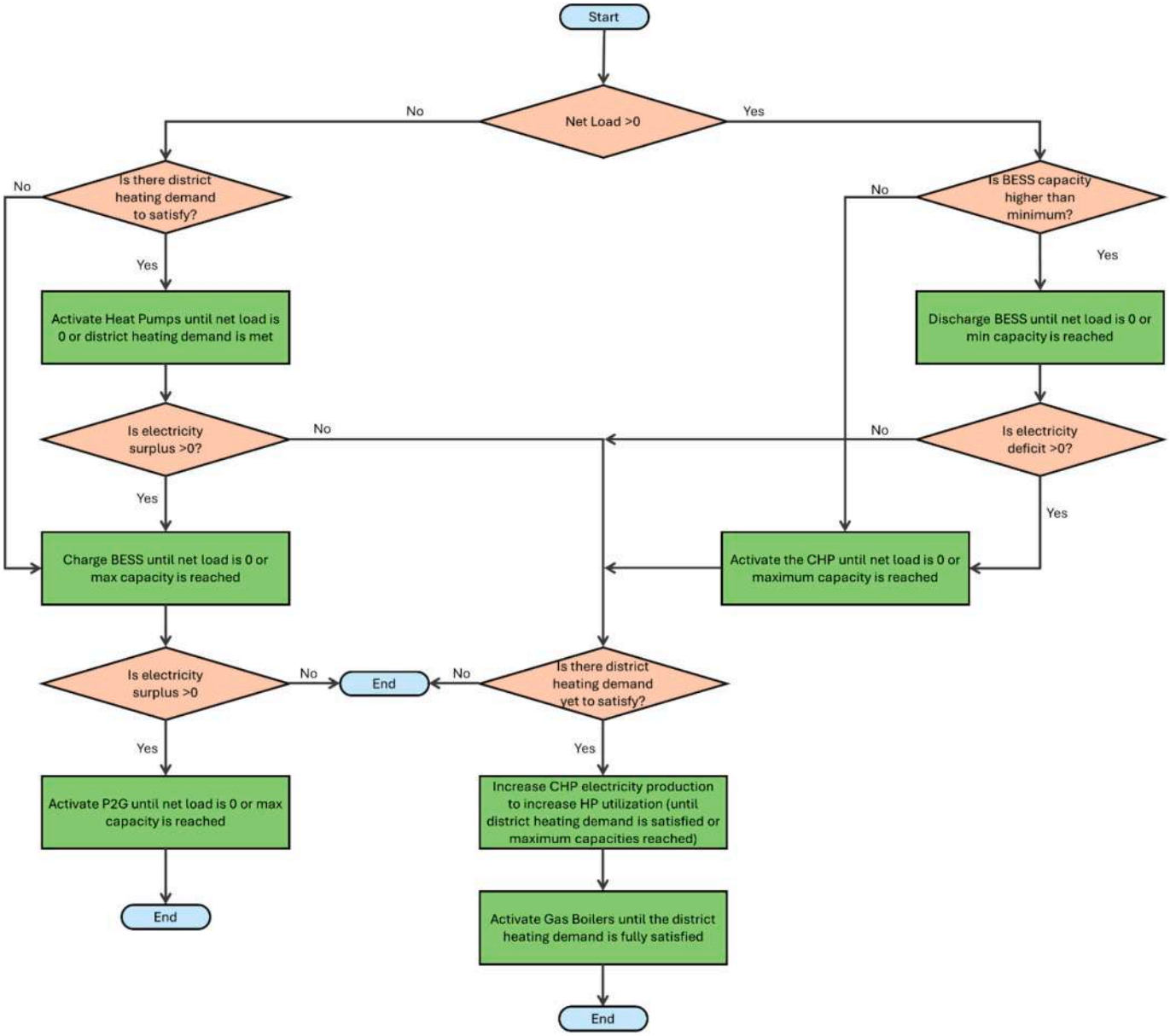


Fig. 12. Flowchart of the optimized rule-based strategy.

Table 9
Comparison of the results over the entire year obtained using the DRL agent and the optimized rule-based algorithm.

Quantity	UoM	Opt. rule-based	DRL Agent	Rel. Diff. [%] (Opt. Rule-Based)
Gas Consumption (CHP)	[TWh]	1.78	1.82	2.2%
Gas Consumption (Boiler)	[TWh]	0.63	0.63	0.0%
Gas Consumption (Total)	[TWh]	2.41	2.44	1.2%
SNG production	[TWh]	0.28	0.34	21.4%
Heat from HP	[TWh]	1.62	1.58	−2.5%
BESS (absorbed)	[TWh]	0.12	0.06	−50.8%
Electricity from the grid	[TWh]	0	0.04	–
Renewable curtailment	[TWh]	0.40	0.39	−2.5%
CO_2 direct emissions	[kton]	428	426	−0.2%
Fuel and electricity costs	[M€]	128.1	130.0	1.5%

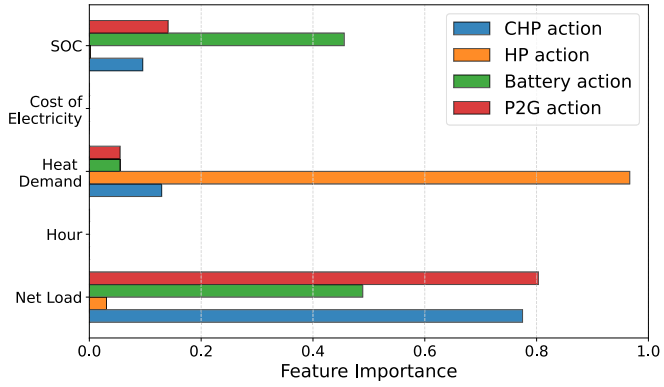


Fig. 13. The relative FIA results for the new rule-based strategy.

prioritization of these two technologies: in the DRL-controlled case, BESS is only charged after P2G has already been fully utilized, while in the optimized rule-based approach, P2G is activated after the BESS charging reaches its maximum. The impact of this operational difference is also evident in Fig. 17a and b, where the utilization of BESS (purple area) and P2G (orange area) differs for the two strategies. Moreover, some inaccuracies can also be noted in the DRL strategy, as in Fig. 16, that is, the CHP (blue area) and BESS (red area) interaction resulted in a slight overproduction of electricity for the evening of March 2nd. On the other hand, discrepancies in Fig. 18 (a and b) are negligible. Despite small differences and flaws, the overall behavior of the DRL agent remains close to the optimal solution, that is, it effectively balances cost reduction and operational flexibility. Nevertheless, it is important to recognize that even the rule-based control strategy, despite its deterministic nature, remains an approximation of reality, as energy systems are inherently complex and subject to uncertainties.

3.3. Sensitivity analysis

As mentioned in Section 2.1, a scenario with a high level of renewable energy integration was considered, in which energy conversion and storage technologies were allowed to exchange energy freely through all the considered grids; a sensitivity analysis was carried out to analyze the trend of the capacity factors of the different technologies considering different renewable penetration scenarios and the diffusion of HP, BESS and P2G using both the unoptimized and optimized rule-based

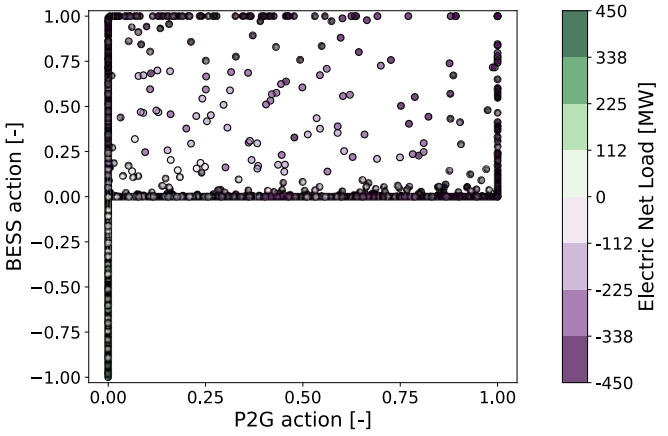


Fig. 15. State-action correlation with focus on the P2G and BESS interaction with respect to the net load in the optimized rule-based strategy.

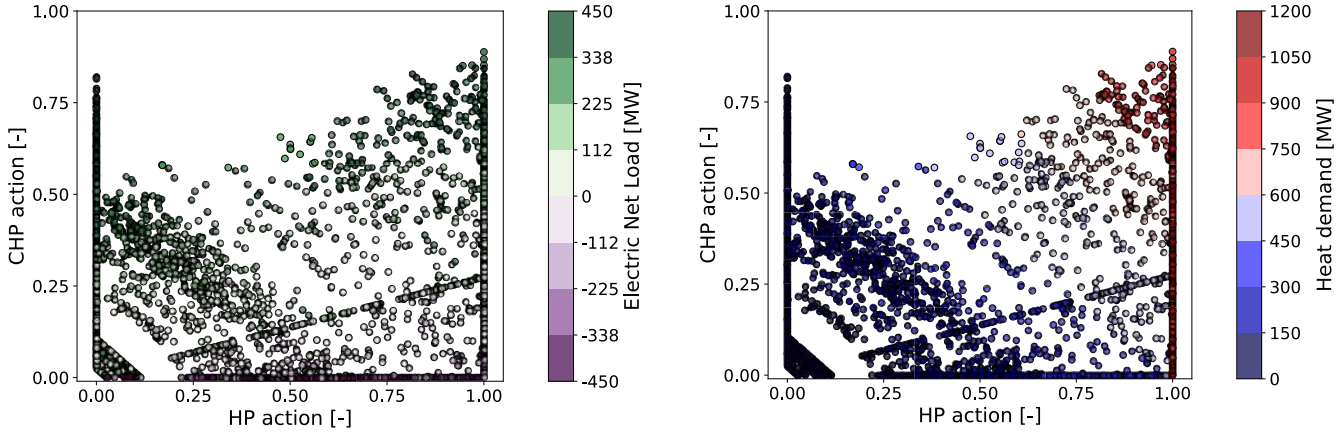


Fig. 14. State-action correlation with focus on the HP-CHP interaction in the optimized rule-based strategy. (a) represents the dependency on the net load (green-to-purple scale), while (b) represents the dependency on the heat demand (blue-to-red scale). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

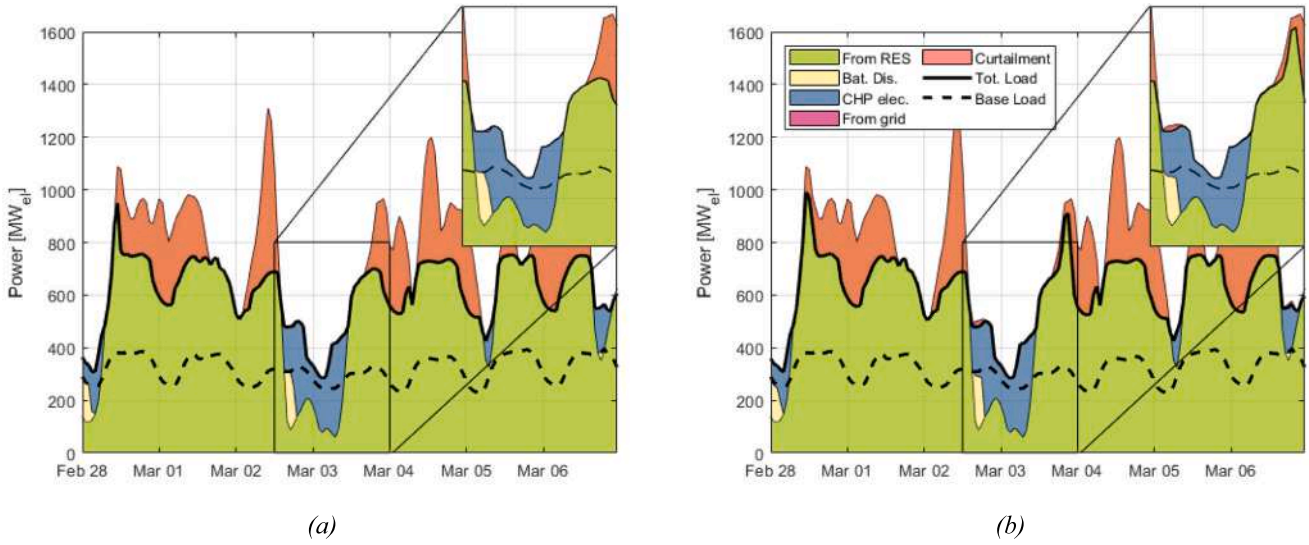


Fig. 16. Comparison of the power management strategies of (a) the Optimized Rule-Based algorithm, and (b) the DRL algorithm for the period from February 28th to March 6th, showing the energy contributions from the CHP, RES and the grid, the battery operations, and the load demands.

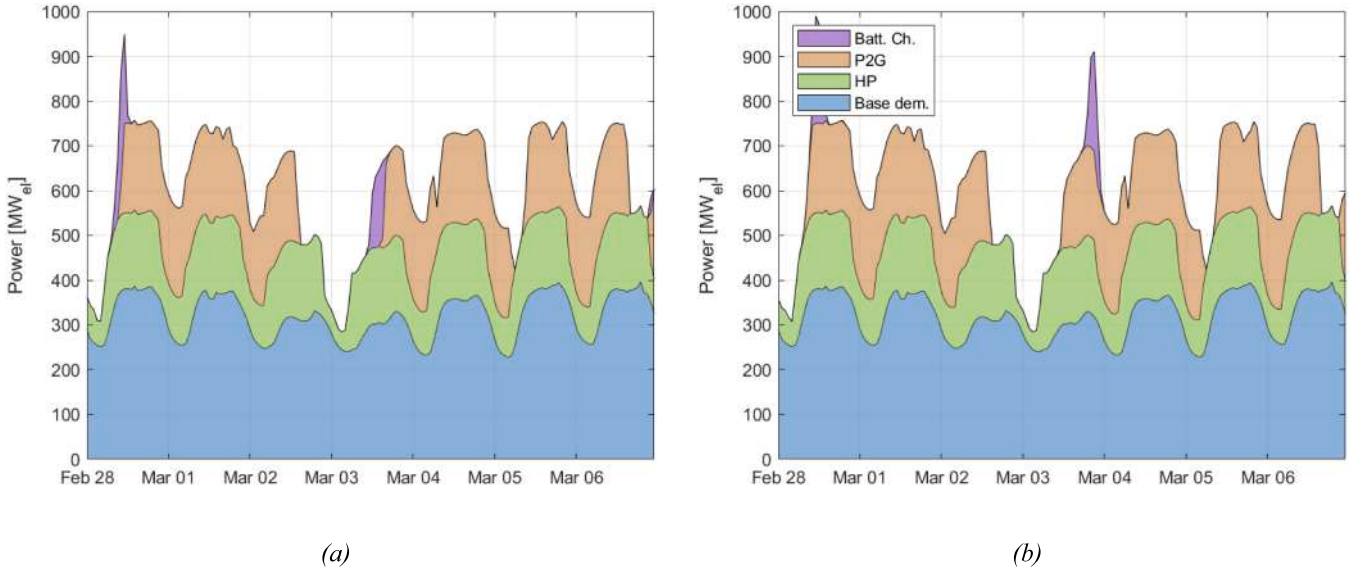


Fig. 17. Comparison of the electric demand of (a) the Optimized Rule-Based algorithm, and (b) the DRL algorithm for the period from February 28th to March 6th, showing the contributions from the HP, BESS and P2G systems to the baseline energy demand.

strategies.

The two variables that had the most impact on the performance of the energy system are the installed RES capacity and HP. This is because the amount of installed RES affects the amount of surplus electricity the agent has to handle, and the agent chooses the HP technology with the highest priority, thereby impacting the downstream technologies. Fig. 19a–d and Fig. 20a–d show the capacity factors of the CHP, HP, P2G and BESS obtained using the two rule-based strategies (the non-optimized rule-based strategy in Fig. 19 and the optimized one in Fig. 20) as a function of the RES installed capacity and HP installed power. The installed capacity values are reported in percentage, where a value of 100% refers to the capacity reported in Section 2.1, which was used in the previous analyses.

When Fig. 19a–b is compared with Fig. 20a–b, which represent the capacity factors of CHP and HP, respectively, similar trends can be observed: Fig. 19a and Fig. 20a show a decrease in the CHP capacity factors when the RES capacity increases, while the HP capacity factors tend to increase in Fig. 19b and Fig. 20b. However, the different strategies greatly impact the importance of a technology for a certain scenario: in fact, it can be noted that the CHP utilization factor in Fig. 19a does not change when the share of HP increases, since there is no interaction between the two components; however, when the interactions (see Section 3.2) are considered, both the CHP and HP capacity factors increase considerably, especially when the RES capacity is lower. In other words, the greater the installation of HP is, the more electrical energy must be produced by the CHP; as a result, less hot water is produced by the GB.

Analyzing how the capacity factor varied on the basis of operational logic and considering the system parameters allowed us to identify under which conditions a given technology could be advantageous. Indeed, a technology with a low-capacity factor, which means it is used marginally, is unlikely to generate sufficient benefits to justify the initial investment costs. Therefore, understanding these dynamics is crucial to evaluate the economic and operational sustainability of different

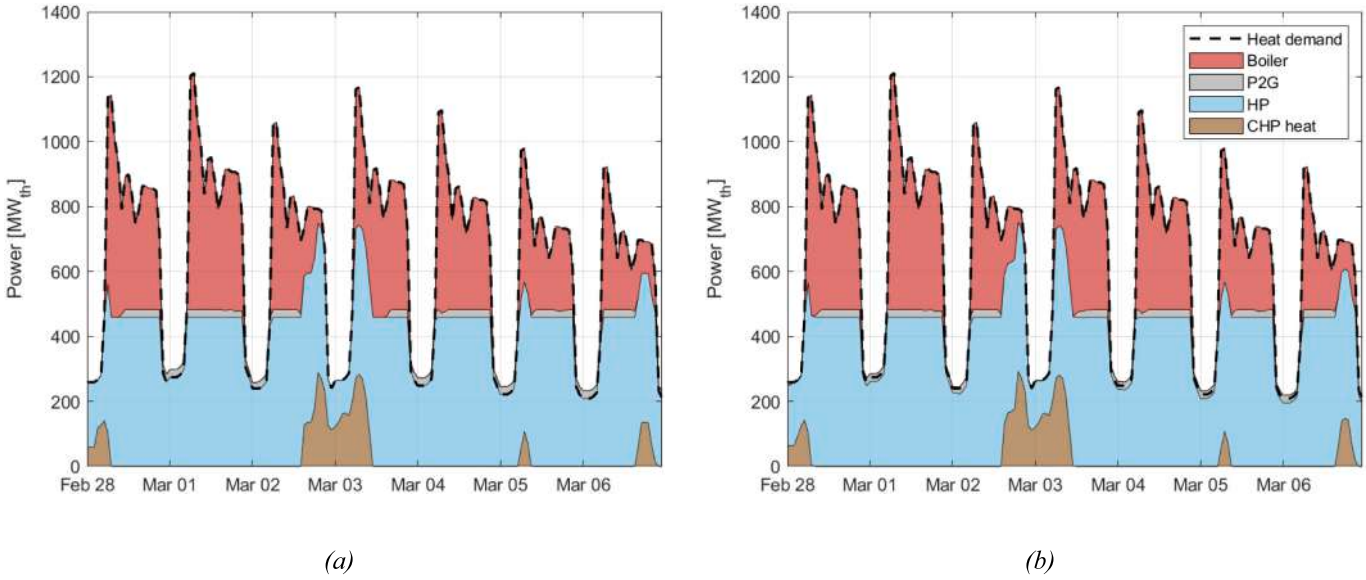


Fig. 18. Comparison of the heat management of (a) the Optimized Rule-Based algorithm, and (b) the DRL algorithm for the period from February 28th to March 6th, showing the contributions from the boilers, HP, P2G, and CHP systems to the overall heat demand.

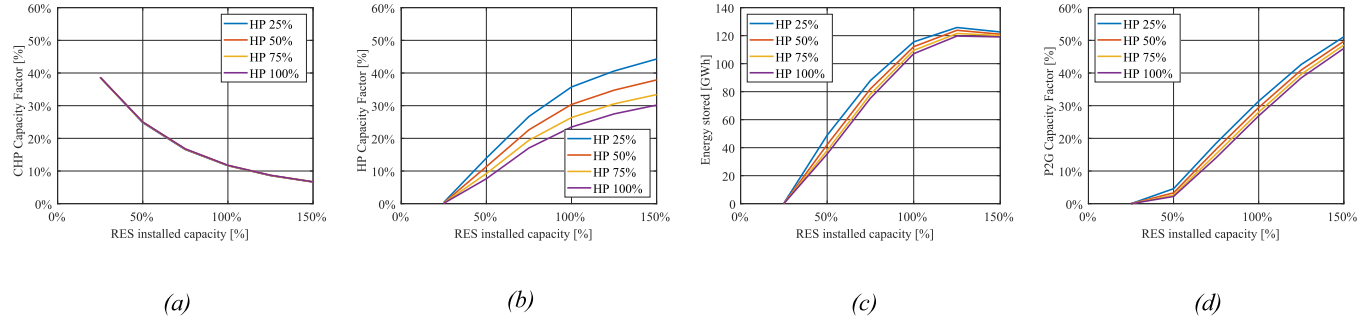


Fig. 19. Unoptimized rule-based strategy – sensitivity analysis of the capacity factors for CHP (a), HP (b), BESS (c), and P2G (d) as a function of the installed RES and HP capacities.

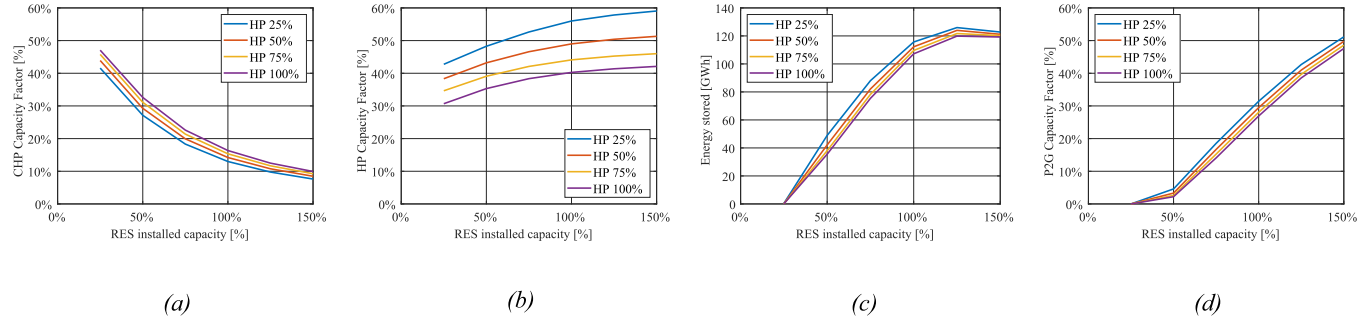


Fig. 20. Optimized rule-based strategy – sensitivity analysis of the capacity factors for CHP (a), HP (b), BESS (c), and P2G (d) as a function of the installed capacities of RES and HP.

technologies. It can also be observed that the control strategy influences the capacity factor. Compared to the unoptimized rule-based approach, the optimized rule-based strategy shows a slight increase in the capacity factor for CHP and a more significant increase for HP. This demonstrates that it is essential to define the correct control strategy to understand the overall real impact of different technologies on the system. It can be noted that as the RES capacity increases, the capacity factor of CHP decreases, thus suggesting that CHP may not be the most suitable technology for high renewable energy generation scenarios. This finding is aligned with the results of [51], which suggests that, in the coming decades, natural gas CHP might no longer be the most efficient power generation method when renewable penetration increases.

The two different control strategies that were analyzed do not significantly alter the activation logic of P2G and BESS, and this has led to the overlapping of Fig. 19c, Fig. 20c, Fig. 19d, and Fig. 20d. Fig. 19c and Fig. 20c show the total stored energy in the battery; it can be noted that the BESS utilization increases in the first part as the installed RES power increases; however, after 125% of installed RES power, the utilization decreases slightly, as the chances of being discharged diminish. As the installed power of HP increases, the use of BESS decreases; this is

because the HP absorbs part of the RES overproduction, thus decreasing the need to store production surpluses. Fig. 19d and Fig. 20d show that the P2G capacity factor increases almost linearly as the RES capacity increases.

Decreasing the HP capacity causes the utilization of P2G to increase, in a similar way to what happens for BESS. Like the CHP systems, P2G plants require a large number of operating hours to be cost-effective. Although PEM electrolyzers offer relatively flexible usage, the methanation reactor operates more efficiently under consistent conditions. As mentioned in [34], the system can be designed to separate the electrolyzer from the methanation unit by inserting a hydrogen storage unit, thereby allowing continuous operation of the methanation unit to be obtained. However, as also highlighted in [52], for this technology to be viable, the selling price of SNG must be significantly higher than that of conventional natural gas. For this reason, the agent did not prioritize the use of this technology. Clearly, if SNG production were to be incentivized, the agent's decisions would be different.

3.4. Discussion

In the present work, a temporal discretization of 1 h was chosen. This choice is in line with the scientific literature on modeling energy flows in multi-energy scenarios. The selected time step represents a compromise between calculation complexity of the calculation and the accuracy of the results. An hourly discretization allows a clear identification of fluctuations in energy production and demand and thus the definition of feasible operational strategies for the different control technologies. At the same time, this discretization does not allow a detailed analysis of the system dynamics (e.g. ramp rate constraints) associated with the technologies under consideration, which significantly reduces the computational effort required to find the optimal solution. A finer temporal resolution would enable the inclusion of such factors and allow the design of control strategies that also optimize the dynamic behavior of the different technologies. However, such an approach is beyond the scope of the energy dispatch optimization addressed in this study.

The developed methodological approach has proven to be effective in improving understanding of the internal decision-making mechanisms of the black-box solution. It enabled the identification of the key input features that influence the selection of actions and the exploration of the dependencies between the actions. Nevertheless, the interpretation of these relationships required the support of expert domain knowledge to avoid inaccurate conclusions arising from spurious or indirect correlations, as discussed in Section 3.1.

Moreover, the interpretation process exposed some limitations of the DRL algorithm, particularly with respect to reward sparsity and credit assignment. Reward sparsity occurs when meaningful feedback (i.e., non-zero rewards) is received only occasionally, while credit assignment issue arises when there is an imbalance in the reward. In this case, both phenomena occur:

- Reward sparsity: the scenario considered is characterized by high renewable penetration and, therefore, long surplus periods which distances the charging and the discharging phase.
- Credit assignment: the economic contribution of the BESS tends to be limited, particularly during the winter season, since it typically generates only marginal savings when compared to the usage of the CHP unit.

These effects combined led to suboptimal BESS utilization that needed to be corrected in the development of the optimized rule-based strategy. This hypothesis is confirmed by the comparison with the MILP optimization model, which slightly outperforms both the DRL and the optimized rule-based approach.

The MILP approach uses a perfect forecast horizon, while the DRL approach can only observe the current time step and avoids forecasts for the following hours. This is important in the context of developing a control strategy, as an analysis of the MILP-optimized strategy could lead to solutions that are difficult to apply. When considering finite horizon approaches that combine MILP with control schemes (such as Linear MPC with a finite prediction window), the performance degrades as reported in [30], leading to similar (or worse) performance compared to DRL.

4. Conclusion

In this paper, DRL has been used as a tool to explore synergies in complex energy systems. The proposed approach allows the black-box policies of DRL to be translated into more intuitive rules that could be generalized to conduct broader studies. An MES scenario with a high renewable energy penetration, was considered to pursue this goal. The MES incorporated CHP, BESS, HP, and P2G to couple the different energy sectors and increase the flexibility of the system. In this system, energy conversion and storage technologies were assumed to be free to exchange energy flows with the grids, and the DRL agent was left free to explore different solutions in the feasibility region in order to find the most economically advantageous one. Initially, the strategy identified by the DRL agent was compared with a simple rule-based control algorithm that has been proposed in the literature for a similar case study. The optimization performed by DRL proved to be more effective than the simplified rule-based approach. However, because of the nature of the DRL algorithms, the proposed solution operates as a black box, thus making it impossible to directly identify the reasons behind the control decisions made by the DRL agent. Thus, the solution generated by the DRL was analyzed to address this limitation. Moreover, a surrogate model was trained, and FIA was developed to interpret the policy adopted by the DRL agent, and this allowed the interactions between the energy devices exploited by the DRL agent to be identified. This analysis has allowed us to understand the following aspects:

- The reason why the DRL proposed strategy is more effective: it was found that this strategy depends on the optimization of the interactions between HP and CHP, which lead to more efficient heat management practices. This resulted in a reduced use of gas boilers (61% less) and a consequent increase in the use of heat pumps and CHP (43% and 70% more, respectively).
- The limitations of the implemented methodology: this technology also revealed some optimization limitations regarding the use of storage systems by the DRL agents, which was found to be suboptimal. In other words, it was observed that the DRL algorithm struggles to find optimal control of the batteries. This issue is already known in the literature, and it is caused by the fact that the action of charging the battery is temporally separated for these devices from the gain in using the device, as the gain occurs during the battery discharge phase. This temporal decoupling makes it difficult for the DRL agent to establish the connection between the actions and rewards.
- Some minor inaccuracies were identified. These inaccuracies concerned the agent's ability to perfectly balance the energy demand and production. However, these imbalances were found to be negligible.

Building on these insights, we have developed a refined rule-based strategy by integrating the most effective operational principles learned from the DRL. This new approach retains the transparency and simplicity of rule-based control, while incorporating the efficiency gains observed in the DRL optimization. A comparison between the two methods has shown that the performance of the optimized rule-based strategy closely matched the performance of the DRL solution, and it even produced slightly better results.

The flexibility and simplicity of the optimized rule-based algorithm enabled us to conduct a comprehensive sensitivity analysis, and this allowed us to obtain a clearer understanding of how different technologies interact and perform in various scenarios. This adaptability is

crucial for the design of robust and efficient MES. By analyzing how the capacity factors of various technologies responded to changes in the installed capacity and operational strategies, we identified several key findings:

- Among the examined technologies, HP, the most efficient conversion technology, exhibited the highest capacity factor. Because of the interconnected nature of MES, changes in one technology affected the other technologies. Indeed, an increase in the HP capacity raised the capacity factor of CHP, as greater electricity demand emerged. Conversely, a higher HP capacity reduced the utilization of BESS and P2G, as HP absorbed any surplus RES, thereby decreasing the flexibility requirements.
- As RES penetration continued to grow, CHP became less viable for cogeneration, since direct RES utilization and RES-powered HP for heat provision were more cost-effective. Indeed, a low CHP capacity factor may fail to justify investment costs. Nonetheless, backup power remains essential, raising the question of whether cogeneration or simpler alternatives would be preferable.
- Increased RES penetration also amplified the flexibility demands, thereby enhancing the utilization of BESS and P2G. However, BESS utilization declined beyond a certain threshold, due to energy saturation. P2G only achieved a 50% capacity factor for particularly high RES penetration levels, and it remained underutilized, due to its low efficiency and the availability of better alternatives, such as HP, therefore necessitating financial or regulatory incentives to make its adoption plausible.

In this study, DRL has been employed to optimize the operational strategies of an energy system. However, DRL has also shown a significant potential concerning the issue of optimal system sizing. Future research will explore the application of DRL as a decision-support tool to optimize the sizing of MES, with the aim of enhancing design-phase decision-making by integrating operational flexibility and long-term performance considerations.

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

The authors only used ChatGPT during the preparation of this paper 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 – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Gabriele Fambri:** Writing – original draft, Supervision, Methodology, Investigation, 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enconman.2025.120095.

Data availability

Data will be made available on request.

References

[1] K. Calvin *et al.*, "IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.," Jul. 2023. doi: 10.59327/IPCC/AR6-9789291691647.

[2] European Commission, "A European Green Deal." Accessed: Aug. 26, 2023. [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en.

[3] European commission, "A European Green Deal." Accessed: Aug. 26, 2023. [Online]. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en.

[4] International Energy Agency, "Renewables 2024," 2024. [Online]. Available: www.iea.org.

[5] Mancarella P. MES (multi-energy systems): an overview of concepts and evaluation models. Energy 2014;65:1–17. https://doi.org/10.1016/j.energy.2013.10.041.

[6] Badami M, Fambri G. Optimising energy flows and synergies between energy networks. Energy 2019;173:400–12. https://doi.org/10.1016/J.ENERGY.2019.02.007.

[7] Mohammadi M, Noorollahi Y, Mohammadi-ivatloo B, Yousefi H. Energy hub: from a model to a concept – a review. Renew Sustain Energy Rev Dec. 2017;80:1512–27. https://doi.org/10.1016/J.RSER.2017.07.030.

[8] Sievers J, Blank T. A systematic literature review on data-driven residential and industrial energy management systems. Energies 2023;16(4):1688. https://doi.org/10.3390/EN16041688.

[9] Perera ATD, Kamalaruban P. Applications of reinforcement learning in energy systems. Renew Sustain Energy Rev 2021;137:110618. https://doi.org/10.1016/J.RSER.2020.110618.

[10] Dhavala SS, Srihari C, Vanishree K, Rashmi R. An extensive review of applications, methods and recent advances in deep reinforcement learning. 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) 2023:1–6. https://doi.org/10.1109/HORA58378.2023.10156687.

[11] R. S. . Sutton and A. G. . Barto, Reinforcement learning : an introduction. The MIT Press, 2020.

[12] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning," Dec. 2013, Accessed: Feb. 06, 2025. [Online]. Available: https://arxiv.org/abs/1312.5602v1.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. K. Openai, "Proximal Policy Optimization Algorithms," Jul. 2017, Accessed: Feb. 06, 2025. [Online]. Available: https://arxiv.org/abs/1707.06347v2.

[14] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[15] T. Haarnoja *et al.*, "Soft Actor-Critic Algorithms and Applications," Dec. 2018, Accessed: Feb. 06, 2025. [Online]. Available: https://arxiv.org/abs/1812.05905v2.

[16] Vamvakas D, Michailidis P, Korkas C, Kosmatopoulos E. Review and evaluation of reinforcement learning frameworks on smart grid applications. Energies 2023;16(14):5326. https://doi.org/10.3390/EN16145326.

[17] Lissa P, Deane C, Schukat M, Seri F, Keane M, Barrett E. Deep reinforcement learning for home energy management system control. Energy AI 2021;3:100043. https://doi.org/10.1016/J.EGYAI.2020.100043.

[18] Brandi S, Piscitelli MS, Martellacci M, Capozzoli A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energ Build 2020;224:110225. https://doi.org/10.1016/J.ENBUILD.2020.110225.

[19] Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy 2021;229:120725. https://doi.org/10.1016/J.ENERGY.2021.120725.

[20] Park JY, Dougherty T, Fritz H, Nagy Z. LightLearn: an adaptive and occupant centered controller for lighting based on reinforcement learning. Build Environ 2019;147:397–414. https://doi.org/10.1016/J.BUILDENV.2018.10.028.

[21] Ding X, Du W, Cerpa A. OCTOPUS: Deep reinforcement learning for holistic smart building control. BuildSys 2019 - Proce 6th ACM Int Conference on Systems for Energy-Efficient Buildings, Cities, and Transp 2019;19:326–35. https://doi.org/10.1145/3360322.3360857.

[22] Silvestri A, et al. Real building implementation of a deep reinforcement learning controller to enhance energy efficiency and indoor temperature control. Appl Energy 2024;368:123447. https://doi.org/10.1016/J.APENERGY.2024.123447.

[23] Dorokhova M, Martinson Y, Ballif C, Wyrsch N. Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation. Appl Energy 2021;301:117504. https://doi.org/10.1016/J.APENERGY.2021.117504.

[24] Zhang F, Yang Q, An D. CDDPG: a deep-reinforcement-learning-based approach for electric vehicle charging control. IEEE Internet Things J 2021;8(5):3075–87. https://doi.org/10.1109/JIOT.2020.3015204.

[25] Abedi S, Yoon SW, Kwon S. Battery energy storage control using a reinforcement learning approach with cyclic time-dependent Markov process. Int J Electr Power Energy Syst 2022;134:107368. https://doi.org/10.1016/J.IJEPES.2021.107368.

[26] Zhou S, et al. Combined heat and power system intelligent economic dispatch: a deep reinforcement learning approach. Int J Electr Power Energy Syst 2020;120:106016. https://doi.org/10.1016/J.IJEPES.2020.106016.

[27] Guo C, Wang X, Zheng Y, Zhang F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. Energy 2022; 238:121873. https://doi.org/10.1016/J.ENERGY.2021.121873.
[28] 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 Building Eng 2023;65:105682. https://doi.org/10.1016/j.jobe.2022.105682.
[29] 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.
[30] Ceusters G, et al. Model-predictive control and reinforcement learning in multi-energy system case studies. Appl Energy 2021;303:117634. https://doi.org/10.1016/J.APENERGY.2021.117634.
[31] Bousnina D, Guerassimoff G. Optimal energy management in smart energy systems: a deep reinforcement learning approach and a digital twin case-study. Smart Energy 2024;16:100163. https://doi.org/10.1016/J.SEGY.2024.100163.
[32] Ali S, et al. Explainable Artificial Intelligence (XAI): what we know and what is left to attain Trustworthy Artificial Intelligence. Inf Fusion 2023;99:101805. https://doi.org/10.1016/J.INFFUS.2023.101805.
[33] Razzano G, Brandi S, Piscitelli MS, Capozzoli A. Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of Heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings. Appl Energy 2025;381: 125046. https://doi.org/10.1016/J.APENERGY.2024.125046.
[34] Fambri G, Diaz-Londono C, Mazza A, Badami M, Weiss R. Power-to-Gas in gas and electricity distribution systems: a comparison of different modeling approaches. J Energy Storage 2022;55:105454. https://doi.org/10.1016/J.EST.2022.105454.
[35] Barco-Burgos J, Bruno JC, Eicker U, Saldaña-Robles AL, Alcántar-Camarena V. Review on the integration of high-temperature heat pumps in district heating and cooling networks. Energy 2022;239:122378. https://doi.org/10.1016/J.ENERGY.2021.122378.
[36] Terna, "Transparency report." Accessed: Jun. 26, 2023. [Online]. Available: https://www.terna.it/it/sistema-elettrico/transparency-report/download-center.
[37] Noussan M, Jarre M, Poggio A. Real operation data analysis on district heating load patterns. Energy 2017;129:70–8. https://doi.org/10.1016/J.ENERGY.2017.04.079.
[38] ARERA, "Dati e Statistiche." Accessed: Jul. 31, 2024. [Online]. Available: https://www.arera.it/dati-e-statistiche.
[39] Noussan M, Fambri G, Negro V, Chiaramonti D. Hourly electricity CO2 intensity profiles based on the real operation of large-scale natural gas combined cycle cogeneration plants. Energy 2024;312:133424. https://doi.org/10.1016/J.ENERGY.2024.133424.
[40] Laveneziana L, Prussi M, Chiaramonti D. Critical review of energy planning models for the sustainable development at company level. Energ Strat Rev 2023;49: 101136. https://doi.org/10.1016/J.ESR.2023.101136.
[41] Badami M, et al. A decision support system tool to manage the flexibility in renewable energy-based power systems. Energies 2019;13:153. https://doi.org/10.3390/EN13010153.
[42] C. Diaz-Londono, G. Fambri, A. Mazza, M. Badami, and E. Bompard, "A Real-Time Based Platform for Integrating Power-to-Gas in Electrical Distribution Grids," *UPEC 2020 - 2020 55th International Universities Power Engineering Conference, Proceedings*. 2020, doi: 10.1109/UPEC49904.2020.9209803.
[43] Hoang AT, Pham VV, Nguyen XP. Integrating renewable sources into energy system for smart city as a sagacious strategy towards clean and sustainable process. J Clean Prod 2021;305:127161. https://doi.org/10.1016/J.JCLEPRO.2021.127161.
[44] Götz M, et al. Renewable Power-to-Gas: a technological and economic review. Renew Energy 2016;85:1371–90. https://doi.org/10.1016/J.RENENE.2015.07.066.
[45] Kötter E, Schneider L, Sehnke F, Ohnmeiss K, Schröer R. The future electric power system: impact of Power-to-Gas by interacting with other renewable energy components. J Energy Storage 2016;5:113–9. https://doi.org/10.1016/J.EST.2015.11.012.
[46] Arpagaus C, Bless F, Uhlmann M, Schiffmann J, Bertsch SS. High temperature heat pumps: market overview, state of the art, research status, refrigerants, and application potentials. Energy 2018;152:985–1010. https://doi.org/10.1016/J.ENERGY.2018.03.166.
[47] Capone M, Guelpa E, Verda V. Optimal installation of heat pumps in large district heating networks. Energies (Basel) 2023;16(3). https://doi.org/10.3390/en16031448.
[48] C. Molnar, *Interpretable Machine Learning*, 2nd ed. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book.
[49] Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. Decision Analytics J 2023;7:100230. https://doi.org/10.1016/J.DAJOUR.2023.100230.
[50] Gurobi Optimization LLC, "Gurobi Optimizer Reference Manual," 2024. [Online]. Available: https://www.gurobi.com.
[51] Franzoso A, Noussan M, Marocco P, Badami M, Fambri G, Gandiglio M. Assessing the role of storage and thermoelectric plants in the energy transition: a short-and medium-term scenario analysis with Italy as a case study. Smart Energy 2025;13: 100186. https://doi.org/10.1016/j.segy.2025.100186.
[52] Fambri G, Diaz-Londono C, Mazza A, Badami M, Sihvonen T, Weiss R. Techno-economic analysis of Power-to-Gas plants in a gas and electricity distribution network system with high renewable energy penetration. Appl Energy 2022;312: 118743. https://doi.org/10.1016/J.APENERGY.2022.118743.