Applied Energy 333 (2023) 120526

Contents lists available at ScienceDirect

Applied Energy

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

Federated reinforcement learning for smart building joint peer-to-peer energy and carbon allowance trading

Dawei Qiu [a], Juxing Xue [b], Tingqi Zhang [c], Jianhong Wang [a], Mingyang Sun [b,*]

[a] Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK
[b] College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
[c] Electric Power Research Institute, State Grid Liaoning Electric Power Company Ltd., Shenyang, China

ARTICLE INFO

Keywords:
Multi-energy system
Smart buildings
Peer-to-peer energy trading
Emission trading scheme
Multi-agent reinforcement learning
Federated learning

ABSTRACT

The multi-energy system (MES), which is regarded as an optimum solution to a high-efficiency, green energy system and a crucial shift towards the future low-carbon energy system, has attracted great attention at the district building level. However, the current exploration of flexible MES operation has been hampered by (1) the increasing penetration of renewable energies and the complicated operation of coupling multi-energy sectors; (2) the privacy concern in the decentralization of the energy system; and (3) the lack of integration of the energy market and carbon emission trading scheme. To address the aforementioned challenges, this paper proposes a joint peer-to-peer energy and carbon allowance trading mechanism for a building community, and then models it as a multi-agent reinforcement learning (MARL) paradigm. In this setting, the flexibility of building local trading and the decarbonization of building energy management can both be fully utilized. To stabilize the training performance, an abstract critic network capturing system dynamics is introduced based on a deep deterministic policy gradient method. The technique of federated learning (FL) is also applied to speed up the training and safeguard the private information of each building in the community. Empirical results on a real-world test case evaluate its superior performance in terms of achieving both economic and environmental benefits, resulting in 5.87% and 8.02% lower total energy and environment costs than the two baseline mechanisms of peer-to-grid energy trading and peer-to-peer energy trading, respectively.

1. Introduction

Buildings account for a large proportion of the world's energy consumption and carbon emissions. In 2020, buildings worldwide consumed 36% of total energy and generated 37% of total carbon emissions, while these two statistics have been put on the agenda and are expected to fall by 50% and 45% respectively till 2030 towards a low-carbon transition [1]. In the above context, smart buildings have received increasing attention in recent years following the deployment of advanced technologies, such as smart meters, machine learning, and big data analysis [2], to provide a sustainable, economical, and comfortable operating environment for energy users. In order to support the above functions, it is critical and urgent to develop an autonomous building energy management scheme (BEMS) that can achieve the best performance in energy consumption, energy cost, carbon emission, and user comfort through the intelligent scheduling of building energy systems [3].

In general, buildings are characterized by a multi-energy system (MES) [4], including electric and heat demand, renewable-based generation, energy storage, multi-energy converters, and demand-side response technologies. The energy consumption and surplus of buildings are both traded with the upstream utility company at high retail import prices (e.g., Time-of-Use (ToU) prices) and low retail export prices (e.g., Feed-in-Tariff (FiT)) [5]. As a result, buildings may complain about such unfair retail pricing schemes and lose economic benefits. In recent years, peer-to-peer (P2P) energy trading [6] has provided localized solutions to such problems. P2P specifically enables buildings to trade energy surpluses with peers who consume energy. Moreover, the energy consumption of buildings is mainly supplied by coal, oil, and natural gas. The burning of these fossil fuels and the losses caused by long-distance power transmission will lead to a series of environmental problems. To this end, trading carbon allowances in emissions trading schemes (ETS) [7] is regarded as an economically effective policy favored by many economists for carbon emission reduction, which has been promoted as a global energy strategy towards low-carbon

[*] Corresponding author.
E-mail address: mingyangsun@zju.edu.cn (M. Sun).

https://doi.org/10.1016/j.apenergy.2022.120526
Received 1 March 2022; Received in revised form 25 November 2022; Accepted 13 December 2022
Available online 28 December 2022
0306-2619/© 2022 Published by Elsevier Ltd.

Nomenclature

A. Indices and Sets

$i \in \mathcal{I}$	Index and set of smart buildings
$t \in T$	Index and set of time steps (hours)

B. Parameters

Δt	Time resolution of market (1 h)
λ_t^b	Grid electricity buy price at time step t (£/kWh)
λ_t^s	Grid electricity sell price at time step t (£/kWh)
λ_t^m	Mid-market rate at time step t (£/kWh)
λ_t^{m+}	Local buy price at time step t (£/kWh)
λ_t^{m-}	Local sell price at time step t (£/kWh)
λ^g	Grid gas price (£/kWh)
λ^c	Carbon price (£/ton)
P_t^l	Electric load at time step t (kW)
Q_t^l	Heat load at time step t (kW)
P_t^{pv}	PV generation at time step t (kW)
$\overline{P}^{ees}$	Power capacity of EES (kW)
$\overline{E}^{ees}$	Energy capacity of EES (kWh)
η^{ees}	Charging/Discharging efficiency of EES
$\overline{Q}^{tes}$	Power capacity of TES (kW)
$\overline{E}^{tes}$	Energy capacity of TES (kWh)
η^{tes}	Charging/Discharging efficiency of TES
η^{chpe}	Energy conversion efficiency from gas to electricity of CHP
η^{chpq}	Energy conversion efficiency from gas to heat of CHP
$\overline{P}^{chp}$	Electric power capacity of CHP (kW)
$\overline{Q}^{chp}$	Heat power capacity of CHP (kW)
η^{ehp}	Energy conversion efficiency from electricity to heat of EHP
$\overline{Q}^{ehp}$	Heat power capacity of EHP (kW)
η^{gb}	Energy conversion efficiency from gas to heat of GB
$\overline{Q}^{gb}$	Heat power capacity of GB (kW)
H_t^{out}	Outdoor temperature at step t (° C)
$\overline{P}^{hvac}$	Power capacity of HVAC (kW)
$\underline{H}^{hvac}$	Minimum temperature limit of HVAC (° C)
$\overline{H}^{hvac}$	Maximum temperature limit of HVAC (° C)
η_i^{hvac}	Conversion efficiency from power to temperature of HVAC of MG i

C. Variables

V_t^{ees}	Binary variable indicating whether EES charges ($V_{i,t}^{ees} = 1$) or discharges ($V_{i,t}^{ees} = 0$) at time step t
P_t^{eesc}	Charging power of EES at time step t (kW)
P_t^{eesd}	Discharging power of EES at time step t (kW)
E_t^{ees}	Energy content in EES at time step t (kWh)
V_t^{tes}	Binary variable indicating whether TES charges ($V_{i,t}^{tes} = 1$) or discharges ($V_{i,t}^{tes} = 0$) at time step t
Q_t^{tesc}	Charging power of TES at time step t (kW)
Q_t^{tesd}	Discharging power of TES at time step t (kW)
E_t^{tes}	Energy content in TES at time step t (kWh)
G_t^{chp}	Gas power input of CHP at time step t (kW)
P_t^{chp}	Electric power output of CHP at time step t (kW)
Q_t^{chp}	Heat power output of CHP at time step t (kW)
P_t^{ehp}	Electric power input of EHP at time step t (kW)
Q_t^{ehp}	Heat power output of EHP at time step t (kW)
G_t^{gb}	Gas power input of GB at time step t (kW)
Q_t^{gb}	Heat power output of GB at time step t (kW)
P_t^{hvac}	Demanding power of HVAC at time step t (kW)
H_t^{in}	Indoor temperature of HVAC at time step t (° C)

transitions. However, tracing and reducing carbon emissions at the district building level is still vacant since trading carbon allowances is normally applied at the national level in the centralized market. On the other hand, separately designing energy and carbon trading systems is not efficient, since energy and environmental concerns may bring about conflicts in view of using cheap natural gas (but damaging the environment) and clean electricity (but increasing the expense), respectively. To address the aforementioned two issues, this paper proposes a joint P2P energy and carbon trading mechanism for a group of MES buildings that can achieve a better balance between energy costs and carbon emissions, resulting in a low-cost and environmentally friendly transition.

1.1. Literature review on model-based optimization approaches

There have been many studies on P2P energy trading among buildings in the existing literature. In which, the centralized optimization approach is the most straightforward method to optimize the energy schedules and trading strategies of buildings. In [8], a P2P trading-based energy cost optimization problem is proposed to find the optimal solutions that result in a fair cost distribution among the participating smart homes. In [9], a centralized optimization problem is proposed to manage the peer energy trading to achieve a net-zero energy community, consisting of university campuses, commercial offices, and high-rise residential buildings. In [10], a joint co-optimization problem is proposed to evaluate the energy complementary effect between residential and industrial prosumers, with the objective of energy cost minimization. Although the energy management and trading strategy can be guaranteed under the centralized optimization approach, the central coordinator needs to acquire all participants' mathematical models and technical parameters, which may destroy their security and privacy [11]. To this end, the decentralized optimization approach is being applied to the P2P energy trading problems, since only a small amount of information is exchanged between each market participant at relatively low computational costs. In [12], a dual-consensus algorithm based on the alternating direction method of multipliers (ADMM) is proposed to solve the global pricing consensus among smart buildings, which can preserve the privacy of participant buildings to a significant extent. In [13], a multi-leader, multi-follower Stackelberg game approach is utilized to model the energy trading process among microgrids. In [14], a decentralized control based on ADMM is proposed to find coordinated energy management for a building community, transiting to an economically and sustainably building community.

Nevertheless, the above three papers only consider energy management and P2P energy trading in electricity, while neglecting the flexibility of multi-energy vectors. The flexibility of multi-energy vectors has been investigated in recent studies. In [15], a scenario-based framework for energy hub (EH) design considering the variable efficiencies of gas-fired converters, wind turbines, and integrated demand response programs is developed. According to [16,17], the optimal scheduling of EH is optimized using information gap decision theory (IGDT). In [18], a two-stage robust operation scheduling of EH based on the worst scenarios is solved. A scenario-based stochastic programming of the EH operation model considering uncertain electric vehicle batteries [19] and renewable power generation [20] is developed. Recently, P2P energy trading under the MES concept has received great attention, since the combination of P2P energy trading and multi-energy conversion can fully maximize the flexibility of the local energy system. In [21], a cooperative game theory approach is proposed to solve the P2P energy trading problem among 4 EHs and also shows its scalability in a larger 32-EHs experiment. In [22], a non-cooperative game theory approach is applied to a multi-energy market that aims at developing a fair and convincing payoff allocation scheme and determining the fair prices for multi-energy trading. In [23], a parametric optimization-based peer-to-peer energy trading model is developed to optimize the energy management of commercial buildings together with the procurement of natural gas from the grid, with the objective of minimizing energy costs and increasing energy efficiency. In [24], a fully decentralized ADMM model is proposed to address the electricity trading within a transactive energy market incorporating industrial, commercial, and residential energy hubs. On the other hand, some existing studies have focused on the joint trading of local energy and carbon markets. In [25], a blockchain-based P2P trading framework is proposed to optimize the trading strategies of prosumers and microgrids for both energy bids in an auction-based market and carbon bids in the ETS. In [26], a transactive energy and carbon market for networked microgrids based on cooperative games is developed. Microgrids in this problem can submit trading energy and carbon allowance data to the centralized operator, which will optimize the energy and carbon allowance trading quantities among microgrids to satisfy power distribution network constraints. However, the system uncertainties and dynamics are extremely difficult to address in the decentralized approach, since the iterative algorithm in a decentralized manner is time-consuming, especially when accounting for a highly complex and dynamic environment. Even though some papers assume uncertain parameters (e.g., normal distribution), such optimal solutions are theoretically possible. This assumption is not realistic since these uncertainties do not simply follow a fixed probability distribution. Furthermore, to solve such problems, all the mathematical models and technical parameters of the studied buildings must be obtained in order to form an optimization, which may destroy their privacy and lead to another unrealistic assumption.

1.2. Literature review on model-free reinforcement learning approaches

Reinforcement learning (RL) [27], as a model-free and data-driven approach in the machine learning era, is used to study the sequential and dynamic decision-making problems of agents that can gradually learn the optimal control decisions by utilizing experiences acquired from their repeated interactions with the environment, without a *prior* knowledge. As a result, RL does not require any knowledge of mathematical models or technical parameters, since they are captured in the environment, which is also assumed to be a black box for RL agents. In addition, RL as an online learning method can make efficient use of increasing data, thereby capturing system uncertainties and adapting to various state dynamics. Finally, once the RL algorithm is well trained, its policy can be delivered to the on-line test set on timescales of milliseconds without requiring any identification. Therefore, RL is claimed to be an efficient tool for real-time automatic control applications.

Previous papers have successfully adopted various RL methods to solve P2P energy trading problems [28]. In general, these methods can be classified into two categories, depending on the learning framework. The first category is concerned with the independent learning algorithm, which applies traditional RL methods to a multi-agent setup. In [29], a deep Q-network (DQN) method is proposed to optimize the local trading strategies of a prosumer in a holistic electricity market model. In [30], a DQN method is proposed to solve a decision-making problem for three microgrids in the local energy market. The proposed method is also generalized to four seasons within the year. In [31], a DQN method is proposed to address the problem of energy sharing among multiple buildings in a zero-energy community. However, directly implementing DQN into a multi-agent setup may suffer from an instability issue without information exchange, since all other agents' policies are implicitly formulated as part of the environment's dynamics while their own policies are continuously adjusted during the training process. In addition, DQN based on the value-based RL method can only generate discrete actions, which is not applicable for the problems in continuous action space, e.g., the joint P2P energy and carbon problem studied in this paper.

To address this issue, the second category of the centralized learning algorithm has also been applied to the P2P energy trading problem. In [32], multi-agent deep deterministic policy gradient (MADDPG) method is proposed to optimize the trading decisions, which can help each microgrid find the optimal policy without requiring the generation and load information of other microgrids. In [33], a modified MADDPG is developed by taking all agents' information into the training process, which has been used to solve the automatic P2P energy trading among residential houses in a double-auction market. To further stabilize the training performance of MADDPG, a multi-agent twin delayed deep deterministic policy gradient (MATD3) method is proposed in [34] to solve the P2P energy trading problem among three multi-energy microgrids in residential, commercial, and industrial areas, respectively. In addition to microgrids, the P2P energy trading for multi-energy prosumers are also investigated in [35] that applies a mean-field MARL algorithm based on MADDPG. This method allows all prosumer agents to learn a single shared policy, with mean-field approximation enhancing system scalability and parameter-sharing achieving accelerated convergence speed. In [36], another popular MARL algorithm named soft actor–critic (SAC) is proposed to optimize the energy management of four buildings with the objectives of cost minimization and peak demand reduction. In [37], a modified SAC method is proposed for an industrial park, which can enhance the stability of policy and enable agents to focus on important energy-related information by introducing an attention mechanism. However, the above conventional CL-based MARL methods may raise privacy concerns since they require the information of other agents' local observations and actions when training the centralized critic network. In order to address this issue, the technique of federated learning (FL) [38] can be applied to the RL concept, forming federated reinforcement learning (FRL) [39]. More specifically, FRL is a distributed learning framework for training the network models of local agents without sharing local information. Despite the aforementioned attempts to investigate the P2P energy trading problem, it is still challenging to trace and trade the carbon emissions caused by these localized buildings, since the ETS in the carbon market is operated in a centralized manner independent of the exchange of carbon allowance among individual buildings.

1.3. Paper contributions

This paper aims at addressing the above challenge by introducing a novel MARL method that can efficiently learn the energy management scheme and trading strategy for a group of MES buildings in a joint P2P energy and carbon trading market. More specifically, the novel contributions of this paper are described below:

(1) Create a **J**oint **P**2P energy and **C**arbon trading (JPC) mechanism for an MES building community wherein each building acts as an agent. A mid-market rate (MMR) pricing scheme is established in P2P energy trading to encourage more local energy trading of buildings. In carbon trading, a cap-and-trade (CAT) approach is applied to allow buildings to trade carbon allowances. As such, economic and environmental benefits can both be achieved in JPC.

(2) Formulate the JPC problem as a *Partially Observable Markov Game* (POMG). To solve this POMG, a novel MARL algorithm named Fed-JPC is proposed by adopting the technique of federated learning (FL) to the deep deterministic policy gradient (DDPG) method, which can safeguard private information. Furthermore, Fed-JPC features an abstract critic network, which can capture the system dynamics and improve the training stability.

(3) Conduct extensive case studies on a real-world dataset to show that Fed-JPC achieves a better balance between low-cost and environmentally friendly transitions than the other two baseline MARL algorithms, e.g., peer-to-grid (P2G) energy trading and sole P2P energy trading. Finally, the superior performance of the proposed model-free learning-based Fed-JPC algorithm is also compared with the two model-based consensus algorithm and centralization optimization approach.

1.4. Paper organization

The rest of this paper is organized as follows. Section 2 presents the mathematical model and the problem formulation of the MES buildings, the MMR pricing scheme, and the CAT approach. Section 3 formulates the building community in the JPC mechanism as a POMG. Section 4 provides the detailed algorithm of the proposed Fed-JPC that can efficiently solve the POMG. The experiment setup and the case studies are presented in Sections 5 and 6, respectively. Finally, Section 7 discusses the conclusions of this work.

2. Problem formulation for building joint P2P energy and carbon trading

MES buildings are normally classified into different types related to their locations in residential, commercial, and industrial areas. On the one hand, they can trade electricity on the main grid and purchase natural gas as fuel from the gas grid, but they can also exchange electricity locally through the P2P trading platform. On the other hand, they are allowed to trade their carbon allowances on the carbon market. In this context, this paper proposes a joint P2P energy and carbon trading (JPC) mechanism for a building community, as depicted in Fig. 1. Specifically, the P2P energy trading mechanism is based on the mid-market rate (MMR) pricing scheme [40], which is adequately incentivized for smart energy buildings to cooperatively participate in local trading, irrespectively of whether they act as buyers or sellers. The buildings under the MMR scheme can benefit from a lower local buy price and a higher local sell price compared to the unattractive grid buy price and grid sell price, respectively. On the other hand, the carbon trading mechanism is based on the cap-and-trade (CAT) approach [41] in emissions trading scheme (ETS), which limits the total amount of carbon that can be emitted by buildings. The CAT approach makes a significant contribution to the low-carbon transition of buildings from fuel-based energy to renewable-based energy.

Each MES building is equipped with an autonomous *building energy management system* (BEMS) [42] (which is equivalent to the home energy management system in smart homes, as studied in [43,44]) that can manage its energy scheduling decisions in MES based on: (1) the grid information of energy and carbon price signals; (2) the local information of its consumption loads, renewable generation, and the status of controllable components; (3) the community information of P2P energy trading quantity; and (4) the environment information of carbon emissions. Furthermore, the above energy scheduling decisions in MES can automatically result in the net demand/generation and natural gas procurement of each building, which also reflect the trading decisions in the P2P trading platform and carbon market, respectively. Finally, the objective of these buildings is to minimize energy and environmental costs by optimally managing their MES components, so as to expand local trading activities and reduce carbon emissions.

2.1. Multi-energy system

MES buildings are operating with a synergy effect of energy consumption, supply, storage, and conversion. Specifically, the set of MES components of the examined buildings mainly includes: (1) two types of typical consumption loads: electric load (EL) and heat load (HL); (2) one flexible demand maintaining the indoor comfort level: a heating, ventilation, and air conditioning (HVAC) system; (3) one renewable-based generator: solar photovoltaic (PV); (4) two types of storage units exploiting respectively the electricity and heat flexibility: electric energy storage (EES) and thermal energy storage (TES); and (5) three types of energy converters making the conventions between multiple energy sectors: combined heat and power (CHP) engine, electric heat pump (EHP), and gas boiler (GB). This section aims at providing detailed mathematical models of the above-mentioned controllable components, as shown below.

2.1.1. Energy storage units

Energy storage units with high flexibility in buildings are characterized by their redistribution ability of off-peak and peak loads as well as their ability to absorb free renewable energy for future usage when energy prices are at their peak. More specifically, the mathematical models of an EES unit can be formulated as follows:

$$0 \le P_t^{eesc} \le \overline{P}^{ees}, \ \forall t \in T \tag{1}$$

$$-\overline{P}^{ees} \le P_t^{eesd} \le 0, \ \forall t \in T \tag{2}$$

$$0 \le E_t^{ees} \le \overline{E}^{ees}, \ \forall t \in T \tag{3}$$

$$E_{t+1}^{ees} = E_t^{ees} + P_t^{eesc} \Delta t \eta^{ees} + P_t^{eesd} \Delta t / \eta^{ees}, \ \forall t \in T \tag{4}$$

where constraints (1) and (2) limit the battery charging and discharging power P_t^{eesc}, P_t^{eesd} within its power capacity $\overline{P}^{ees}$. Constraint (3) expresses the upper bound of battery energy content, while the storage dynamic transition of battery energy content is presented in (4) taking into account the energy losses caused by the powering efficiencies $\eta^{ees} \in (0, 1]$. Then, the charging and discharging power Q_t^{tesc}, Q_t^{tesd} as well as the storage dynamic transition E_t^{tes} of a TES unit from time step t to $t+1$ can be derived similarly to the EES model (1)–(4).

2.1.2. HVAC system

The HVAC system is typically installed with buildings and is controlled to keep the indoor temperature within a desired comfort range; thus, the operation of the HVAC system is a transition from electricity power to thermal comfort, the details of which are discussed in [45]. The mathematical models of an HVAC system can be expressed as follows:

$$0 \le P_t^{hvac} \le \overline{P}^{havc}, \ \forall t \in T \tag{5}$$

$$\underline{H}^{havc} \le H_t^{in} \le \overline{H}^{havc}, \ \forall t \in T \tag{6}$$

$$H_{t+1}^{in} = H_t^{in} - \frac{(H_t^{in} - H_t^{out} + \eta^{hvac} R^{hvac} P_t^{hvac})\Delta t}{C^{hvac} R^{hvac}}, \ \forall t \in T \tag{7}$$

where the adjusted HVAC electric power P_t^{hvac} and the indoor temperature H_t^{in} are restricted in constraints (5) and (6), respectively. The equality (7) indicates the HVAC dynamic model of indoor temperature, which is related to the outdoor temperature H_t^{out}, the indoor temperature H_t^{in}, the power demand P_t^{hvac} as well as the HVAC system energy efficiency η^{hvac}, thermal capacity C^{hvac} and resistance R^{hvac}.

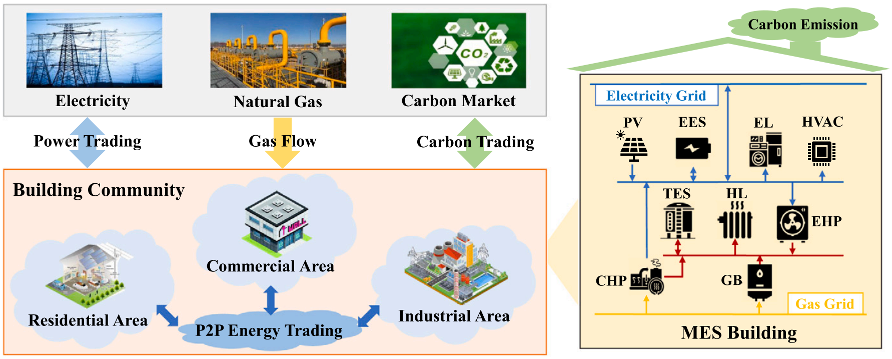


Fig. 1. Paradigm of the joint P2P energy and carbon trading (JPC) mechanism for a building community. The utilized buildings are characterized by a multi-energy system including electricity, heat, and gas sectors. Apart from local energy trading, buildings can also trade on the electricity grid, gas grid, and carbon market.

2.1.3. Energy converters

The key to modeling an MES is to capture the energy conversion relationships between different energy carriers. The examined buildings mainly consist of three types of energy converters. The CHP engine, a single-input-multi-output converter, is typically characterized by its high energy efficiency compared to independent electricity and heat sources. Therefore, it is considered a critical converter in the MES. More specifically, a CHP engine's coupled heat and electricity generation can be modeled as follows:

$$P_t^{chp} = \eta^{chpe} G_t^{chp}, \ \forall t \in T \tag{8}$$

$$Q_t^{chp} = \eta^{chpq} G_t^{chp}, \ \forall t \in T \tag{9}$$

$$0 \leq P_t^{chp} \leq \overline{P}^{chp}, \ \forall t \in T \tag{10}$$

$$0 \leq Q_t^{chp} \leq \overline{Q}^{chp}, \ \forall t \in T \tag{11}$$

where constraints (8) and (9) respectively indicate the efficiency of the CHP engine to convert natural gas into electricity and heat power, as determined by the conversion efficiency from gas to electricity η^{chpe} and the conversion efficiency from gas to heat η^{chpq}, respectively. The electric and heat output power of the CHP engine are limited by its power capacities expressed in (10) and (11), respectively.

Aside from CHP, EHP is essential for a building energy management system because it can generate heat energy while consuming electricity, as shown in Eq. (12), where η^{ehp} represents the energy conversion efficiency from electricity to heat power.

$$Q_t^{ehp} = \eta^{ehp} P_t^{ehp}, \ \forall t \in T \tag{12}$$

$$0 \leq Q_t^{ehp} \leq \overline{Q}^{ehp}, \ \forall t \in T \tag{13}$$

The vessel GB converts natural gas into heat energy. The generation of heat from natural gas via GB is given in (14) that is affected by the energy conversion efficiency from natural gas to heat power η^{gb}. Furthermore, the boiler has a limitation $\overline{Q}^{gb}$ for outputting heat power constrained by (15).

$$Q_t^{gb} = \eta^{gb} G_t^{gb}, \ \forall t \in T \tag{14}$$

$$0 \leq Q_t^{gb} \leq \overline{Q}^{gb}, \ \forall t \in T \tag{15}$$

2.2. Peer-to-peer energy trading

The mid-market-rate (MMR) method [40] is designed as an appropriate P2P trading pricing mechanism to adequately incentivize these buildings to cooperatively participate in local trading, irrespectively of whether they act as buyers or sellers. Consider a set of buildings $\mathcal{I}$, and define the community's net demand P_t^{md} and net generation P_t^{mg}, as well as its remaining energy deficit (positive) or surplus (negative) P_t^{mn} as functions of individual building net demand and generation $P_{i,t}^n$.

$$P_t^{md} = \sum_{i \in \mathcal{I}^d} P_{i,t}^n, \ P_t^{mg} = \sum_{i \in \mathcal{I}^g} P_{i,t}^n, \ P_t^{mn} = \sum_{i \in \mathcal{I}} P_{i,t}^n, \ \forall t \in T \tag{16}$$

where $\mathcal{I}^d = \{i \in \mathcal{I} : P_{i,t}^n > 0\}$ and $\mathcal{I}^g = \{i \in \mathcal{I} : P_{i,t}^n \leq 0\}$ are the sets of buildings as buyers and sellers, respectively. The net demand (positive) / generation (negative) $P_{i,t}^n$ of building i is determined by its energy portfolios and component schedules, which can be expressed as the sum of its electric power demand and generation:

$$P_{i,t}^n = P_{i,t}^l - P_{i,t}^{pv} + P_{i,t}^{hvac} + P_{i,t}^{ehp} - P_{i,t}^{chp} + P_{i,t}^{eesc} + P_{i,t}^{eesd}, \ \forall i \in \mathcal{I}, \ \forall t \in T \tag{17}$$

MMR calculates the local buy price λ_t^{m+} and local sell price λ_t^{m-} as the average of the grid sell price λ_t^- and grid buy price λ_t^+, denoted by λ_t^m. However, the total demand P_t^{md} and total generation P_t^{mg} are not always equally matched within the community. As a result, the unbalanced quantity (i.e., the community's net demand/generation) P_t^{mn} must still be traded on the main grid at the unattractive grid buy and sell prices, resulting in new adjustments of λ_t^{m+} and λ_t^{m-}. Overall, the pricing scheme of MMR can be defined as the below two scenarios:

$$\lambda_t^{m+} = \begin{cases} \lambda_t^m & \text{if } P_t^{mn} \leq 0 \\ (\lambda_t^m |P_t^{mg}| + \lambda_t^+ P_t^{mn}) / P_t^{md} & \text{if } P_t^{mn} > 0, \end{cases} \quad \forall t \in T \tag{18}$$

$$\lambda_t^{m-} = \begin{cases} \lambda_t^m & \text{if } P_t^{mn} \geq 0 \\ (\lambda_t^m P^{md} + \lambda_t^- |P_t^{mn}|) / |P_t^{mg}| & \text{if } P_t^{mn} < 0, \end{cases} \quad \forall t \in T \tag{19}$$

2.3. Carbon market

The emissions trading scheme (ETS) is a mandatory scheme for greenhouse gases and is central to the European Union's climate change target of reducing emissions by 40% by 2030 compared to 2005 levels [7]. Every building that participates in the ETS must use its carbon allowance to fully cover its emissions. In particular, ETS works on the cap-and-trade (CAT) approach [41]. On the one hand, a cap is set on the total amount of greenhouse gases that can be emitted by buildings, and these buildings can receive a certain number of free allowances under this cap. In MES, these free allowances can be allocated to the heat components using natural gas or renewable resources that generate clean energy. The remaining carbon allowance, on the other hand, with a surplus (deficit), can be sold (bought) in the carbon market. Thus, buildings must make use of their individual resources wisely to cooperatively reduce emissions, e.g., by increasing the usage of EHP using electric power rather than GB using natural gas. We assume in the community that all buildings can participate in the ETS and that those equipped with heating components and/or renewable energy sources can be assigned a certain number of free carbon allowances and trade in the carbon market. With the guidance for agents participating in the ETS, a free carbon allowance allocation method and a carbon trading mechanism are designed for these buildings in the community, with the overall structure being shown in Fig. 2.

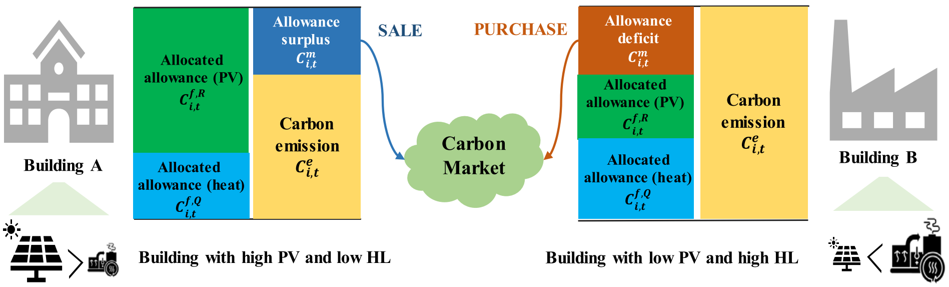

Building A
Allocated allowance (PV) $C^{f,R}_{i,t}$
Allowance surplus $C^{m}_{i,t}$
Carbon emission $C^{e}_{i,t}$
Allocated allowance (heat) $C^{f,Q}_{i,t}$
SALE
Carbon Market
PURCHASE
Allowance deficit $C^{m}_{i,t}$
Allocated allowance (PV) $C^{f,R}_{i,t}$
Allocated allowance (heat) $C^{f,Q}_{i,t}$
Carbon emission $C^{e}_{i,t}$
Building B
Building with high PV and low HL
Building with low PV and high HL

Fig. 2. Operations of the cap-and-trade (CAT) approach for different buildings. On the one hand, buildings with high PV and low HL that have an allowance surplus can sell it to the carbon market. On the other hand, buildings with low PV and high HL that have an allowance deficit need to purchase it from the carbon market.

2.3.1. Allocation of free carbon allowance

Heat energy components. In the CAT approach, it is assumed that buildings with energy components can receive a certain level of carbon allowances for free. The continuation of free allocation allows the policymaker to pursue ambitious emissions reduction targets while shielding an internationally competing industry from carbon leakage. Accordingly, energy components that use fuel to generate electricity do not receive any free allowance, while energy components that use fuel to generate heat energy will receive a free carbon allowance based on the heat benchmark. Specifically, the initial allocation of free carbon allowances for each heat benchmark component is expressed as:

$$C^{f,Q}_{i,t} = Q_{i,t} \times \delta^{Q} \times \xi^{Q}, \ \forall i \in \mathcal{I}, \ \forall t \in T \tag{20}$$

where $Q_{i,t}$ is the total heat generation of a building's energy components, i.e., $Q_{i,t} = Q^{gb}_{i,t} + Q^{chp}_{i,t}, \forall i \in \mathcal{I}$; δ^{Q} is the heat benchmark; and ξ^{Q} is the carbon leakage exposure factor (CLEF) for the heat benchmark component that depends on carbon leakage status.

PV power generation. Meanwhile, in order to incentivize emission reduction, buildings with PV sources can also receive a free carbon allowance based on grid carbon intensity to encourage them to generate more clean energy. The initial allocation of free carbon allowance for PV power generation can be expressed as:

$$C^{f,R}_{i,t} = P^{PV}_{i,t} \times \delta^{R} \times \xi^{R}, \ \forall i \in \mathcal{I}, \ \forall t \in T \tag{21}$$

where $P^{PV}_{i,t}$ denotes the total PV power generation by building i at time step t; δ^{R} denotes the carbon intensity of the electricity grid; and ξ^{R} denotes the free allocation ratio of carbon allowance for PV resources.

2.3.2. Carbon trading

After deducting the free carbon allowance allocation, the remaining carbon allowance with a surplus (deficit) can be sold (bought) in the carbon market, which can be expressed as:

$$C^{m}_{i,t} = C^{e}_{i,t} - C^{f,Q}_{i,t} - C^{f,R}_{i,t}, \ \forall i \in \mathcal{I}, \ \forall t \in T \tag{22}$$

where $C^{e}_{i,t} = \delta^{G} \times Q_{i,t}$ represents the carbon emission emitted by building i at time t based on the gas quantities purchased from the natural gas grid, and δ^{G} represents the averaged carbon emission (carbon intensity) of using natural gas.

3. Partially observable Markov game

Solving the proposed building's JPC problem faces various challenges. First, buildings operate in a decentralized manner and cannot acquire the explicit models of P2P energy trading and carbon markets, as well as the energy schedules and trading strategies of other buildings in the local community. Second, a vast number of system uncertainties and dynamics (e.g., PV generation, demand profiles, outdoor temperature) need to be handled by buildings in a decision-making process. Solving such a stochastic optimization problem is normally time-consuming. Third, even though some techniques (e.g., scenario-reduction) can accelerate the optimization process, the optimal schedules and trading decisions need to be re-optimized in a new state. In order to address the above three challenges, this paper adopts a data-driven, model-free MARL-based algorithm in a decentralized decision-making framework without a *prior* knowledge. As a result, knowledge exchanges among buildings can be avoided. Meanwhile, a coordinated scheme for these buildings can be achieved through the proper design of the reward function and a reasonable cooperative mechanism. Furthermore, the system uncertainties can be captured through the learning process by the vast interactions with the environment of the real-world dataset. Finally, once the MARL algorithms have been properly trained, the learned policies can be applied in milliseconds to practical energy scheduling and trading decisions for buildings, as well as to various system dynamics and state conditions.

Specifically, the proposed JPC problem for a building community can be formulated as a multi-agent coordination problem in the form of a finite partially observable Markov game (POMG) [27] with discrete time steps. The POMG is then defined with a set of global state $\mathcal{S}$, a collection of local observations $\{\mathcal{O}_{1:I}\}$, a collection of action sets $\{\mathcal{A}_{1:I}\}$, a collection of reward functions $\{\mathcal{R}_{1:I}\}$ and a state transition function $\mathcal{T}(s, a_{1:I}, \omega)$, where ω is the parameter describing the uncertainty of the environment. The time interval between two consecutive time steps $\Delta t = 1$ h. At time step t, each agent i chooses an action $a_{i,t}$ according to its policy $\pi_i(a_{i,t}|o_{i,t})$ conditional on its local observation $o_{i,t}$, and then executes this action $a_{i,t}$ to the environment. The environment then moves into the next state according to the transition function $\mathcal{T}$. Each agent i obtains the reward $r_{i,t}$ and the next local observation $o_{i,t+1}$. The objective of each agent i is to maximize the cumulative discounted reward $R_i = E_{s \sim \mathcal{T}, a_i \sim \pi_i}[\sum_{t=0}^{T} \gamma^t r_{i,t}]$, where $\gamma \in [0,1)$ is the discount factor and T is the daily horizon of 24 h. The detailed components of POMG in the proposed problem are expressed in the following subsections:

3.1. Agents

The agents $i \in \mathcal{I}$ are defined as the individual buildings that can directly manage their controllable components in the MES and trading strategies in the joint P2P energy trading and emission trading scheme.

3.2. Environment

The problem includes joint energy and carbon trading activities. The environment can be organized into two sectors: (1) P2P energy trading (Sections 2.1 and 2.2), which includes internal MES management, local electricity trading, unbalanced trading with the external electricity grid, and natural gas procurement from the gas grid; and (2) carbon allowance trading (Section 2.3), which includes free carbon allowance allocations and remaining carbon market trading.

3.3. Observation set

The local observation $o_{i,t}$ of each building agent i can be defined as an 8-dimensional vector:

$$o_{i,t} = [\lambda^{+}_{t}, H^{out}_{t}, H^{in}_{i,t}, P^{l}_{i,t}, Q^{l}_{i,t}, P^{pv}_{i,t}, E^{ees}_{i,t}, E^{tes}_{i,t}],\ \forall i \in \mathcal{I}, \forall t \in T \tag{23}$$

which consists of two parts: (1) the exogenous state unaffected by the action includes the sensor data of electricity grid buy prices λ^{+}_{t} and outdoor temperature H^{out}_{t} as well as the measured data of EL $P^{l}_{i,t}$, HL $Q^{l}_{i,t}$ and PV generation $P^{pv}_{i,t}$; (2) the endogenous state that serves as the feedback signals of agents' executed action and represents the system dynamics, including the energy content of EES $E^{ees}_{i,t}$ and TES $E^{tes}_{i,t}$ as well as the indoor temperature $H^{in}_{i,t}$.

3.4. Action set

Each building agent i controls its action $a_{i,t}$ that can be defined as a 5-dimensional vector:

$$a_{i,t} = [a^{ees}_{i,t}, a^{tes}_{i,t}, a^{hvac}_{i,t}, a^{chp}_{i,t}, a^{ehp}_{i,t}],\ \forall i \in \mathcal{I}, \forall t \in T \tag{24}$$

where $a^{ees}_{i,t}, a^{tes}_{i,t} \in [-1, 1]$ indicating the mutually exclusive charging (positive) and discharging (negative) power rate of EES and TES as a percentage of their power capacity $[-\overline{P}^{ees}, \overline{P}^{ees}]$ and $[-\overline{P}^{tes}, \overline{P}^{tes}]$, respectively. $a^{hvac}_{i,t}, a^{chp}_{i,t}, a^{ehp}_{i,t} \in [0, 1]$ indicating the magnitude of power schedules as a percentage of their power capacity $[0, \overline{P}^{hvac}]$, $[0, \overline{Q}^{chp}]$, and $[0, \overline{Q}^{ehp}]$, respectively.

It is noted that, as a backup component, the power out of GB is not directly controlled by action. Instead, it can be automatically derived once the power outputs of all other heat components are determined, i.e., $Q^{gb}_{i,t} = Q^{l}_{i,t} - Q^{ehp}_{i,t} - Q^{chp}_{i,t} + Q^{tesc}_{i,t} - Q^{tesd}_{i,t}, \forall i \in \mathcal{I}, \forall t \in T$. In this setting, the heat demand-supply balance constraint of each building i at each time step t can be always guaranteed in the RL environment.

3.5. State transition

The state transition from time step t to $t+1$ is governed by $s_{t+1} = \mathcal{T}(s_t, a_{1:\mathcal{I},t}, \omega_t)$, influenced by the combination of environment state s_t, all agents' actions $a_{1:\mathcal{I},t}$, and environment stochasticity ω_t. In the examined problem, $\omega_t = [\lambda^{+}_{t}, H^{out}_{t}, P^{l}_{t}, Q^{l}_{t}, P^{pv}_{t}]$ is decoupled from the agents' actions and is characterized by inherent variability. In the machine learning area, RL translates this problem to a data-driven approach that learns the stochastic characteristics directly from the data sources [27].

By contrast, the state transitions of endogenous states $E^{ees}_{i,t}, E^{tes}_{i,t}$ are determined by actions $a^{ees}_{i,t}, a^{tes}_{i,t}$. Given EES as an example, the mutually charging and discharging power quantities $P^{eesc}_{i,t}, P^{eesd}_{t}$ are managed by action $a^{ees}_{i,t}$, and are also restricted by its technical parameters of power and energy capacities $\overline{P}^{ees}_{i}, \overline{E}^{ees}_{i}$, and the powering efficiency η^{ees}_{i}, which can be expressed as:

$$P^{eesc}_{i,t} = [\min(a^{ees}_{i,t}\overline{P}^{ees}_{i}, (\overline{E}^{ees}_{i} - E^{ees}_{i,t})/(\eta^{ees}_{i}\Delta t))]^{+},\ \forall i \in \mathcal{I}, \forall t \in T \tag{25}$$

$$P^{eesd}_{i,t} = [\max(a^{ees}_{i,t}\overline{P}^{ees}_{i}, (-E^{ees}_{i,t}\eta^{ees}_{i})/\Delta t)]^{-},\ \forall i \in \mathcal{I}, \forall t \in T \tag{26}$$

where operators $[\cdot]^{+/-} = \max/\min\{\cdot, 0\}$. Then, given the charging and discharging power $P^{eesc}_{i,t}, P^{eesd}_{i,t}$ and efficiency η^{ees}_{i}, the state transition of $E^{ees}_{i,t}$ from time step t to $t+1$ can be expressed as Eq. (4). As a consequence, charging and discharging power $Q^{tesc}_{i,t}, Q^{tesd}_{i,t}$ as well as the state transition of $E^{tes}_{i,t}$ from time step t to $t+1$ of TES can be derived in the similar manner as the EES model. Finally, based on the HVAC model formulated in Section 2.1.2 and give the HVAC power $P^{hvac}_{i,t}$, which is defined by action $a^{hvac}_{i,t}$, the state transition of $H^{in}_{i,t}$ from time step t to $t+1$ of HVAC can be expressed as Eq. (7).

3.6. Reward function

After determining the energy schedules of all MES components, the electricity net demand or generation of each building $P^{n}_{i,t}$ defined in (17) can be calculated and submitted to the P2P energy trading market. The market operator then clears the local energy trading and calculates the total demand P^{md}_{t}, total generation P^{mg}_{t}, and net quantity P^{mn}_{t} for the building community, as well as the corresponding MMR prices $\lambda^{m+}_{t}, \lambda^{m-}_{t}$, as defined in Section 2.2. Meanwhile, the CAT approach (described in Section 2.3) allocates free carbon allowances for heat energy $C^{f,Q}_{i,t}$ and renewable energy $C^{f,R}_{i,t}$ before balancing the remaining allowance quantities $C^{m}_{i,t}$ in the carbon market at the daily carbon price λ^{c}. As a result of the foregoing, the reward function for each building agent i at time step t can be designed in three parts: (1) the negative electricity cost; (2) the negative gas cost; and (3) the negative environmental cost.

$$r_{i,t} = -(\lambda^{m+}_{t}[P^{n}_{i,t}]^{+} + \lambda^{m-}_{t}[P^{n}_{i,t}]^{-}) - \lambda^{g}G^{g}_{i,t} - \lambda^{c}C^{m}_{i,t},\ \forall i \in \mathcal{I}, \forall t \in T \tag{27}$$

where $G^{g}_{i,t} = G^{chp}_{i,t} + G^{gb}_{i,t}$ indicates the natural gas quantity purchased from the gas grid through the CHP and GB, and λ^{g} is the gas price.

4. Federated multi-agent reinforcement learning method

To efficiently solve the above POMG, we propose a novel MARL method named Fed-JPC, with its general architecture shown in Fig. 3. Specifically, Fed-JPC derives three concrete implementation details that are insightful and particularly critical to our proposed problem: (1) featuring an actor–critic architecture based on the deep deterministic policy gradient (DDPG) method [46], with a policy (actor) network outputting continuous actions and a Q-value (critic) network correcting the policy network' weights; (2) integrating the federated reinforcement learning (FRL) framework [39] with DDPG method to achieve distributed control while maintaining privacy; and (3) introducing an abstract critic network that additionally involves the community net quantity and individual carbon emission, apart from the local observations and actions, in order to stabilize the training performance by capturing the system dynamics.

For a better understanding of Fed-JPC in Fig. 3, we show five steps inside the algorithm, which can be summarized as follows.

Step 1: The building agent i executes an action a_i to the environment based on the output of the actor network $\mu_{\phi_i}(a|o)$ while observing the local observation o_i.

Step 2: Once the action a_i is executed, the energy flows in the building will be calculated based on the MES model (Section 2.1). The P2P energy trading price and quantity will be calculated based on the MMR method (Section 2.2). The carbon emission and trading quantity will be calculated based on the CAT approach (Section 2.3). Afterwards, each building agent i can observe its local observation o_i. The community net quantity P^{mn} and carbon emission C^{e}_{i}, as well as the local observation o_i, executed action a_i, resulted reward r_i, and next local observation o'_i will be stored to the experience tuple.

Step 3: Each building agent i samples a mini-batch of experiences from the replay buffer and uses them to update the local model ω_i of both the actor network ψ_i and the critic network θ_i.

Step 4: After training the local model, each building agent i will transits its model ω_i to the central aggregator.

Step 5: The central aggregator collects the local model ω_i from individual agents, creates the aggregation ω^{a}, and then broadcasts it to individual agents for action execution.

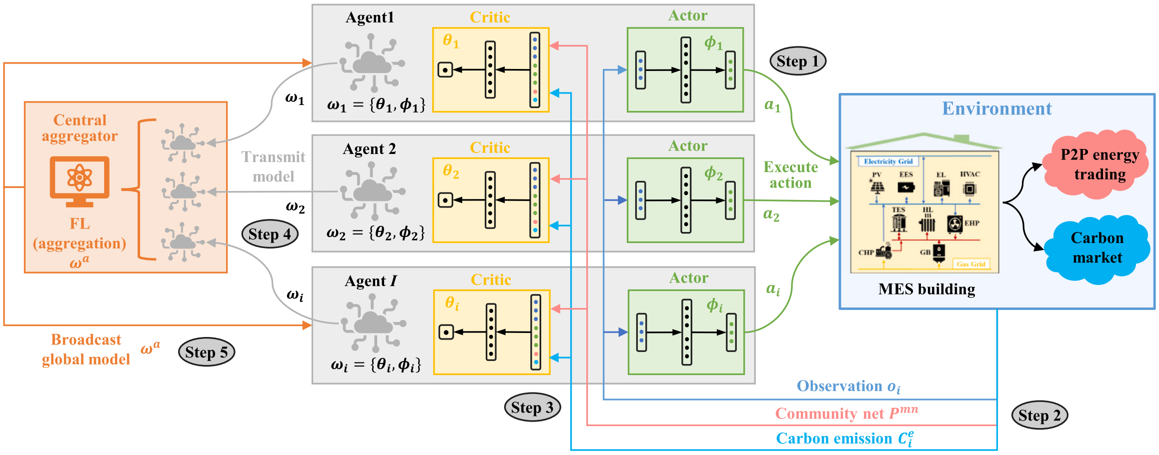


Fig. 3. Architecture of the proposed Fed-JPC method.

4.1. Federated reinforcement learning

Since we are considering a POMG of I agents with the same observation, action and reward function, we can train their policies more efficiently using the federated reinforcement learning (FRL) concept [39]. More specifically, FRL is a distributed learning framework for training actor–critic network models of local buildings without sharing local information. The training process for FRL consists of local model training and global aggregation of updated local models. As defined in Section 3, let $\mathcal{I}$ be the set of community buildings. Using a dataset $\mathcal{D}_i$, each building i constructs and trains its own actor–critic network model $\omega_i = \{\phi_i, \theta_i\}$. Following the completion of the training process for each local building, the model of each building is transmitted and aggregated by the central aggregator in order to estimate a global model $\omega^a = \{\phi^a, \theta^a\}$ of all local buildings, which is expressed as follows: $\omega^a = f(\omega_1, \omega_2, \dots, \omega_I)$. This model ω^a is then broadcast to all local buildings whose models are replaced by ω^a: $\omega^a = \omega_1 = \omega_2 = \dots = \omega_I$. Finally, each local building restarts its training based on ω^a until they all obtain the desired model.

It is noteworthy that there are two major motivations behind the FRL algorithm adopted for our building JPC problem. Firstly, many MARL algorithms have the problem of low learning efficiency caused by complex tasks and incomplete information. Although centralized training allowing information exchange between agents can improve learning efficiency, it needs to collect all the local information and thus break the agents' privacy. FRL, as a safe algorithm, introduces a central aggregator that can complete information exchange between the central aggregator and local agents while avoiding direct information exchange among local agents, thereby protecting privacy. Second, agents in a POMG can only observe partial local information (e.g., demand and PV generation), and some sufficient information (e.g., community trading quantity) capturing system critical dynamics cannot be observed. FRL makes it possible to integrate this information through aggregation and broadcast it to the individual agents so they can make informed decisions.

4.2. Deep deterministic policy gradient

The proposed method is based on the actor–critic architecture that contains two networks for different purposes. The actor network $\mu_{\phi_i}(o_i)$, parameterized by ϕ_i, takes as input the local observation o_i and outputs the continuous action a_i for each agent i. The critic network $Q_{\theta_i}(o_i, a_i)$, parameterized by θ_i, takes as input the concatenation of local observation o_i and executed action a_i of agent i, and outputs a scalar estimate of the Q-value to perform the policy evaluation task. More specifically, we update the weights of the critic network with temporal difference (TD) learning [27] as:

$$\mathcal{L}(\theta_i) = \mathbb{E}\left[(r_i + \gamma Q'_{\theta'_i}(o'_i, a'_i) - Q_{\theta_i}(o_i, a_i))^2\right], \ \forall i \in \mathcal{I} \tag{28}$$

where $Q'_{\theta'_i}(\cdot)$ is the target critic network whose parameters θ'_i are updated by softly tracking the online critic network θ_i, to give consistent target during TD learning. Furthermore, a'_i is the executed action according to the next observation o'_i of agent i using its target actor network $\mu'_{\phi'_i}(o'_i)$ whose parameters ϕ'_i are also updated by softly tracking the online actor network ϕ_i. As opposed to TD learning in the critic network, the actor network is updated via the deterministic policy gradient theorem [27] as:

$$\nabla_{\phi_i} J(\mu_{\phi_i}) = \nabla_{\phi_i} \mu_{\phi_i}(o_i) \nabla_{a_i} Q_{\theta_i}(o_i, a_i)|_{a_i = \mu_{\phi_i}(o_i)}, \ \forall i \in \mathcal{I} \tag{29}$$

4.3. Abstract critic network

Applying the above DDPG method directly to a multi-agent setup may be difficult because the independently learning algorithm treats other agents as part of the environment, which appears non-stationary from the perspective of any agent. As a result, in existing MARL algorithms that can stabilize training performance, a centralized Q-value $Q_{\theta_i}(o_{1:I}, a_{1:I})$ with access to all agents' local observations and actions is widely used. However, it is difficult to directly acquire other agents' local observations and actions in our proposed problem, since buildings with privacy concerns are not willing to exchange their energy schedules and trading activities with each other. To this end, this paper assumes the market operator as a trusted third party that can make use of community information and carbon emissions and then incorporate them into the centralized Q-value function to epitomize the key information of the system dynamics. The centralized Q-value function for each agent i in this context can be approximated as follows:

$$Q_{\theta_i}(o_{1:I}, a_{1:I}) \approx Q_{\theta_i}(o_i, a_i, P^{mn}, C_i^e), \ \forall i \in \mathcal{I} \tag{30}$$

where P^{mn} is the amount of community net demand and generation and C_i^e is the amount of carbon emissions emitted by building i. It is clear that P^{mn} is an embedded function that captures the trading dynamics through the community net quantity balanced with the upstream utility company. Specifically, it can be observed from (16) that P^{mn} is the summation of all buildings' net load P_i^n defined in (17), while P_i^n is also a result of the individual's local observations (e.g., EL and PV generation) and actions (e.g., power schedules of HVAC systems, power input of EHP, power output of CHP, and charging/discharging power of EES). As a result, P^{mn} can somehow capture the local observations and actions of other buildings in the community. More importantly, P^{mn} is critical for buildings to adjust their energy schedules and trading strategies, as its positive value represents the community's net demand, encouraging buildings to generate (sell) more energy surplus, while its negative value represents the community's net generation, encouraging buildings to consume (buy) more energy deficit. On the other hand, C_i^e is also an informed index that can capture the effect of a building's energy schedules on carbon emissions and the associated environmental costs. The higher value of C_i^e encourages building i to reduce the procurement from the natural gas grid, and vice versa. In conclusion, by incorporating P^{mn}, C_i^e into the centralized Q-value function, agent i can make informed decisions based on the impact of other agents' actions in the community, while albeit not knowing their specific information, thereby protecting the privacy of the buildings and improving scalability.

4.4. Parameter update for actor–critic networks

Fed-JPC is an off-policy MARL method that requires past experience to update the networks. As a result, an experience replay buffer $\mathcal{D}_i$ is employed. The buffer is a cache storing the past experiences of all agents acquired from the environment. In detail, an experience is a transition tuple that contains $e_{i,t} = (o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1})$ used to update policy and P^{mn}, C_i^e used to abstract the Q-value function. On every iteration of training process, we sample uniformly a minibatch of J mixed experiences from the replay buffer $\{(e_j, P_j^{mn}, C_j^e)\}_{j=1}^J \sim \mathcal{D}_i$ to compute the mean-squared TD error of online critic network:

$$\mathcal{L}(\theta_i) = \frac{1}{J}\sum_{j=1}^{J}\Big[\big(y_j - Q_\theta(o_j, a_j, P_j^{mn}, C_j^e)\big)^2\Big],\ \forall i \in \mathcal{I} \tag{31}$$

where the target Q-value:

$$y_j = r_j + \gamma Q'_{\theta'_i}\big(o_{j+1}, \mu'_{\phi'_i}(o_{j+1}), P_{j+1}^{mn}, C_{j+1}^e\big),\ \forall i \in \mathcal{I} \tag{32}$$

here $Q'_{\theta'_i}(\cdot)$ and $\mu'_{\phi'_i}(\cdot)$ are respectively the target critic and actor networks of agent i, softly updated with their online networks. P_{j+1}^{mn} is calculated from the P2P trading market given the target actions $\mu'_{\phi'}(o_{j+1})$ conditioned on next observations o_{j+1}.

On the other hand, the online actor network employs the deterministic policy gradient theorem, which can be expressed as:

$$\nabla_{\phi_i} J(\mu_{\phi_i}) = \frac{1}{J}\sum_{j=1}^{J}\big[\nabla_{\phi_i}\mu_{\phi_i}(o_j)\nabla_{a_j} Q_{\theta_i}(o_j, a_j, P_j^{mn}, C_j^e)|_{a_j=\mu_{\phi_i}(o_j)}\big],\ \forall i \in \mathcal{I} \tag{33}$$

The following updates are then applied to the weights of the online and target networks, where $\alpha^\theta, \alpha^\phi$ are the learning rates of gradient descent algorithm for online critic and actor networks, and τ is the soft update rate for target networks.

$$\theta_i \leftarrow \theta_i - \alpha^\theta \nabla_{\theta_i}\mathcal{L}(\theta_i) \ \text{ and } \ \theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i,\ \forall i \in \mathcal{I} \tag{34}$$

$$\phi_i \leftarrow \phi_i + \alpha^\phi \nabla_{\phi_i} J(\mu_{\phi_i}) \ \text{ and } \ \phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i,\ \forall i \in \mathcal{I} \tag{35}$$

Moreover, in order to help the agents explore the environment and acquire more valuable experiences, we add a random Gaussian noise $\mathcal{N}(0, \sigma_{i,t}^2)$ to the online actor network (policy) $\mu_{\phi_i}(o_{i,t})$, constructing an exploration policy:

$$\hat{\mu}(o_{i,t}) = \mu_{\phi_i}(o_{i,t}) + \mathcal{N}(0, \sigma_{i,t}^2),\ \forall i \in \mathcal{I}, \forall t \in T \tag{36}$$

Finally, the pseudo-code of the proposed Fed-JPC is presented in Algorithm 1:

Algorithm 1 Training process of Fed-JPC for I agents

```
Initialize parameters $\phi_i$ and $\theta_i$ for online actor and critic networks for each agent $i$
Copy online networks to their respective target network weights $\phi'$ and $\theta'$ for each agent $i$
Initialize the replay buffer $\mathcal{D}_i$ for each agent $i$
for episode (i.e., day) = 1 to $M$ do
    Initialize a random process $\mathcal{N}(0, \sigma_{i,t}^2)$ for action exploration
    Initialize global state $s_0$ and local observation $o_{i,0}$
    for time step (i.e., hour) $t = 1$ to $T$ do
        For each agent $i$, selects action $a_{i,t} = \hat{\mu}(o_{i,t})$ according to current local observation $o_{i,t}$ using (36)
        Execute all agents' actions $a_t = [a_{1,t}, ..., a_{I,t}]$ to the environment
        Calculate energy flows, MMR prices, carbon emission, energy and trading quantities, reward values
        for agent (i.e., building) $i = 1$ to $I$ do
            Observes current reward $r_{i,t}$ and next local observation $o_{i,t}$, and constructs an experience tuple $(o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1})$
            Concatenates community net quantity $P_t^{mn}$, carbon emission $C_{i,t}^e$ and local experience $e_{i,t}$ together and stores them $(e_{i,t}, P^{mn}, C_{i,t}^e)$ to replay buffer $\mathcal{D}_i$
            Samples a minibatch $(e_j, P_j^{mn}, C_j^e)_{j=1}^J$ from replay buffer $\mathcal{D}_i$
            Updates the online critic and actor networks in (31) and (33)
            Updates the weights of online and target networks in (34)-(35)
            Transits local model $\omega_i = \{\phi_i, \theta_i\}$ to the central aggregator
        end for
        Update state $s_t \leftarrow s_{t+1}$ and local observation $o_{i,t} \leftarrow o_{i,t+1}$
        Central aggregator makes the model aggregation $\omega^a = \{\phi^a, \theta^a\}$ and then broadcasts it to all local buildings
    end for
end for
```

5. Experiment setup and dataset

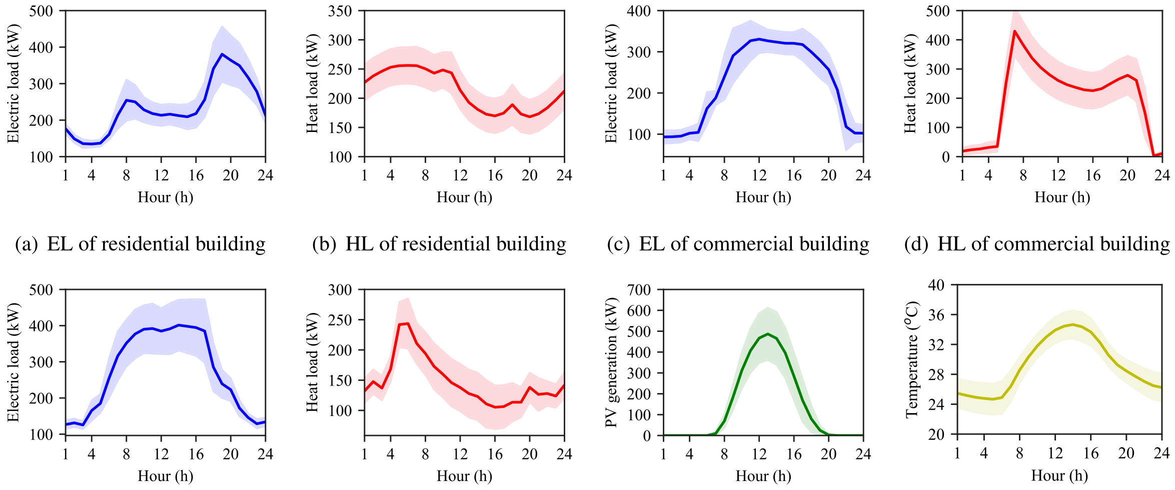
5.1. Data source

We implement all the simulations on a real-world open-source dataset recorded by the Open Energy Data Initiative (OEDI) [47] and RWTH Aachen University [48]. We collect the corresponding one-year electric and heat loads and PV generation of residential, commercial, and industrial users with hourly resolution for our simulations. The energy users can then be classified and aggregated into three types of MES buildings, with their load and generation profiles depicted in Fig. 4(a)–(g). The outdoor temperature data is collected from [49] and is plotted in Fig. 4 (h). The technical parameters of MES controllable components are derived from [50] and presented in Table 1. Table 2 shows the Time-of-Use (ToU) tariff [51] as the grid electricity buy price varying for the time, while the Feed-in-Tariff (FiT) as the grid electricity sell price and the natural gas price are flat over the day at 0.0395 £/kWh and 0.0325 £/kWh, respectively. The carbon parameters are taken from the policy documents [52] and are shown in Table 3. In order to evaluate the impact of system uncertainties and the generalization of the MARL algorithm, we split the one-year dataset into the training set (Jan.–Nov.) and the test set (Dec.).

5.2. Benchmarks

To validate the proposed Fed-JPC's superior performance in achieving economic and environmental benefits, we compare it to two benchmark mechanisms:

(1) Ind-P2G (independent peer-to-grid): Buildings are only allowed to buy (sell) energy deficits (surplus) in the electricity grid, while local energy trading does not exist. The energy supplies for heating loads are purchased from the natural gas grid. The carbon allowance is not considered in this mechanism, so the environment cost in Eq. (27) will be removed when designing the reward function. Under the P2G mechanism, each building agent uses an independent actor–critic network to optimize the energy schedules of each MES controllable component. The critic network for agent i involves the local observation o_i and action a_i in such a way that $Q_{\theta_i}(o_i, a_i)$, where θ_i denotes its parameters. The actor network $\pi_{\phi_i}(o_i)$, parameterized by ϕ_i, takes the local observation o_i as input and outputs the continuous action a_i.



(a) EL of residential building (b) HL of residential building (c) EL of commercial building (d) HL of commercial building

(e) EL of industrial building (f) HL of industrial building (g) PV generation (h) Outdoor temperature

Fig. 4. Mean (line) and standard deviation (error) of electrical load (a,c,e), heat load (b,d,f), and PV generation (g) for three buildings, as well as the outdoor temperature (h).

Table 1
Technical parameters of MES controllable components.

Component	Capacity	Efficiency
EES	$\overline{P}^{ees}, \overline{E}^{ees}$ = 90 kW, 350 kWh	η^{ees} = 0.95
TES	$\overline{Q}^{tes}, \overline{E}^{tes}$ = 75 kW, 300 kWh	η^{tes} = 0.9
GB	$\overline{Q}^{gb}$ = 500 kW	η^{gb} = 0.8
EHP	$\overline{Q}^{ehp}$ = 400 kW	η^{ehp} = 3
CHP	$\overline{P}^{chp}, \overline{Q}^{chp}$ = 200 kW, 300 kW	$\eta^{chp,e}, \eta^{chp,q}$ = 0.3, 0.45
HVAC[a]	$\overline{P}^{hvac}, \underline{H}^{hvac}, \overline{H}^{hvac}$ = 120 kW, 19 °C, 24 °C	$\eta^{hvac}, C^{hvac}, R^{hvac}$ = 2.2, 0.33 kWh/°F, 13.5°F/kW

[a] °F = °C * 1.8 + 32.

Table 2
Structure of ToU tariff.

Structure	Shoulder	Peak	Off-peak
Time (h)	9:00-16:00	17:00-20:00	21:00-8:00 (next day)
Value (£/kWh)	0.12	0.25	0.05

Table 3
Carbon parameters.

Parameter	Value
δ^Q, δ^R, δ^G	0.17028, 0.21233, 0.368 kgCO2e/kWh
ξ^Q, ξ^R	0.3, 0.45

(2) Fed-P2P (federated learning-based peer-to-peer): Buildings can form a local community for P2P energy trading under the MMR pricing scheme. Besides, they are still allowed to trade the unbalanced quantities in the electricity grid. In an analogy to the P2G mechanism, the energy supplies for heating loads are purchased from the natural gas grid, and the carbon allowance is ignored. Each P2P building agent, on the other hand, employs an abstract critic that also includes the community net quantity P^{mn}, resulting in $Q_{\theta_i}(o_i, a_i, P^{mn})$.

In addition to the above two data-driven learning-based methods, the two classic model-based optimization methods will be also evaluated to compare with our proposed Fed-JPC method.

(1) Consensus [12,14,24]: Each building employs a consensus-based alternating direction method of multipliers (ADMM) approach to solve the distributed coordination energy management problem of the building community, assuming perfect knowledge of the building system models and technical parameters, as well as perfect forecasting of system uncertainties.

(2) Centralization [9]: A central daily optimization is proposed for the building community, with the objective function being the minimization of the community's joint energy cost and carbon cost, subject to the operation constraints of the MES model and the community net demand-supply balance, assuming perfect knowledge of the system models and technical parameters, as well as perfect forecasting of system uncertainties.

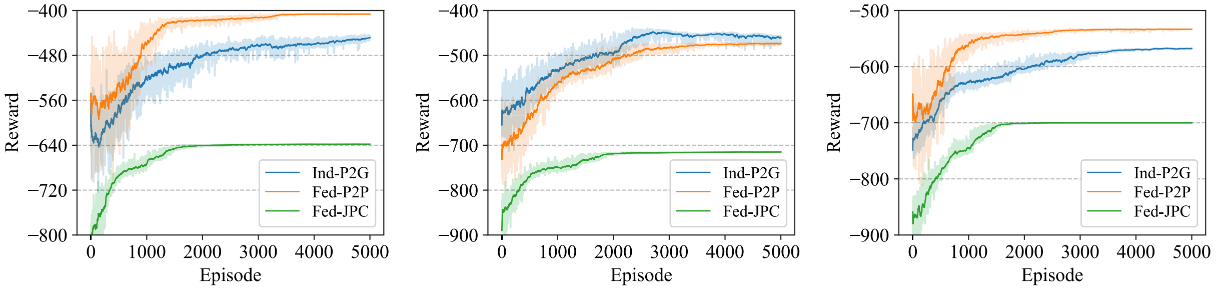
5.3. MARL algorithms settings and training details

Similar to the Fed-JPC algorithm proposed in Section 4, we also adopt the FRL framework and DDPG method to the P2P mechanism, resulting in a new benchmark MARL algorithm, Feb-P2P. We do not adopt the FRL framework to the P2G mechanism, since it is an independent trading scheme without forming into a community. As a result, in the P2G mechanism, we use the conventional independent DDPG method to solve the energy management decisions for three building agents, forming Ind-P2G.

Network structures. The detailed specifications of actor and network networks for each MARL algorithm are presented in Table 4. For all three MARL methods, both actor and critic networks have one hidden layer with 64 units using RELU as the activation function. By constructing the deterministic policy, the actor network for all three MARL algorithms takes in the (S_DIM = 8 dimensions) local observations and outputs the (A_DIM = 5 dimensions) continuous actions. For critical networks, linear is used without an activate function for the output layer, while its input, however, varies for different mechanisms. In particular, Ind-P2G inputs local observation and action; Feb-P2P additionally incorporates the community net quantity; and Fed-JPC further involves carbon emission. When executing actions to the environment, the values of the actions for EES and TES are normalized to $[-1, 1]$ from $[0, 1]$ that represent their respective discharging $[-1, 0)$ (negative)

Table 4
The general specifications of three MARL algorithms.

Mechanism	Actor Network	Critic Network
Ind-P2G	linear(s_dim, 64) → ReLU() → sigmoid(64, a_dim)	linear(s_dim+a_dim, 64) → ReLU() → linear(64, 1)
Feb-P2P		linear(s_dim+a_dim+1, 64) → ReLU() → linear(64, 1)
Fed-JPC		linear(s_dim+a_dim+2, 64) → ReLU() → linear(64, 1)



(a) residential building (b) commercial building (c) industrial building

Fig. 5. Learning curves of three building agents (a–c) for different MARL algorithms.

and charging $(0, 1]$ (positive) power schedules. For all three MARL algorithms, we run 5000 episodes with the same 10 random seeds for the environment and model initialization.

Hyperparameters. All three MARL algorithms are trained with online training (i.e., the online actor and critic networks are updated per time step), and the target actor and critic networks are updated every two intervals, which is twice the update interval of the online actor and critic networks introduced above. We use the Adam optimizer [53] for both actor and critic networks with a learning rate of α^{ϕ} = 1e-4 and α^{θ} = 1e-3, respectively. The size of the replay buffer is 5000, and the batch size of training is 32. The discount rate $\gamma = 0.9$ is used to expect a long-term return within a trading day of 24 time steps. We update the network parameters after the replay buffer is full via a random policy. During the training process, a fixed standard deviation of 0.1 is applied to the exploration. The soft update rate is τ = 1e-2. A sigmoid activation function is used for the actor output.

5.4. Evaluation metrics

In simulations, we evaluate the performance of algorithms for the building community with the following three metrics: (1) ***Electricity cost (EC)***: It calculates the electricity cost (the first two terms in Eq. (27)); (2) ***Gas cost (GC)***: It calculates the natural gas cost (the third term in Eq. (27)); (3) ***Carbon cost (CC)***: It calculates the environment cost in carbon market (the last term in Eq. (27)). We aim to find the MARL control policy that corresponds to low EC, GC, and CC.

6. Case studies

6.1. Training performance

This section aims at comparing the training performance of three examined MARL algorithms for three building agents. Fig. 5 illustrates the evolution of the episodic reward of three buildings over 5000 episodes for different MARL algorithms, where the solid lines and the shaded areas respectively depict the moving average over 100 episodes and the oscillations of the reward during the training process.

The first observation from Fig. 5 is that the reward levels in the three MARL algorithms exhibit an increasing trend for all three buildings within 5000 episodes, which means all three building agents can learn an optimal control policy to optimize their energy management problem and trading decisions for each MARL algorithm. Specifically, for each building agent, the reward in Fed-P2P is at the highest level, followed by Ind-P2G, and the lowest in Fed-JPC. This is because when training Ind-P2G and Fed-P2P, the term “environment cost” in Eq. (27) is not considered in their reward functions.

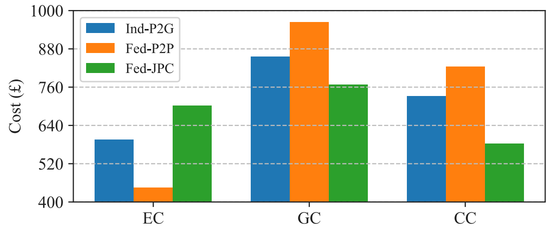


Fig. 6. Comparison of three metrics (EC, GC, CC) for different MARL algorithms.

As a result, to better evaluate the performance of three MARL algorithms, we collect the three examined metrics for different algorithms, as illustrated in Fig. 6. It can be observed that Fed-P2P, with its P2P energy trading mechanism, exhibits a lower EC compared to Ind-P2G. This, however, leads to a higher GC along with CC due to the declining dependence on the electricity grid and the increasing usage of natural gas. Fed-JPC, on the other hand, can alleviate the high GC and CC in Fed-P2P and produce a more environmentally friendly policy through the CAT approach, but with an increasing EC after importing more clean but expensive electricity energy. In conclusion, we can find that each MARL algorithm has its own focus (e.g., Fed-P2P only cares about EC), but Fed-JPC overall exhibits the best performance among the three algorithms that can achieve the lowest energy and environment cost (i.e., the sum of EC, GC, and CC). In relative terms, the proposed Fed-JPC achieves for the community 5.87% and 8.02% lower total energy and environmental costs over Ind-P2G and Fed-JPC, respectively.

6.2. Test performance

Once the control policies of the three examined MARL algorithms are well-trained, we collect the learned models and their associated weights from the actor networks and deploy them to three building agents to execute their energy scheduling actions $a_{i,t} = \mu_{\phi_i}(o_{i,t})$ to the environment in observing local observations $o_{i,t}$ at each time step t over the 31 test days. We further use the model-based consensus algorithm and the centralization optimization approach discussed in Section 5.2 to solve the studied building coordination problem per test day. Table 5 records the averaged daily EC, GC, CC and total (energy

Table 5
Averaged daily costs and computation time over 31 test days for different MARL and optimization methods.

Method	Ind-P2G	Fed-P2P	Fed-JPC	Consensus	Centralization
EC (£)	618	469	731	695	722
GC (£)	871	988	782	812	768
CC (£)	755	851	597	665	581
Total Cost (£)	2,244	2,308	2,110	2,172	2,071
Computation (sec.)	1.03	1.18	1.07	152.26	3.36

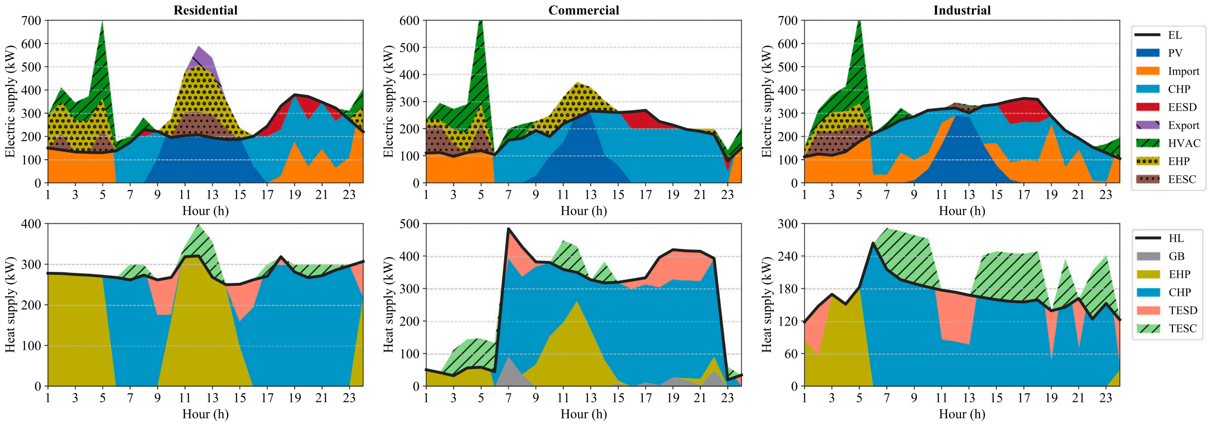


Fig. 7. Energy sources for the electric and heat power of three buildings in Ind-P2G. On the top three plots for the electric side, the black line represents the electric load of each building; the blue area represents the PV generation; the orange area represents the electricity import from the main grid; the cyan area represents the electric power output of CHP; the red area represents the discharge power of EES; the purple area represents the electricity export to the main grid; the green area represents the electricity consumption of HVAC; the olive area represents the electric power input of EHP; and the brown area represents the charge power of EES. On the bottom three plots for the heat side, the black line represents the heat load of each building; the gray area represents the heat power output of GB; the olive area represents the heat power output of EHP; the cyan area represents the heat power output of CHP; the salmon area represents the discharge power of TES; and the lightgreen area represents the charge power of TES.

and environmental) cost as well as the computation time over the 31 test days for three MARL algorithms and two optimization methods.

Regarding test performance in Table 5, it can be observed that EC, GC, and CC, and the associated total cost follow the same trends as training performance in Fig. 6. Specifically, among the three MARL algorithms, Fed-JPC obtains the lowest total cost (2,110 £) driven by the lowest gas cost GC (782 £) and the lowest carbon cost CC (597 £). It can also be found that the proposed Fed-JPC achieves a near-to-optimal performance (1.88% higher total cost than Centralization) and outperforms Consensus at 62 £the averaged total cost over 31 test days. On the other hand, all three MARL methods can be deployed in real-time at around 1.0 s, while the optimization-based Consensus and Centralization require around 150.3 and 3.4 s, respectively, to compute solutions.

6.3. Building multi-energy system management

Having evaluated the superiority of Fed-JPC over the other two baseline (Ind-P2G and Fed-P2P) algorithms during the training process, this section provides more detailed results on both electric and heat energy supplies of three examined buildings for different MARL algorithms, as depicted in Figs. 7–9. Before analyzing the multi-energy schedules, we try to provide the characteristics of three different buildings in the community. First, the residential building features abundant PV resources at midday and high EL peaks at night, as well as a relatively flat HL profile. Second, the commercial building with scarce PV resources is characterized by a high HL and a relatively flat EL profile. Third, the industrial building with scarce PV resources is characterized by high EL at midday and night and a relatively flat and low HL profile.

Ind-P2G. First of all, the HVAC system in the residential building is learned to operate during the evening and early morning to keep the building warm since the outdoor temperature is still relatively low. Second, due to the low off-peak grid buy prices (0.05 £/kWh), a large amount of electricity is purchased from the main grid (import) during the evening and early morning. And this amount of electricity is used to charge EES and supply HL via EHP. Thirdly, the residential building also learns to charge EES and use EHP during the midday periods when PV is abundant. It is noted that a proportion of PV resources are sold back (exported) to the main grid since PV is not allowed to be balanced in the P2G mechanism. The commercial building's HVAC system, like that of the residential building, is learned to operate in the evening and early morning hours. There are scarce PV resources, so CHP is activated to produce more electricity to support EL. In addition, CHP is also used to supply a large amount of HL. Since HL is very low in the evening, the commercial building learns to charge TES via EHP in the evening and discharge to HL at midday and at night. Finally, GB is used as a backup component to supply the left half of HL for a few hours during the day. In comparison to the residential and commercial buildings, the industrial building has PV resources that are even lower than the EL. Thus, the use of EHP, which converts electricity to heat, is reduced. The industrial building then increases its usage of CHP and also has to purchase more electricity from the main grid.

Fed-P2P. Following the introduction of the P2P energy trading mechanism in Fed-P2P, free PV resources are prioritized for distribution to the entire building community. In this algorithm, the residential building learns to reduce: (1) the PV surplus sold to the main grid; (2) the EES self-charging behavior in the mid-day; and (3) the usage of EHP for supplying HL. As a result, the procurement of natural gas (CHP) for the residential building has increased. The interesting result shows that the commercial building increases the usage of EHP due to the possibility of P2P trading with the residential building. This is because the HL in the commercial building is much higher. It is prioritized to support the HL of the commercial building via the EHP. The scheduling behaviors of the industrial building do not change much.

Fed-JPC. The first observation in Fed-JPC is that the use of EHP has increased significantly in all three buildings. This is largely driven by the introduction of the CAT approach, which has made the building community more concerned about the carbon emissions raised by the

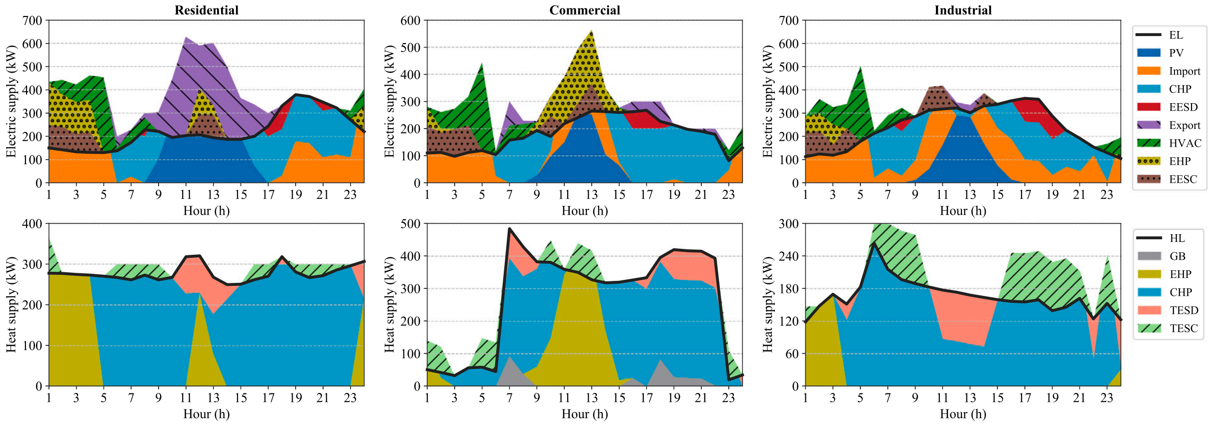


Fig. 8. Energy sources for the electric and heat power of three buildings in Fed-P2P. The two legends' black lines and different colored areas represent the same information as Fig. 7.

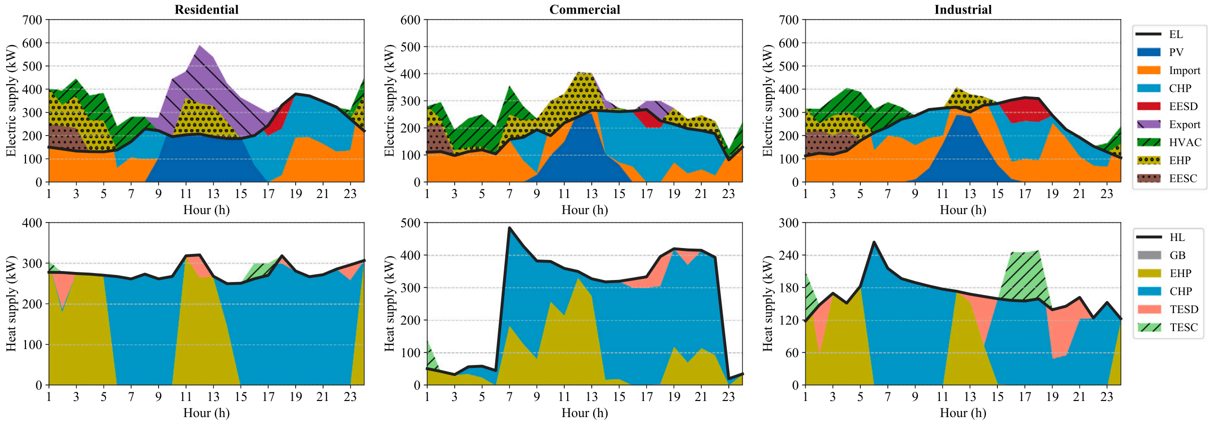


Fig. 9. Energy sources for the electric and heat power of three buildings in Fed-JPC. The two legends' black lines and different colored areas represent the same information as Fig. 7.

use of CHP and GB. As a result, the residential building accordingly reduces the P2P energy transactions among the building community and instead decides to use more PV resources for supplying HL via EHP.

6.4. Energy flows in community

Given the above hourly analysis of three buildings' energy management strategies, this section aims at analyzing and comparing the daily aggregated energy (electricity, gas, and heat) flows within the building community for different MARL algorithms, as depicted in Fig. 10.

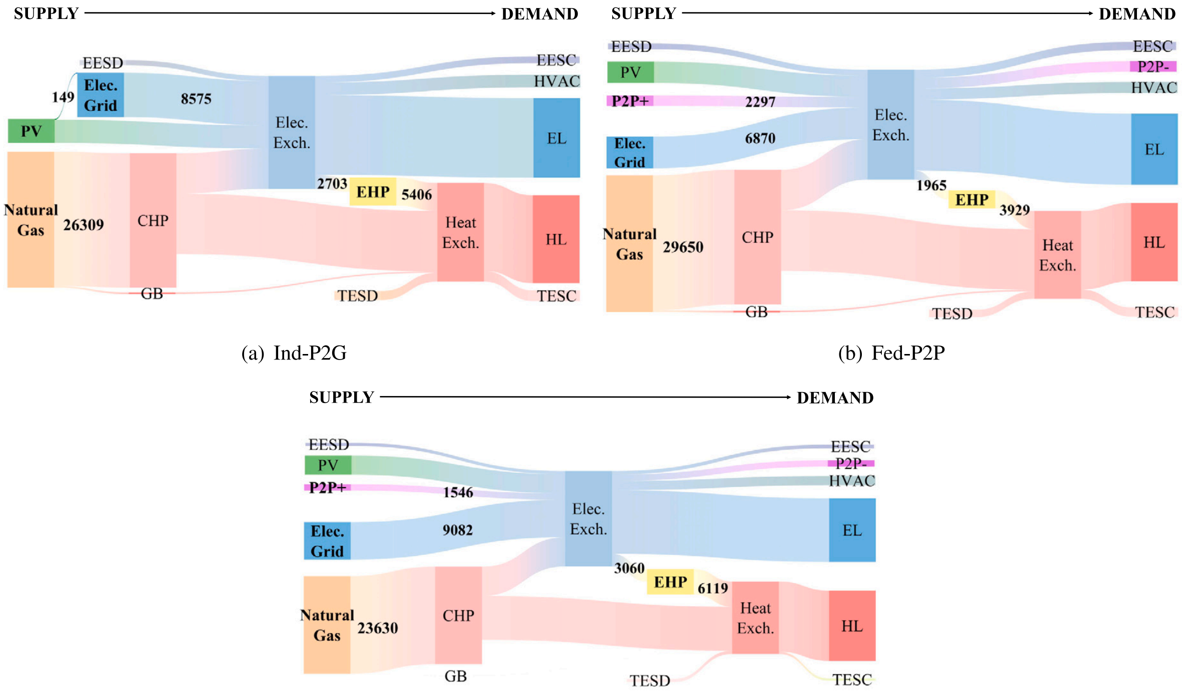
Ind-P2G. Buildings under Ind-P2G operate in an uncoordinated scenario that individually manages their MES controllable components to reach the system demand-supply balance with the objective of energy cost minimization. It can be observed from Fig. 10(a) that the major energy resources are coming from the electricity grid (8575 kWh) and natural gas (26,309 kWh), while PV resources can be used only by the individual buildings without a P2P energy trading mechanism. As a result, the free and clean PV resources may not be fully exploited, which leads to a 149 kWh PV energy surplus sold back to the electricity grid at the unattractive (cheap) grid sell price. When compared to the large amount of energy procured from the electricity grid at the expensive grid buy price, such a phenomenon will result in some economic loss.

Fed-P2P. When P2P energy trading is allowed in the building community, residential buildings with abundant PV energy resources are incentivized to sell PV directly to the commercial and industrial buildings at a high demand deficit, which leads to a 2,297 kWh local energy trading quantity, as illustrated in Fig. 10(b). It is also noted that PV resources are completely snuffed out and there is no energy surplus being sold back to the electricity grid. Fig. 10(b) also shows that the energy resources from the electricity grid (6,870 kWh) are relatively lower than those (8,575 kWh) under Ind-P2G. This reduced procurement is also the perfect proof of the buildings' declining dependence on the external electricity grid in the P2P energy trading mechanism. As such, in order to meet the community's overall demand, more natural gas (29,650 kWh) is purchased on the supply side.

Fed-JPC. In addition to the peer-to-peer energy trading mechanism, Fed-JPC buildings consider the ETS, which allows buildings to trade their carbon allowance in the carbon market. As a result, in order to reduce carbon costs, buildings are incentivized to reduce natural gas energy supplies (23,630 kWh) and increase electricity grid energy supplies (9082 kWh), as illustrated in Fig. 10(c). It is noted that the P2P energy trading behavior among three buildings still exists under Fed-JPC. The local energy trading quantity (1546 kWh) is relatively reduced compared with that under Fed-P2P (2297 kWh). This is mainly driven by the increasing dependence on EHP (6119 kWh), characterized by generating heat energy without carbon emissions.

6.5. Building carbon emission and trading activity

In this section, we show the carbon emission of three buildings for different MARL algorithms in Figs. 11–13. Fig. 13 depicts the free allocation of carbon allowances as well as carbon trading in the carbon market for Fed-JPC.



(a) Ind-P2G

(b) Fed-P2P

(c) Fed-JPC

Fig. 10. Daily aggregated energy flows from supply side (left) to demand side (right) within the building community for (a) Ind-P2G, (b) Fed-P2P, and (c) Fed-JPC. The number on the flow indicates the energy quantity in kWh over 24 h.

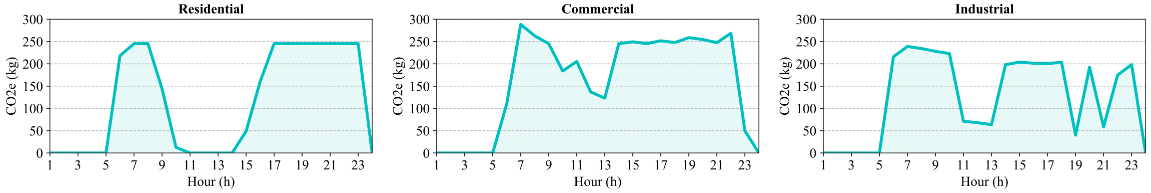


Fig. 11. Carbon emissions of three buildings in Ind-P2G.

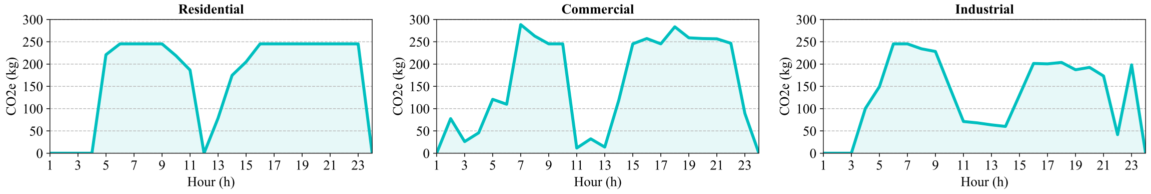


Fig. 12. Carbon emissions of three buildings in Fed-P2P.

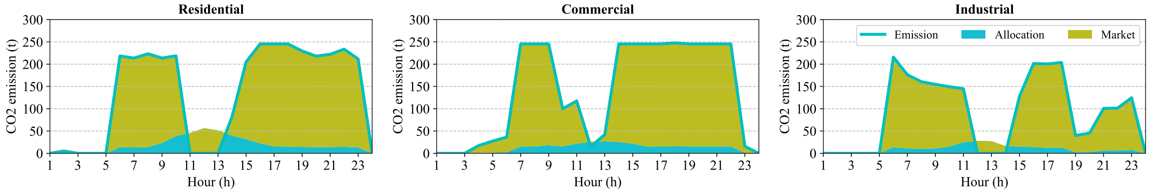
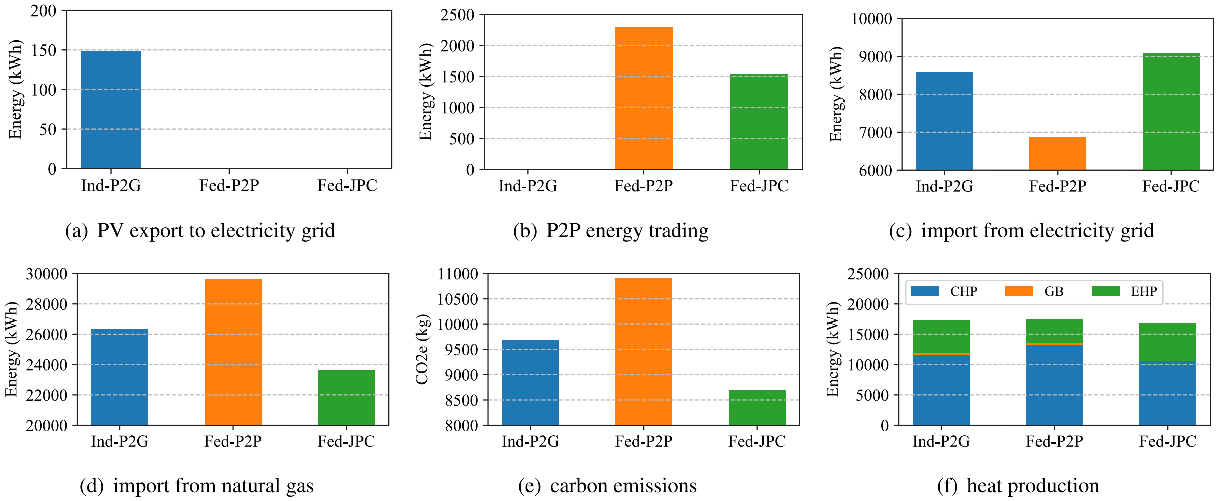


Fig. 13. Carbon emission, allowance allocations, and trading activities of three buildings in Fed-JPC.

Ind-P2G and fed-P2P. Since buildings under Ind-P2G and Fed-P2P do not participate in ETS, there is no carbon allowance free allocation and carbon trading activity in the carbon market. We only plot the carbon emissions of three buildings, as depicted in Figs. 11 and 12. In general, there are not many differences between Ind-P2G and Fed-P2P except the carbon emissions of residential and commercial buildings. Because PV is supplied to the residential building's own EHP for HL, the residential building's carbon emissions are zero at midday under



(a) PV export to electricity grid

(b) P2P energy trading

(c) import from electricity grid

(d) import from natural gas

(e) carbon emissions

(f) heat production

Fig. 14. Comparison of daily results in the community for different algorithms: (a) PV surplus sold to electricity grid; (b) P2P energy trading quantity; (c) energy procurement from electricity grid; (d) energy procurement from natural gas; (e) carbon emission; and (f) energy supplies for heat demand.

Fed-P2G (Figs. 11). However, in Fed-P2P, a residential building's PV is not supplied to its own EHP but rather to the commercial HL via EHP, resulting in a significant reduction in carbon emissions for the commercial building (Fig. 12).

Fed-JPC. Different from Ind-P2G and Fed-P2P, buildings under Fed-JPC are more concerned about the environment. As a result, as shown in Fig. 13, the carbon emissions of all three buildings are significantly reduced. Furthermore, under the CAT approach, a certain level of free carbon allowance is also allocated to each building to cover their carbon emissions, and the remaining quantity is still purchased from the carbon market. It is noted that residential and industrial buildings with PV resources and zero emissions in the midday can sell free carbon allowances back to the carbon market for the extra economic benefit.

6.6. Discussion

We now discuss the key physical insights observed from the experiment results. Specifically, we try to investigate how the building community benefits from the proposed Fed-JPC algorithm through six key results, which are illustrated in Fig. 14.

Overall, we can summarize the following five key physical insights through the analysis in Fig. 14:

(1) Fig. 14(a) shows that in Ind-P2G, there is still a 150 kWh daily PV surplus sold back to the electricity grid, whereas the introduction of P2P energy trading in both Fed-P2P and Fed-JPC reduces this value to zero. That means such an amount of PV surplus is traded among the buildings for the community's local demand-supply balance.

(2) It can be observed from Fig. 14(b) that a large number of local energy transactions exist for Fed-P2P and Fed-JPC. However, the introduction of the CAT approach in Fed-JPC accordingly reduces the volumes of local energy transactions, since buildings with abundant PV resources are prioritized to supply their own heat load via EHP for the purpose of carbon emission reduction rather than sell to other buildings in P2P energy trading. Such a phenomenon is also discussed in Section 6.3.

(3) Despite the introduction of P2P energy trading in Fed-JPC, the building community still requires a significant amount of energy from the electricity grid, as shown in Fig. 14(c). This is because Fed-JPC buildings are more concerned about carbon emissions, which will increase the use of zero-emission electricity (Fig. 14(c)) and, as a result, reduce the use of natural gas with carbon emissions (Fig. 14(d)). Fed-JPC emits the least amount of CO2 among the three algorithms, as shown in Fig. 14(e).

(4) Fed-P2P, on the other hand, is only concerned with P2P energy trading, which will increase the use of natural gas (Fig. 14(d)) and, consequently, carbon emissions (Fig. 14(e)). This is because the residential building with a PV surplus gives priority to selling it to the commercial and industrial buildings rather than supplying its own EHP, which increases the usage of natural gas to meet the remaining heat load requirement.

(5) Fig. 14(f) depicts the energy supplies for three different components of the community heat load. It can be found that the introduction of the CAT approach in Fed-JPC can mitigate carbon emissions by increasing/reducing the usage of EHP/GB, leading to an environmentally friendly transition.

Given the above analysis of realistic insights, we now discuss the potential advantages that each party can enjoy from the proposed Fed-JPC. (1) Regulators and system operators can use Fed-JPC to assess the value of P2P energy trading in a deregulated power system, informing a detailed cost–benefit analysis. (2) Policymakers may use Fed-JPC to provide a localized solution for small-size entities that want to participate in ETS and contribute to the low-carbon transition. (3) The fruitful results of MARL algorithms shown in this paper promote the investigation into the deployment of Fed-JPC on MARL in the real world in the future.

7. Conclusions and future work

This paper has proposed a novel MARL method to address the joint P2P energy and carbon trading (JPC) problem of three different types of MES buildings in a local community. The examined MES buildings feature various demand and renewable characteristics and complex MES operation models that are categorized into residential, commercial, and industrial areas. In order to solve this JPC problem, we first formulate it into a POMG wherein each building is regarded as an agent and the JPC problem is the environment. Then, we propose a novel MARL algorithm named Fed-JPC to solve this POMG. Specifically, the proposed Fed-JPC (1) constructs the centralized critic by abstracting the other agents' local observations and actions via the P2P market net trading quantity and carbon emission, thereby stabilizing the training performance by capturing the system dynamics; and (2) employs the federated learning (FL) technique to achieve distributed control in a privacy-protected manner. Experiment results involving a real-world MES scenario demonstrate the effectiveness of energy management and coordination among three buildings and the superior performance of the proposed Fed-JPC in reducing the total cost of energy and emissions with respect to the benchmark Ind-P2G and Fed-P2P algorithms. Finally, the proposed Fed-JPC achieves a near-to-optimal performance as the theoretical centralized optimization approach and outperforms the conventional consensus-based ADMM approach in terms of both lower total (energy and environmental) cost and computational performance.

However, this paper only focuses on the joint P2P energy and carbon trading problem of three smart buildings. Future work will explore the coordination of a large-scale building community and investigate the scalability of the proposed federated deep reinforcement learning algorithm. Generally, there are two key challenges to scaling up the number of agents in MARL: (1) agent-agent interactions are critical in multi-agent systems while the number of interactions grows quadratically with the number of agents, causing the non-stationary issue and difficulty in policy-stabilization; and (2) the dimensions of concatenating all agents' local observations and actions will increase proportionally with the number of agents, causing the curse of dimensionality and making it impractical to train the neural networks. Future work will try to address the above two challenges and develop a scalable federated deep multi-agent reinforcement learning algorithm.

CRediT authorship contribution statement

Dawei Qiu: Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Juxing Xue:** Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Tingqi Zhang:** Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Jianhong Wang:** Methodology, Data curation, Formal analysis. **Mingyang Sun:** Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Project administration, Supervision, Funding acquisition.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the the National Natural Science Foundation of China under Grants 62103371, 52161135201, U20A20159, 62061130220.

References

[1] Programme UNE. 2021 Global status report for buildings and construction: towards a zero–emission, efficient and resilient buildings and construction sector. 2021, https://globalabc.org/index.php/resources/publications/2021-global-status-report-buildings-and-construction.

[2] Kylili A, Fokaides PA. European smart cities: The role of zero energy buildings. Sustain Cities Soc 2015;15:86–95.

[3] Mariano-Hernández D, Hernández-Callejo L, Zorita-Lamadrid A, Duque-Pérez O, García FS. A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. J Build Eng 2021;33:101692.

[4] Mancarella P. MES (multi-energy systems): An overview of concepts and evaluation models. Energy 2014;65:1–17.

[5] Qiu D, Ye Y, Papadaskalopoulos D. Exploring the effects of local energy markets on electricity retailers and customers. Electr Power Syst Res 2020;189:106761.

[6] Morstyn T, Farrell N, Darby SJ, McCulloch MD. Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants. Nat Energy 2018;3(2):94–101.

[7] Ellerman AD, Convery FJ, De Perthuis C. Pricing carbon: the european union emissions trading scheme. Cambridge University Press; 2010.

[8] Alam MR, St-Hilaire M, Kunz T. Peer-to-peer energy trading among smart homes. Appl Energy 2019;238:1434–43.

[9] Liu J, Yang H, Zhou Y. Peer-to-peer energy trading of net-zero energy communities with renewable energy systems integrating hydrogen vehicle storage. Appl Energy 2021;298:117206.

[10] Si F, Wang J, Han Y, Zhao Q, Han P, Li Y. Cost-efficient multi-energy management with flexible complementarity strategy for energy internet. Appl Energy 2018;231:803–15.

[11] Gan W, Yan M, Wen J, Yao W, Zhang J. A low-carbon planning method for joint regional-district multi-energy systems: From the perspective of privacy protection. Appl Energy 2022;311:118595.

[12] Lyu C, Jia Y, Xu Z. Fully decentralized peer-to-peer energy sharing framework for smart buildings with local battery system and aggregated electric vehicles. Appl Energy 2021;299:117243.

[13] Yan M, Shahidehpour M, Paaso A, Zhang L, Alabdulwahab A, Abusorrah A. Distribution network-constrained optimization of peer-to-peer transactive energy trading among multi-microgrids. IEEE Trans Smart Grid 2020;12(2):1033–47.

[14] Cui S, Wang Y-W, Shi Y, Xiao J-W. A new and fair peer-to-peer energy sharing framework for energy buildings. IEEE Trans Smart Grid 2020;11(5):3817–26.

[15] Mansouri SA, Ahmarinejad A, Nematbakhsh E, Javadi MS, Jordehi AR, Catalão JP. Energy hub design in the presence of P2G system considering the variable efficiencies of gas-fired converters. In: 2021 international conference on smart energy systems and technologies. IEEE; 2021, p. 1–6.

[16] Jordehi AR, Javadi MS, Shafie-khah M, Catalão JP. Information gap decision theory (IGDT)-based robust scheduling of combined cooling, heat and power energy hubs. Energy 2021;231:120918.

[17] Javadi MS, Anvari-Moghaddam A, Guerrero JM. Robust energy hub management using information gap decision theory. In: IECON 2017-43rd annual conference of the IEEE industrial electronics society. IEEE; 2017, p. 410–5.

[18] Shams MH, MansourLakouraj M, Shahabi M, Javadi MS, Catalão JP. Robust scenario-based approach for the optimal scheduling of energy hubs. In: 2021 IEEE madrid PowerTech. IEEE; 2021, p. 1–6.

[19] Jordehi AR, Javadi MS, Catalão JP. Day-ahead scheduling of energy hubs with parking lots for electric vehicles considering uncertainties. Energy 2021;229:120709.

[20] Javadi MS, Anvari-Moghaddam A, Guerrero JM, Nezhad AE, Lotfi M, Catalão JP. Optimal operation of an energy hub in the presence of uncertainties. In: Proc IEEE int conf environ elect eng IEEE ind commercial power syst eur. IEEE; 2019, p. 1–4.

[21] Gan W, Yan M, Yao W, Wen J. Peer to peer transactive energy for multiple energy hub with the penetration of high-level renewable energy. Appl Energy 2021;295:117027.

[22] Jing R, Xie MN, Wang FX, Chen LX. Fair P2P energy trading between residential and commercial multi-energy systems enabling integrated demand-side management. Appl Energy 2020;262:114551.

[23] Zhang H, Zhang S, Hu X, Cheng H, Gu Q, Du M. Parametric optimization-based peer-to-peer energy trading among commercial buildings considering multiple energy conversion. Appl Energy 2022;306:118040.

[24] Javadi MS, Nezhad AE, Jordehi AR, Gough M, Santos SF, Catalão JP. Transactive energy framework in multi-carrier energy hubs: A fully decentralized model. Energy 2022;238:121717.

[25] Hua W, Jiang J, Sun H, Wu J. A blockchain based peer-to-peer trading framework integrating energy and carbon markets. Appl Energy 2020;279:115539.

[26] Yan M, Shahidehpour M, Alabdulwahab A, Abusorrah A, Gurung N, Zheng H, et al. Blockchain for transacting energy and carbon allowance in networked microgrids. IEEE Trans Smart Grid 2021;12(6):4702–14.

[27] Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT Press; 2018.

[28] Jogunola O, Adebisi B, Ikpehai A, Popoola SI, Gui G, Gačanin H, et al. Consensus algorithms and deep reinforcement learning in energy market: A review. IEEE Internet Things J 2020;8(6):4211–27.

[29] Chen T, Su W. Local energy trading behavior modeling with deep reinforcement learning. IEEE Access 2018;6:62806–14.

[30] Chen T, Bu S. Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning. In: 2019 IEEE PES innovative smart grid technologies europe. IEEE; 2019, p. 1–5.

[31] Prasad A, Dusparic I. Multi-agent deep reinforcement learning for zero energy communities. In: 2019 IEEE PES innovative smart grid technologies europe. Bucharest, Romania: IEEE; 2019, p. 1–5.

[32] Xu Y, Yu L, Bi G, Zhang M, Shen C. Deep reinforcement learning and blockchain for peer-to-peer energy trading among microgrids. In: 2020 international conferences on internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData) and IEEE congress on cybermatics (Cybermatics). IEEE; 2020, p. 360–5.

[33] Qiu D, Wang J, Wang J, Strbac G. Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market. In: Proc. 30th int. jt. conf. artif. intell. 2021, p. 2913–20.

[34] Chen T, Bu S, Liu X, Kang J, Yu FR, Han Z. Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning. IEEE Trans Smart Grid 2021;13(1):715–27.

[35] Qiu D, Wang J, Dong Z, Wang Y, Strbac G. Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading. IEEE Trans Power Syst 2022.

[36] Pinto G, Kathirgamanathan A, Mangina E, Finn DP, Capozzoli A. Enhancing energy management in grid-interactive buildings: A comparison among cooperative and coordinated architectures. Appl Energy 2022;310:118497.

[37] 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 2022;311:118636.

[38] Li L, Fan Y, Tse M, Lin K-Y. A review of applications in federated learning. Comput Ind Eng 2020;149:106854.

[39] Qi J, Zhou Q, Lei L, Zheng K. Federated reinforcement learning: Techniques, applications, and open challenges. 2021, arXiv preprint arXiv:2108.11887.

[40] Qiu D, Ye Y, Papadaskalopoulos D, Strbac G. Scalable coordinated management of peer-to-peer energy trading: A multi-cluster deep reinforcement learning approach. Appl Energy 2021;292:116940.

[41] Hirst D, Keep M. Carbon Price Floor (CPF) and the price support mechanism. House Commons Libr Brief Pap 2018;20.

[42] Mason K, Grijalva S. A review of reinforcement learning for autonomous building energy management. Comput Electr Eng 2019;78:300–12.

[43] Javadi M, Lotfi M, Osório GJ, Ashraf A, Nezhad AE, Gough M, et al. A multi-objective model for home energy management system self-scheduling using the epsilon-constraint method. In: 2020 IEEE 14th international conference on compatibility, power electronics and power engineering, vol. 1. IEEE; 2020, p. 175–80.

[44] Javadi MS, Nezhad AE, Nardelli PH, Gough M, Lotfi M, Santos S, et al. Self-scheduling model for home energy management systems considering the end-users discomfort index within price-based demand response programs. Sustain Cities Soc 2021;68:102792.

[45] Javadi M, Nezhad AE, Firouzi K, Besanjideh F, Gough M, Lotfi M, et al. Optimal operation of home energy management systems in the presence of the inverter-based heating, ventilation and air conditioning system. In: Proc IEEE int conf environ elect eng IEEE ind commercial power syst eur. IEEE; 2020, p. 1–6.

[46] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: Proc. 4th int. conf. learn. represent. San Juan, Puerto Rico; 2016, p. 1–14.

[47] US Department of Energy's Programs, Offices, and National Laboratories. Open energy data initiative (OEDI). 2021, https://openei.org/datasets/dataset.

[48] University RA. FIWARE lab. 2021, https://data.lab.fiware.org/organization/rwth-aachen-university.

[49] Rainfall and temperature forecast and observations - verification 2016-05 to 2017-04. 2017, URL https://data.gov.au/data/dataset/0bfba2bc-2042-4ae3-91a1-17e4414e4391.

[50] Huang W, Zhang N, Yang J, Wang Y, Kang C. Optimal configuration planning of multi-energy systems considering distributed renewable energy. IEEE Trans Smart Grid 2019;10(2):1452–64.

[51] Qiu D, Dong Z, Zhang X, Wang Y, Strbac G. Safe reinforcement learning for real-time automatic control in a smart energy-hub. Appl Energy 2022;309:18403.

[52] UK Department for Business, Energy & Industrial Strategy. Greenhouse gas reporting: conversion factors 2021. 2021, https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2021.

[53] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. 3rd int. conf. learn. represent. San Diego, USA; 2015, p. 1–15.