Coordination for Multienergy Microgrids Using Multiagent Reinforcement Learning

Dawei Qiu, *Member, IEEE*, Tianyi Chen, *Student Member, IEEE*, Goran Strbac, *Member, IEEE*, and Shengrong Bu, *Member, IEEE*

***Abstract*—Multienergy microgrids (MEMGs) have significant potential to offer high energy utilization efficiency and system flexibility. The coordination of these MEMGs poses challenges due to the various system dynamics and uncertainties and the need to preserve privacy. This article proposes a double auction (DA)-market-based coordination framework. As such, MEMGs can not only schedule their own energy components but also trade energy with others in the DA market. After that, we formulate this problem as Markov games and propose a multiagent reinforcement learning method by making use of the DA market public information to enhance the stability with privacy perseverance. Case studies involving a real-world scenario validate the superior performance of the proposed method in reducing both the energy costs and the carbon emissions.**

***Index Terms*—Carbon emissions, energy coordination, multienergy microgrid (MEMG), multiagent reinforcement learning (MARL).**

NOMENCLATURE

Indices and Sets

$t \in T$	Index and set of time steps (hours).
$i \in \mathcal{I}_{\mathrm{RG}}$	Index and set of residential MEMGs.
$i \in \mathcal{I}_{\mathrm{CG}}$	Index and set of commercial MEMGs.
$i \in \mathcal{I}_{\mathrm{IG}}$	Index and set of industrial MEMGs.

Parameters

$\lambda^b, \lambda^s, \lambda^g$	Grid electricity buy, electricity sell, and gas prices ($/kWh).
λ^c	Carbon price ($/kg).
P^l, Q^l	Electric and heat load (kW).
$P^{\mathrm{pv}}, P^{\mathrm{wd}}$	PV and wind power generation (kW).
$\overline{P}^{\mathrm{ees}}, \overline{Q}^{\mathrm{tes}}$	Power capacity of EES and TES (kW).
$\underline{E}^{\mathrm{ees}}, \overline{E}^{\mathrm{ees}}$	Energy capacity of EES (kWh).
$\eta^{\mathrm{eesc}}, \eta^{\mathrm{eesd}}$	Charging and discharging efficiency of EES.
$\overline{G}^{\mathrm{chp}}, \overline{G}^{\mathrm{fc}}$	Gas power capacity of CHP and FC (kW).
$\eta^{\mathrm{chpe}}, \eta^{\mathrm{chpq}}$	Conversion efficiency from gas to electricity and heat of CHP.
$\overline{G}^{\mathrm{gb}}$	Gas power capacity of GB (kW).
η^{gb}	Conversion efficiency from gas to heat of GB.
$\overline{P}^{\mathrm{ehp}}$	Electric power capacity of EHP (kW).
η^{ehp}	Conversion efficiency from electricity to heat of EHP.

Variables

$P^{\mathrm{eesc}}, P^{\mathrm{eesd}}$	Charging and discharging power of EES (kW).
$Q^{\mathrm{tesc}}, Q^{\mathrm{tesd}}$	Charging and discharging power of TES (kW).
$E^{\mathrm{ees}}, E^{\mathrm{tes}}$	Energy content in EES and TES (kWh).
$G^{\mathrm{chp}}, G^{\mathrm{fc}}$	Gas power input of CHP and FC (kW).
$P^{\mathrm{chp}}, Q^{\mathrm{chp}}$	Electric and heat power output of CHP (kW).
$P^{\mathrm{fc}}, Q^{\mathrm{fc}}$	Electric and heat power output of FC (kW).
G^{gb}	Gas power input of GB (kW).
$Q^{\mathrm{gb}}, Q^{\mathrm{ehp}}$	Heat power output of GB and EHP (kW).
P^{ehp}	Electric power input of EHP (kW).
E^c	Carbon emissions from natural gas (kg).

Manuscript received 24 February 2022; accepted 7 April 2022. Date of publication 19 April 2022; date of current version 22 March 2023. This work was supported in part by the EU Horizon 2020 Research and Innovation Program under project TradeRES under Grant 864276, in part by the U.K. Engineering and Physical Sciences Research Council under project "Integrated Development of Low-Carbon Energy Systems: A Whole-System Paradigm for Creating a National Strategy" under Grant EP/R045518/1, and in part by start-up funds provided by Brock University. Paper no. TII-22-0847. *(Corresponding author: Tianyi Chen.)*

Dawei Qiu and Goran Strbac are with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: d.qiu15@imperial.ac.uk; g.strbac@imperial.ac.uk).

Tianyi Chen is with the James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K. (e-mail: t.chen.1@research.gla.ac.uk).

Shengrong Bu is with the Department of Engineering, Brock University, St. Catharines, ON L2S 3A1, Canada (e-mail: sbu@brocku.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2022.3168319.

Digital Object Identifier 10.1109/TII.2022.3168319

I. INTRODUCTION

A. Background and Motivation

POWER systems are undergoing a significant transition from fossil fuel resources to the decarbonization of *renewable energy resource* (RES), promising to address the environmental concerns [1]. However, the less controllable and predictable RES introduces new challenges to power system planning and operation [2]. In this respect, there has been a significant increase in developing *multienergy systems* (MESs) that interact electricity, gas, and heat with each other, constituting a significant opportunity to provide the flexibility of shifting across multiple energy vectors and resulting in a cost-effective and reliable system [3]. Currently, an increasing attention has been made to study the MES inside a microgrid, forming the *multienergy microgrids* (MEMGs) [4], [5]. An MEMG is composed of various energy loads, generators, storages, and converters under the microgrid

concept. Currently, the benefits of using the MEMG have been discussed in many studies [5]. Instead of independently scheduling each energy vector, the integrated manner is more efficient to deal with the complementary and synergistic effects of the MES, therefore boosting the operation efficiency of the MEMG.

Gas and electricity are the two main input energy sources for MEMGs. The gas retail market is normally indifferent to MEMGs, allowing them to buy gas but not sell it back [6]. The electricity retail market under the deregulation is more active and flexible, where MEMGs with the RES can sell electricity back to the grid at feed-in tariff (FiT) [7]. However, under scenarios where MEMGs need to import energy from the grid, the higher rated time-of-use (ToU) prices, compared with the lower FiT issued by the same utility company, can present a dilemma for MEMGs' net import decision making [8]. Furthermore, when MEMGs participate in the traditional market, they act independently to manage their supply–demand balance. This is, however, not optimal as the lack of coordination with others leaves untapped the full potential of energy flexibility for achieving overall system supply–demand balance [9]. To this end, an efficient coordinated management of local MEMGs is urgent to maximize the economic benefit and the system flexibility.

B. Literature Review

So far, the existing literature on the coordinated management of multiple MEMGs can be classified into two categories. The first one focuses on the design of a *centralized* framework that employs a central operator to manage all the local resources [10]. Although such a framework provides a theoretical solution for social welfare maximization, it exhibits various drawbacks in practice. Specifically, the central operator needs to acquire mathematical models and collect all the technical parameters of local resources, thereby raising privacy concerns. The second one focuses on the design of a *decentralized* framework that allows the MEMGs to manage their own resources independently with limited information exchange, preserving their privacy. Currently, alternating direction method of multipliers [11], [12], Lagrangian relaxation [13], [14], consensus algorithm [15], and bilateral contract [16] are popular methods in the decentralized framework for solving the coordination management of multiple MEMGs. However, the optimality of solutions is not guaranteed under such a decentralized framework without a central coordinator [10].

To this end, a double auction (DA) market [17] is a kind of framework that takes advantage of both the *centralized* and *decentralized* frameworks, which is potential to be considered to form local coordination among a group of MEMGs. More specifically, an auctioneer, as a third-party coordinator, is responsible for clearing the market to ensure the market efficiency, which is close to optimal in a *centralized* framework [18]. On the other hand, MEMGs can manage their resources independently and submit only the bidding information (i.e., price–quantity bids) to the auctioneer. As such, the privacy can be preserved that is similar to the *decentralized* framework. However, MEMGs in the DA market are faced with a complex quotation decision process. Thus, an appropriate trading strategy is challenging to select in such a complicated market environment. Zero intelligence (ZI) is a fundamental trading strategy adopted by traders in the DA market [19]. Specifically, ZI selects the price bid uniformly at random values between FiT and ToU and runs a day-ahead self-optimization problem for quantity bid submitted to the DA market. However, the randomized price bid does not capture the market dynamics [20]. Furthermore, preoptimized quantity bid requires the complete MEMG mathematical models, technical parameters, and accurate forecasting information of uncertainties, which are generally impractical in real-world applications [21].

In view of the above drawbacks in the ZI strategy, *reinforcement learning* (RL) [22] is a model-free and data-driven control method to study the sequential decision-making problem, where the agents within MEMGs gradually learn the optimal trading strategies by utilizing experiences acquired from their repeated interactions with the environment (MEMGs and DA market), without a *prior* knowledge of MEMGs. In addition, RL as an online learning method can make use of increasing data acquired from the environment to learn the optimal control strategies and to cope with the uncertainties that are encapsulated in the data [23].

Previous works have successfully applied various RL methods to energy management problems in power systems, as reviewed in [24]. The majority of them, however, only consider the energy management problem of a single entity, e.g., a smart energy hub [25] and a residential multienergy home [23], and employ *single-agent reinforcement learning* methods. On the other hand, the research efforts on the application of *multiagent reinforcement learning* (MARL) on power systems are still sparse, particularly for our studied MEMG coordination management problem. The most straightforward approach to solve a multiagent problem is independent reinforcement learning (IRL) that each agent trains its independent control policy depending on the local information. Independent deep Q-network [26] and independent deep deterministic policy gradient (IDDPG) [27] have been applied to the energy management problems of the multiple MGs, where each agent treats others as part of the environment and learns its own policy without considering others' policies. However, directly applying IRL methods to a multiagent setting is problematic, since the environment appears nonstationary from the view of every agent [28]. To overcome this issue, multiagent deep deterministic policy gradient (MADDPG), an extension of IDDPG to a multiagent setting, has been proposed to address the energy trading problem among the microgrids [29]. Each agent in MADDPG trains a centralized Q-value function (critic) with access to all agents' observations and actions to stabilize the training performance. During the execution, the decentralized actor of each agent makes decisions based on its local observation value. However, MADDPG mainly suffers from the following: 1) *privacy concern:* knowing the local observations and actions of all the other agents and 2) *stability concern:* the learned Q-values may be overestimated, which can lead to the suboptimal polices [30].

C. Article Contributions

To address the limitations of privacy and instability issues discussed above, this article proposes a novel MARL method

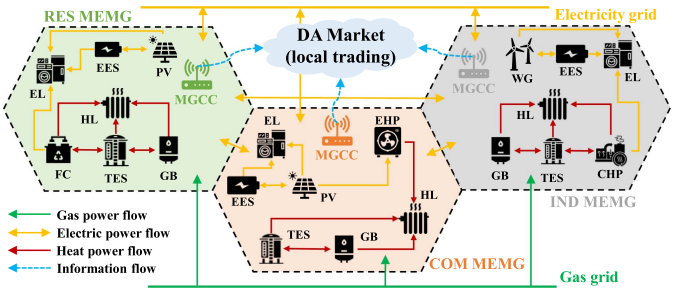

Fig. 1. Energy coordination framework and the MES of considered MEMGs.

for multiple MEMGs to provide autonomous control and trading policies for local energy coordination in a DA market. Specifically, a list of contributions can be provided as follows.

1) The flexibility due to the local electricity trading among different MEMGs and the coupled energy conversions in each MEMG is explored. The examined problem is complex because of various system dynamics and uncertainties. A DA-market-based coordination framework has been proposed to obtain good performance with privacy preservation. To the best of our knowledge, this is the first work to adopt the DA market mechanism to a local energy community with multiple MEMGs.

2) A novel DA-MATD3 method is proposed, which inherits the ability of the multi-agent twin delayed deep deterministic policy gradient (MATD3) to perform well in a multiagent environment with various system dynamics and uncertainties and addresses privacy concerns using a DA market framework. Specifically, the DA-MATD3 method integrates the key information of the DA market into the state-of-the-art MATD3 algorithm by connecting the critic networks of the agents with the DA market order books. To the best of our knowledge, this is the first work to integrate the DA market information into the MATD3 algorithm.

D. Article Organization

The rest of this article is organized as follows. Section II formulates the examined coordination problem of multiple MEMGs in a DA market. Section III proposes the DA-MATD3 method. Section IV presents the case studies to evaluate the effectiveness of the proposed method. Finally, Section V concludes this article.

II. Coordination of MEMGs in the DA Market

A. Problem Setting

We focus on a local energy community consisting of a group of MEMGs, as depicted in Fig. 1. In detail, the set of components of the proposed MEMGs includes: 1) two types of consumption loads: electric load (EL) and heat load (HL); 2) two types of RES generators: solar photovoltaic (PV) and wind generator (WG); 3) two types of storage units: electric energy storage (EES), and thermal energy storage (TES); and 4) four types of energy converters: combined heat and power (CHP) engine, fuel cell (FC), electric heat pump (EHP), and gas boiler (GB). The MEMGs are categorized into three groups: 1) residential MEMGs with the energy portfolio of EL, HL, PV, EES, TES, FC, and GB; 2) commercial MEMGs with the energy portfolio of EL, HL, PV, EES, TES, EHP, and GB; and 3) industrial MEMGs with the energy portfolio of EL, HL, WG, EES, TES, CHP, and GB.

In order to incentivize MEMGs to cooperatively participate in local trading, we introduce a DA market driven by its high trading efficiency [17]. As shown in Fig. 1, the options of each MEMG to supply its consumption loads are diverse. First, MEMGs can manage their own installed energy resources to supply EL and HL. Second, MEMGs can trade their electricity with each other in the DA market. Third, MEMGs are allowed to buy/sell their unbalanced electricity with the utility company at the grid buy/sell prices. Finally, MEMGs can purchase natural gas from the gas grid. The decision-making problem is processed for each hour across a daily horizon, with the objective of minimizing energy cost and carbon emission. At each hour, each *microgrid central controller* (MGCC) [31] equipped in an MEMG can manage its energy schedules and trading decisions based on: 1) the grid information of energy and carbon price signals; 2) the local information of its consumption loads, renewable generations, and the status of controllable components; and 3) the community information of DA market trading prices and quantities.

B. Multienergy Microgrids

This section aims at providing the detailed mathematical models of four energy converters (CHP, FC, EHP, and GB) and two storage energy units (EES and TES).

1) Energy Converters: CHP as a single-input multioutput converter is characterized by its high energy efficiency compared to independent electricity and heat sources, of which the coupled heat and electricity generation can be modeled as

$$P_t^{\text{chp}} = \eta^{\text{chpe}} G_t^{\text{chp}} \tag{1}$$

$$Q_t^{\text{chp}} = \eta^{\text{chpq}} G_t^{\text{chp}} \tag{2}$$

$$0 \le G_t^{\text{chp}} \le \overline{G}^{\text{chp}} \tag{3}$$

where constraints (1) and (2) indicate the efficiency of CHP to convert natural gas into electric and heat power, respectively. The gas input is limited by its power capacity expressed in (3). Like CHP engines, an FC is also a single-input multioutput converter, characterized by its higher combined efficiency and lower emissions. Given the high thermal efficiency and low operating temperature, the FC is more suitable for individual residents with high heat demands. The mathematical model of the FC is similar to the CHP model (1)–(3).

Apart from CHP and the FC, the studied MEMGs also include the energy converters of the EHP and the GB. The EHP produces heat energy by consuming electricity, as presented in (4). The GB is a vessel converting natural gas into heat energy. The generation of heat from natural gas via the GB is given in (6). The power inputs of the EHP and the GB are limited by their individual capacity expressed in (5) and (7), respectively

$$Q_t^{\text{ehp}} = \eta^{\text{ehp}} P_t^{\text{ehp}} \tag{4}$$

$$0 \le P_t^{\text{ehp}} \le \overline{P}^{\text{ehp}} \tag{5}$$

$$Q_t^{\text{gb}} = \eta^{\text{gb}} G_t^{\text{gb}} \tag{6}$$

$$0 \le G_t^{\text{gb}} \le \overline{G}^{\text{gb}}. \tag{7}$$

2) Energy Storage Units: The energy storage units with the high flexibility are characterized by their redistribution ability of off-peak and peak loads and the ability to absorb free RES for the future usage when energy prices are at the peak. The mathematical models of an EES unit can be formulated as

$$E_{t+1}^{\text{ees}} = E_t^{\text{ees}} + P_t^{\text{eesc}} \Delta t \eta^{\text{eesc}} + P_t^{\text{eesd}} \Delta t / \eta^{\text{eesd}} \tag{8}$$

$$\underline{E}^{\text{ees}} \le E_t^{\text{ees}} \le \overline{E}^{\text{ees}} \tag{9}$$

$$0 \le P_t^{\text{eesc}} \le \overline{P}^{\text{ees}} V_t^{\text{ees}} \tag{10}$$

$$\overline{P}^{\text{ees}} (V_t^{\text{ees}} - 1) \le P_t^{\text{eesd}} \le 0 \tag{11}$$

where equality (8) corresponds to the storage dynamic transition of battery energy content, taking into account the charging and discharging energy losses. Constraint (9) expresses the lower and upper bounds of battery energy content. The following constraints (10) and (11) ensure that charging and discharging powers P_t^{eesc} and P_t^{eesd} are under their power capacity $\overline{P}^{\text{ees}}$ and operate mutually exclusive by introducing a binary variable $V_t^{\text{ees}} \in \{0, 1\}$. Then, the power rate $Q_t^{\text{tesc}}, Q_t^{\text{tesd}}$ and battery energy content E_t^{tes} of the TES unit can be derived similarly to the EES model (8)–(11).

C. DA Market

The DA market matches multiple buyers (MEMGs with energy deficit) and sellers (MEMGs with energy surplus) who are interested in local trading and is deemed as a highly efficient mechanism [17]. It is widely used in the trading of a variety of commodities, including equities and electricity. In this article, we apply the DA market to the local electricity trading, while the heat energy cannot be traded in the community. In general, a DA market lasts for a fixed period of time, known as the *auction period* (1 h). It allows traders to submit their bids/offers at the beginning of each auction period; then, the *auctioneer* (DA market operator) clears the market and publishes the public market outcomes (trading prices and quantities) at the end of each auction period. More specifically, a DA market comprises the following:

1) a set of *buyers* $\mathcal{B}$, where each buyer $b \in \mathcal{B}$ defines its trading price p_b and quantity q_b, which means that the buyer b would like to buy q_b amount of energy at price p_b;
2) a set of *sellers* $\mathcal{S}$, where each seller $s \in \mathcal{S}$ defines its trading price p_s and quantity q_s, which means that the seller s would like to sell q_s amount of energy at price p_s; and
3) a public *order book* managed by an auctioneer, where all the accepted bids and offers are listed. Bids submitted by buyers are sorted by decreasing the submitted buy prices and queue in the buy order book $k^b(b, p_b, q_b)$, while offers submitted by sellers are sorted by increasing the submitted sell prices and queue in the sell order book $k^s(s, p_s, q_s)$.

Algorithm 1: DA Market Clearing Algorithm.

1: Collect price–quantity bids/offers at auction period t
2: Allocate order books $k_t^b(b, p_{b,t}, q_{b,t})$ and $k_t^s(s, p_{s,t}, q_{s,t})$ at auction period t
3: Initialize $b = s = 1$
4: **while** $p_{b,t} \ge p_{s,t}$ **do**
5: match the trading energy: $q_t^l = \min(q_{b,t}, q_{s,t})$
6: calculate the trading price: $p_t^l = (p_{b,t} + p_{s,t})/2$
7: update buy order book $q_{b,t} \leftarrow q_{b,t} - q_t^l$
8: **if** $q_{b,t} = 0$ **then**
9: $b \leftarrow b + 1$
10: update sell order book $q_{s,t} \leftarrow q_{s,t} - q_t^l$
11: **if** $q_{s,t} = 0$ **then**
12: $s \leftarrow s + 1$
13: **break if**
14: $b >$ length of k_t^b or $s >$ length of k_t^s
15: **end while**
16: Balance unmatched quantity at FiT (λ_t^s) and ToU (λ_t^b) prices

The pseudocode of the DA market clearing process is given in Algorithm 1. Once an auction period begins, traders submit their order information with a trading price and a corresponding energy quantity to the market, collected by the auctioneer (step 1). All the submitted orders are allocated in the order book (step 2). The clearing process iterates down the order books and attempts to match each buy order with sell order (steps 3–12) until the buy price is less than the sell price or no unmatched order exists anymore (steps 13 and 14). Specifically, when two orders get matched, the auctioneer calculates the trading price between the matched buy price and sell price, using the traditional mid-pricing method [17] (step 6), while the trading quantity is equal to the lower value between the two matched orders (step 5). Owing to the sorting principle and the clearing algorithm, the clearing results promise the social welfare maximization [17]. Finally, at the end of the auction period, the remaining quantity of energy and the unmatched orders are balanced with the utility company at the grid electricity prices. It should be noted that the submitted prices of all the traders are bounded between the grid sell (FiT) and sell (ToU) prices to guarantee the economic benefits in the DA market instead of directly trading with the utility company [21].

D. Energy Coordination as the Markov Decision Process

The above-introduced DA market can be formulated as a multiagent coordination problem in the form of a finite *partially observable Markov decision process* (POMDP) [22] with discrete time steps. The POMDP is, then, defined with a set of state $\mathcal{S}$ describing the global state of environment $\mathcal{E}$ (DA market), 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 environment stochasticity representing uncertain parameters. The time interval between two consecutive time steps is one auction period ($\Delta t = 1$ 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 executes this $a_{i,t}$ to the environment $\mathcal{E}$. 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 maximizing the cumulative discounted reward $R_i = \mathbb{E}_{s\sim\mathcal{T},a_i\sim\pi_i}[\sum_{t=0}^{T}\gamma^t r_{i,t}]$, where $\gamma \in [0,1)$ is the discount factor and T is the daily horizon of 24 h. In detail, the components of the POMDP for the proposed coordination problem are defined as follows.

1) Agents: An agent is a computation entity within each MGCC of the MEMG, who can directly manage the controllable components in each MEMG and the trading strategies in the DA market.

2) Environment: The environment includes MEMGs defined in Section II-B, and the DA market defined in Section II-C.

3) Observation: Each MGCC agent i at time step t observes its local observation $o_{i,t}$ that varies for different MEMG categories and can be defined as

$$o_{i,t} = \begin{cases} [\lambda_t, L_{i,t}, P^{\text{pv}}_{i,t}, E^{\text{es}}_{i,t}] & \forall i \in \mathcal{I}_{\text{RG}}, \forall t \in T \\ [\lambda_t, L_{i,t}, P^{\text{pv}}_{i,t}, E^{\text{es}}_{i,t}] & \forall i \in \mathcal{I}_{\text{CG}}, \forall t \in T \\ [\lambda_t, L_{i,t}, P^{\text{wg}}_{i,t}, E^{\text{es}}_{i,t}] & \forall i \in \mathcal{I}_{\text{IG}}, \forall t \in T \end{cases} \tag{12}$$

where the observation $o_{i,t}$ consists of two parts: 1) the exogenous state unaffected by the action includes the sensor data of price signals $\lambda_t = [\lambda^b_t, \lambda^s_t, \lambda^g_t, \lambda^c_t]$ representing the grid electricity buy and sell prices, the gas price, and the carbon price, as well as the measured data of consummation loads $L_{i,t} = [P^l_{i,t}, Q^l_{i,t}]$ representing EL and HL, the renewable generation of PV $P^{\text{pv}}_{i,t}$, and WG $P^{\text{wg}}_{i,t}$; and 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 and TES $E^{\text{es}}_{i,t} = [E^{\text{ees}}_{i,t}, E^{\text{tes}}_{i,t}]$.

4) Action: Each MGCC agent i at time step t controls its action $a_{i,t}$ that varies for different MEMG categories and can be defined as

$$a_{i,t} = \begin{cases} [a^p_{i,t}, a^{\text{ees}}_{i,t}, a^{\text{tes}}_{i,t}, a^{\text{fc}}_{i,t}, a^{\text{gb}}_{i,t}] & \forall i \in \mathcal{I}_{\text{RG}}, \forall t \in T \\ [a^p_{i,t}, a^{\text{ees}}_{i,t}, a^{\text{tes}}_{i,t}, a^{\text{ehp}}_{i,t}, a^{\text{gb}}_{i,t}] & \forall i \in \mathcal{I}_{\text{CG}}, \forall t \in T \\ [a^p_{i,t}, a^{\text{ees}}_{i,t}, a^{\text{tes}}_{i,t}, a^{\text{chp}}_{i,t}, a^{\text{gb}}_{i,t}] & \forall i \in \mathcal{I}_{\text{IG}}, \forall t \in T \end{cases} \tag{13}$$

where the action $a_{i,t}$ consists of two parts: 1) the price decision $a^p_{i,t} \in [0,1]$ representing the magnitude of willing price submitted to the DA market as a ratio of FiT and ToU price differentials $p_{i,t} = \lambda^s_t + a^p_{i,t}(\lambda^b_t - \lambda^s_t)$; and 2) the energy decisions that comprise of $a^{\text{ees}}_{i,t}, a^{\text{tes}}_{i,t} \in [-1,1]$ indicating the mutually exclusive charging (positive) and discharging (negative) power rate of EES $P^{\text{ees}}_{i,t} = P^{\text{eesc}}_{i,t} + P^{\text{eesd}}_{i,t}$ and TES $Q^{\text{tes}}_{i,t} = Q^{\text{tesc}}_{i,t} + Q^{\text{tesd}}_{i,t}$ as a percentage of their power capacity $[-\overline{P}^{\text{ees}}_i, \overline{P}^{\text{ees}}_i]$ and $[-\overline{Q}^{\text{tes}}_i, \overline{Q}^{\text{tes}}_i]$ (as EES and TES cannot charge and discharge simultaneously), and $a^{\text{fc}}_{i,t}, a^{\text{gb}}_t, a^{\text{ehp}}_t, a^{\text{chp}}_t \in [0,1]$ indicating the magnitude of power schedules as a percentage of their power capacity for FC $G^{\text{fc}}_{i,t} \in [0, \overline{G}^{\text{fc}}_i]$, GB $G^{\text{gb}}_{i,t} \in [0, \overline{G}^{\text{gb}}_i]$, EHP $P^{\text{ehp}}_{i,t} \in [0, \overline{P}^{\text{ehp}}_i]$, and CHP $G^{\text{chp}}_{i,t} \in [0, \overline{G}^{\text{chp}}_i]$.

5) State Transition: The state transition from time step t to $t+1$ is governed by $s_{t+1} = \mathcal{T}(s_t, a_{1:I,t}, \omega_t)$, influenced by the combination of the environment state s_t, all agents' actions $a_{1:I,t}$, and environment stochasticity ω_t. In the examined problem, this corresponds to the exogenous states $\omega_t = [L_{1:I,t}, P^{\text{pv}}_{1:I,t}, P^{\text{wg}}_{1:I,t}]$ that are decoupled from the agents' actions and are 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 [22].

By contrast, the state transitions of endogenous states $S^{\text{ees}}_{i,t}$ and $S^{\text{tes}}_{i,t}$ are determined by actions $a^{\text{ees}}_{i,t}$ and $a^{\text{tes}}_{i,t}$, respectively. Given EES as an example, the mutually quantities $P^{\text{eesc}}_{i,t}$ and P^{eesd}_t are managed by action $a^{\text{ees}}_{i,t}$ and are also restricted by its parameters of the minimum/maximum energy level $\underline{E}^{\text{ees}}_i$,$\overline{E}^{\text{ees}}_i$, and the charging/discharging efficiencies η^{eesc}_i and η^{eesd}_i, which are expressed as

$$P^{\text{eesc}}_{i,t} = [\min(a^{\text{ees}}_{i,t}\overline{P}^{\text{ees}}_i, (\overline{E}^{\text{ees}}_i - E^{\text{ees}}_{i,t})/(\eta^{\text{eesc}}_i \Delta t)]^+ \tag{14}$$

$$P^{\text{eesd}}_{i,t} = [\max(a^{\text{ees}}_{i,t}\overline{P}^{\text{ees}}_i, (\underline{E}^{\text{ees}}_i - E^{\text{ees}}_{i,t})\eta^{\text{eesd}}_i/\Delta t]^- \tag{15}$$

where $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$. Given the charging and discharging powers $P^{\text{eesc}}_{i,t}$ and $P^{\text{eesd}}_{i,t}$ and efficiencies η^{eesc}_i and η^{eesd}_i, the state transition of $E^{\text{ees}}_{i,t}$ from t to $t+1$ can be expressed as

$$E^{\text{ees}}_{i,t+1} = E^{\text{ees}}_{i,t} + P^{\text{eesc}}_{i,t}\Delta t\eta^{\text{eesc}}_i + P^{\text{eesd}}_{i,t}\Delta t/\eta^{\text{eesd}}_i. \tag{16}$$

Then, the charging and discharging powers $Q^{\text{tesc}}_{i,t}$ and $Q^{\text{tesd}}_{i,t}$ as well as the state transition $E^{\text{tes}}_{i,t}$ of TES can be derived in the similar manner as the EES model (14)–(16).

To this end, the electricity quantity $q_{i,t}$ submitted to the DA market of each agent i at time step t can be expressed as the summation of its individual electric demand and supply power, where the positive value represents the electricity demand to buy, while the negative value represents the electricity generation to sell in the DA market

$$q_{i,t} = \begin{cases} (P^l_{i,t} - P^{\text{pv}}_{i,t} - P^{\text{fc}}_{i,t} + P^{\text{ees}}_{i,t})\Delta t & \forall i \in \mathcal{I}_{\text{RG}}, \forall t \in T \\ (P^l_{i,t} - P^{\text{pv}}_{i,t} + P^{\text{ehp}}_{i,t} + P^{\text{ees}}_{i,t})\Delta t & \forall i \in \mathcal{I}_{\text{CG}}, \forall t \in T \\ (P^l_{i,t} - P^{\text{wd}}_{i,t} - P^{\text{chp}}_{i,t} + P^{\text{ees}}_{i,t})\Delta t & \forall i \in \mathcal{I}_{\text{IG}}, \forall t \in T. \end{cases} \tag{17}$$

After collecting the price–quantity bids $(p_{i,t}, q_{i,t})$ from all the participating agents, the auctioneer allocates the order books $k^b_t(i, p_{i,t}, q_{i,t}), \forall i \in \mathcal{B}$ and $k^s_t(i, p_{i,t}, q_{i,t}), \forall i \in \mathcal{S}$, clears the DA market (see Algorithm 1), and publishes the market outcomes $[p^l_{1:I,t}, q^l_{1:I,t}, q^g_{1:I,t}, k^b_t, k^s_t]$, which comprises: 1) the local information of cleared trading price $p^l_{i,t}$, cleared trading quantity $q^l_{i,t}$, and the remaining/unmatched quantity balanced with the utility company $q^g_{i,t}$ for each agent i; and 2) the public market information of updated order books k^b_t and k^s_t.

6) Reward Function: The reward function for each agent i at time step t is designed as two parts: 1) the energy and environment costs and 2) the penalty imposed to avoid the constraint violations of the MES operation model. Specifically, for these agents who are successfully matched in the DA market will receive the cleared local trading price $p^l_{i,t}$ and quantity $q^l_{i,t}$, then each agent i can calculate its corresponding electricity cost/revenue in the DA market, and the remaining/unmatched quantity $q^g_{i,t}$ will be bought or sold with the utility company

at ToU λ_t^b or FiT λ_t^s. For these agents who are unsuccessfully matched in the DA market, their quantity $q_{i,t}^g = q_{i,t}$ (i.e., $q_{i,t}^l = 0$) will be directly traded at λ_t^b or λ_t^s. As a result, the reward term corresponding to the electricity cost for each agent i at time step t can be formulated as

$$r_{i,t}^e = -(p_{i,t}^l q_{i,t}^l \cdot \mathbb{1}_{i,t} + \lambda_t^b [q_{i,t}^g]^+ + \lambda_t^s [q_{i,t}^g]^-) \tag{18}$$

where the indicator $\mathbb{1}_{i,t} = 1$ if $i \in \mathcal{B}$ and $\mathbb{1}_{i,t} = -1$ if $i \in \mathcal{S}$. Furthermore, the reward terms corresponding to the gas cost and the environment cost out of the DA market for each agent i at time step t can be, respectively, formulated as

$$r_{i,t}^g = -\lambda_t^g G_{i,t}^g \Delta t, \quad r_{i,t}^c = -\lambda_t^c E_{i,t}^c \tag{19}$$

where the gas quantity purchased from the natural gas grid varies for three kinds of MEMGs: $G_{i,t}^g = G_{i,t}^{\text{gb}}\ \forall i \in \mathcal{I}_{\text{RG}} \cup \mathcal{I}_{\text{CG}}$, $G_{i,t}^g = G_{i,t}^{\text{chp}} + G_{i,t}^{\text{gb}}\ \forall i \in \mathcal{I}_{\text{IG}}$.

Note that, in (17), the electricity demand and supply in each MEMG can always be balanced through the internal system together with the external DA market at each time step. However, the heat demand and supply may not be balanced, since extra heat cannot sell back to the grid. More specifically, the power schedules of components (i.e., FC, GB, EHP, CHP, and TES) controlled by actions only respect their individual operation models (e.g., power capacity). However, they do not make sure that the heat demand and supply are always balanced. The main factor leading to this issue is that the action selections in the RL algorithm for different dimensions are independent, decoupling the correlation in the optimization-based approach. To adequately account for such operation constraints of heat demand–supply balance, we introduce a penalty term $r_{i,t}^p$ for each agent in the reward function, which penalizes the extent of violation of the heat demand–supply balance constraint, with κ denoting a large (negative) penalty weighting factor to ensure its feasibility

$$r_{i,t}^p = \begin{cases} \kappa |Q_{i,t}^l - Q_{i,t}^{\text{fc}} - Q_{i,t}^{\text{gb}} + Q_{i,t}^{\text{tes}}| & \forall i \in \mathcal{I}_{\text{RG}}\ \forall t \in T \\ \kappa |Q_{i,t}^l - Q_{i,t}^{\text{ehp}} - Q_{i,t}^{\text{gb}} + Q_{i,t}^{\text{tes}}| & \forall i \in \mathcal{I}_{\text{CG}}\ \forall t \in T \\ \kappa |Q_{i,t}^l - Q_{i,t}^{\text{chp}} - Q_{i,t}^{\text{gb}} + Q_{i,t}^{\text{tes}}| & \forall i \in \mathcal{I}_{\text{IG}}\ \forall t \in T \end{cases}. \tag{20}$$

Thus, the final reward function $r_{i,t}$ of each MGCC agent i at time step t can be expressed as

$$r_{i,t} = r_{i,t}^e + r_{i,t}^g + r_{i,t}^c + r_{i,t}^p \quad \forall i \in I\ \forall t \in T. \tag{21}$$

III. Proposed MARL Method

To solve the POMDP defined above, we propose a novel MARL method named DA-MATD3 with its general flowchart being shown in Fig. 2. DA-MATD3 derives three concrete implementation details that are insightful and particularly critical to our proposed MEMG energy management coordination problem: 1) learning an abstracted Q-value function for each agent through the DA market public order books to protect the private information of each MEMG; 2) forming an actor-critic architecture to handle the high-dimensional continuous state and the action spaces of the MEMGs; and 3) taking advantage of double critic networks in the twin delayed deep deterministic policy gradient (DDPG) (TD3) algorithm [32] to address the Q-value overestimation problem, thereby stabilizing the training performance.

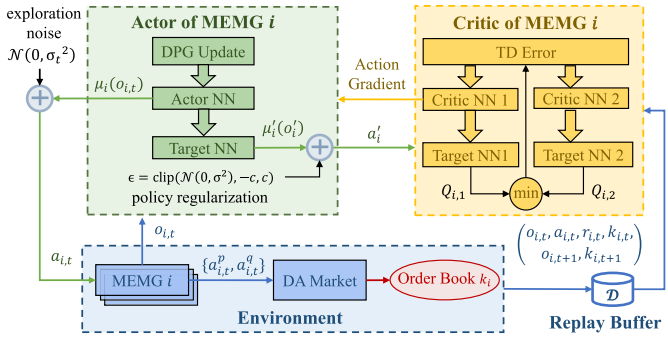

Fig. 2. Flowchart of the proposed DA-MATD3 method.

A. Abstracted Q-Value Function

As discussed in Section I-B, it is challenging to directly acquire the local observations and actions by other agents in our proposed problem since the MEMGs are not willing to share their energy portfolios, technical parameters, and energy usage behaviors. This article, thus, assumes that the agents can use the public order books that epitomize the key information of the DA market (thereby abstracting all agents' price–quantity bid information) in the centralized training process. This substantial improvement protects the privacy of all the agents. To this effect, we approximate the centralized Q-value as

$$Q_i(o_{1:I}, a_{1:I}) \approx Q_i(o_i, a_i, k_i) \tag{22}$$

where $k_i = \{k_j^b, k_j^s\ \forall j \in \mathcal{I} \setminus \{i\}\}$ denotes the combination of buy and sell order books of all the agents other than agent i in the DA market. k_i is an embedded function of order books k_j^b and k_j^s that not only abstracts all other agents' observations (e.g., E_j^l, P_j^{pv}, and P_j^{wd}) as well as actions of the price bids a_j^p and the quantity bids resulting from their energy decisions (e.g., a_j^{ees}, a_j^{fc}, a_j^{ehp}, and a_j^{chp}) but also displays the DA market dynamics of local trading activities. As a result, this combination provides a good approximation of agents' observations and actions as well as the DA market dynamics. Incorporating k_i into the critic estimation, each agent can make acquainted decisions on the basis of the impact of other agents' actions, albeit not knowing their energy portfolios and usage activities, protecting the privacy of each MEMG.

B. MATD3

MATD3 [30], an extension of TD3 to multiagent setup, addresses the stability concern that occurred in conventional MADDPG by three key features: 1) using a pair of critics that estimate the current Q-value via a separate target value function; 2) updating the policy less frequently (delayed update) than the Q-value function; and 3) smoothing the target policy by using a (noise) regularization technique.

1) Twin Critic Networks: The overestimation bias in the conventional MADDPG method has been discussed in [30]. Inspired

by the technique in *double Q-learning* [33] using a separate target Q-value function to estimate the current Q-value, thus reducing the bias, we introduce for each agent i two separate online critic networks ($Q_{i,1}$ and $Q_{i,2}$) parameterized by $\theta_{i,1}$ and $\theta_{i,2}$, along with two target critic networks ($Q'_{i,1}$ and $Q'_{i,2}$) parameterized by $\theta'_{i,1}$ and $\theta'_{i,2}$. Then, the two target values used to update the critic can be written as

$$\begin{aligned} y_{i,1} &= r_i + \gamma Q_{i,1}\left(o'_i, \mu'_i(o'_i), k'_i\right) \\ y_{i,2} &= r_i + \gamma Q_{i,2}\left(o'_i, \mu'_i(o'_i), k'_i\right). \end{aligned} \tag{23}$$

However, the values of $Q_{i,1}$ and $Q_{i,2}$ cannot be equal, and it is inevitable that the high value may be overestimated. Therefore, we make a slight change on the basis of *double Q-learning* and take the minimum value between these two estimates to get the target Q-value for each agent i

$$y_i = r_i + \gamma \min_{k=1,2} Q_{i,k}\left(o'_i, \mu'_i(o'_i), k'_i\right). \tag{24}$$

With this improvement, MATD3 can simultaneously train two critic networks and pick the minimum value of them, thus alleviating the overestimation phenomenon.

2) Delayed Policy Updates: Another potential failure in MADDPG is the variance, which generates noisy gradients during the policy update, thus slowing down the update speed and leading to poor performance [30]. Similar to MADDPG, MATD3 also introduces the target networks to achieve stability. Apart from this, the algorithm also proposes to delay the actor network update until the critic network is updated after a fixed number of time steps. As such, the updates of actor and critic networks are decoupled, i.e., the actor network is updated at a lower frequency than the critic network, to first achieve an accurate Q-value before it is used to update the policy. This less frequent policy update will have a Q-value estimate with lower variance, resulting in better policy performance.

3) Target Policy Smoothing Regularization: The final technique of MATD3 is smoothing the target policy. Deterministic policies trend to produce the high variance of the target when updating the critic; this is caused by overfitting to narrow peaks in the Q-value estimate [30]. MATD3 reduces this variance by adding a clipped Gaussian noise $\epsilon = \text{clip}(\mathcal{N}(0, \sigma^2), -c, c)$ to the actions in the critic update: $a'_i = \mu'_i(o'_i) + \epsilon$. This serves as a regularization, such that all the actions within this small area have similar Q-values, thereby reducing the variance in the associated estimations. The complete target, then, resolves to

$$y_i = r_i + \gamma \min_{k=1,2} Q_{i,k}\left(o'_i, \mu'_i(o'_i) + \epsilon, k'_i\right). \tag{25}$$

C. Training Process

DA-MATD3 is an off-policy MARL method that requires the past experiences to update the networks. To this end, an experience replay buffer $\mathcal{D}_i$ is employed for each agent i. The buffer is a cache storing the past experiences of agent i acquired from the environment (an experience is a transition tuple $(o_{i,t}, a_{i,t}, r_{i,t}, k_{i,t}, o_{i,t+1}, k_{i,t+1})$. For each time step t, we sample uniformly a minibatch of N experiences from each agent's corresponding replay buffer $\{(o_i^n, a_i^n, r_i^n, k_i^n, o_i^{n+1}, k_i^{n+1})\}_{n=1}^N \sim \mathcal{D}_i$ to compute the mean-squared temporal difference (TD) error of two online critic networks as

$$\mathcal{L}(\theta_{i,1}) = \frac{1}{N}\sum_{n=1}^{N}\left[\left(y_i^n - Q_{i,1}(o_i^n, a_i^n, k_i^n)\right)^2\right] \tag{26}$$

$$\mathcal{L}(\theta_{i,2}) = \frac{1}{N}\sum_{n=1}^{N}\left[\left(y_i^n - Q_{i,2}(o_i^n, a_i^n, k_i^n)\right)^2\right] \tag{27}$$

where

$$y_i^n = r_i^n + \gamma \min_{k=1,2} Q_{i,k}\left(o_i^{n+1}, \mu'_i(o_i^{n+1}) + \epsilon, k_i^{n+1}\right). \tag{28}$$

The online actor network employs the delayed update after d critic updates, its policy gradient can be expressed as

$$\begin{aligned} \nabla_{\phi_i} J(\mu_i) = \frac{1}{N}\sum_{n=1}^{N}\big[&\nabla_{\phi_i}\mu_i(o_i^n)\nabla_{a_i^n} Q_{i,1} \\ &\times(o_i^n, a_i^n, k_i^n)|_{a_i^n=\mu(o_i^n)}\big]. \end{aligned} \tag{29}$$

The target networks of two critic and one actor are also employed as the delayed updates after d critic updates

$$\theta'_{i,1} \leftarrow \tau\theta_{i,1} + (1-\tau)\theta'_{i,1} \tag{30}$$

$$\theta'_{i,2} \leftarrow \tau\theta_{i,2} + (1-\tau)\theta'_{i,2} \tag{31}$$

$$\phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i \tag{32}$$

where τ is the soft update rate for their target networks. Moreover, in order to help the agents explore the environment and acquire more valuable experiences, we add a random Gaussian noise $\mathcal{N}(0, \sigma_t^2)$ to the online policy $\mu_i(o_{i,t})$ of each agent i, constructing an exploration policy

$$\hat{\mu}_i(o_{i,t}) = \mu_i(o_{i,t}) + \mathcal{N}(0, \sigma_t^2). \tag{33}$$

Finally, the overall training process of the proposed DA-MATD3 is summarized in Algorithm 2.

IV. Case Studies

A. Experimental Setup and Implementation

1) Experiment Setup: We implement experiments on a real-world dataset recorded from Open Energy Data Initiative [34] and RWTH Aachen University [35]. We collect the corresponding EL, HL, and PV and wind power of residential, commercial, and industrial users with hourly resolution for our experiments. Then, these energy users can be classified and aggregated into three MEMGs. To further account for the uncertainties, we add the Gaussian noise [zero mean and 5% standard deviation (std)] to the original collected data as the train set, while using the original collected data as the test set. The operating parameters of MEMGs' controllable components are derived from [36]. ToU tariff [37] selected as the grid electricity buy price varying for the time: 0.1129 \$/kWh at 20:01–17:00 (next day) and 0.2499 \$/kWh at 17:01–20:00. FiT as the grid electricity sell price, natural gas price, and carbon price are flat over the day at 0.04 \$/kWh [38], 0.0338 \$/kWh [39], and 0.0316 \$/kg [40], respectively. The averaged carbon emission of using natural gas is 0.245 kg CO_2/kWh [40].

Algorithm 2: DA-MATD3 for I Agents.

Initialize weights $\theta_{i,1}, \theta_{i,2},$ and ϕ_i for the online networks and copy them to the target network weights $\theta'_{i,1}, \theta'_{i,2},$ and ϕ'_i for each agent i
Initialize replay buffer $\mathcal{D}_i$ for each agent i
for episode (i.e., trading day) $= 1$ to M **do**
  Initialize the environment $\mathcal{E}$ and Gaussian noise $\mathcal{N}(0, \sigma_t^2)$
  for time step (i.e., 1 h) $t = 1$ to T **do**
    For agent i, select action $a_{i,t} = \hat{\mu}_i(o_{i,t})$ in (33)
    Execute actions $a_{1:I,t}$ to the DA market, then observe reward $r_{i,t}$, next observation $o_{i,t+1}$, and order books $k_{i,t+1}$
    For agent i, store $(o_{i,t}, a_{i,t}, r_{i,t}, k_{i,t}, o_{i,t+1}, k_{i,t+1})$ in $\mathcal{D}_i$
    Update local observations for next time step $o_{i,t} \leftarrow o_{i,t+1}$
    for agent $i = 1$ to I **do**
      Sample uniformly a minibatch of N experiences $(o_i^n, a_i^n, r_i^n, k_i^n, o_i^{n+1}, k_i^{n+1})$ from $\mathcal{D}_i$
      Compute critic target value in (28)
      Update two online critic networks in (26) and (27)
      if t mod $d = 0$ **then**
        Update online actor network in (29)
        Update parameters of target networks in (30)–(32)
      end if
    end for
  end for
end for

2) Benchmarks: We compare the proposed DA-MATD3 with the conventional ZI strategy and three state-of-the-art MARL methods of IDDPG, MADDPG, and MATD3. To further evaluate the benefit of the energy coordination architecture, we benchmark the performance against one scenario that each MGCC agent trades independently with the utility company using DDPG without MEMG energy coordination (UDDPG).

3) Implementations and Hyperparameter Selections: For all the examined five MARL methods, we use an Adam optimizer [41] for both the actor and critic networks with the same learning rate $\alpha = 10^{-3}$ [30]. The sizes of replay buffer $\mathcal{D}$ and batch N are 10^5 and 10^2 [30], respectively. We employ a soft update rate $\tau = 10^{-2}$ [30] and a discount rate $\gamma = 0.9$. The delayed step $d = 2$ [30] for DA-MATD3. For all the networks, we use multilayer perceptron (MLPs) with two hidden layers with 400 and 300 units, respectively. The sigmoid activation function is used as the actor outputs. The outputs are, then, scaled linearly to their individual action space. For all the examined methods, we run 5×10^3 episodes to evaluate their training performance with ten random seeds for both environment and network initialization. The values of the hyperparameters $\alpha, \tau,$ and d were set based on the original MATD3 [30] paper. The `grid search` function [42] was used to determine the value of hyperparameter γ to obtain the best performance.

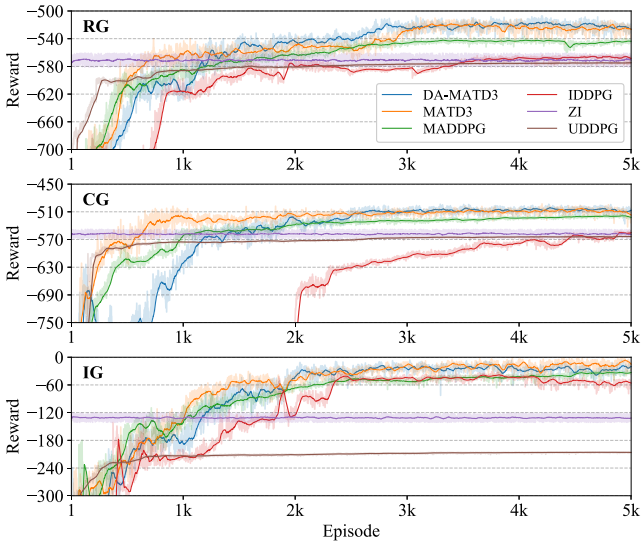

Fig. 3. Episodic reward of three MEMGs for different control methods.

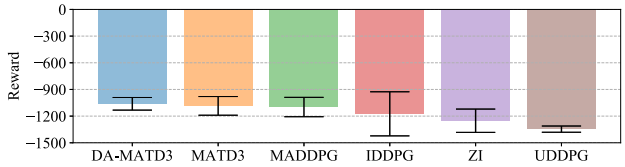

Fig. 4. Mean and std of three MEMGs' aggregated reward at convergence for different control methods.

TABLE I
TEST ENERGY COST AND CARBON EMISSION OF THREE MEMGS AND COMMUNITY FOR DIFFERENT CONTROL METHODS

Method	Energy Cost ($)				Carbon Emission (kg)			
MEMGs	RG	CG	IG	Total	RG	CG	IG	Total
UDDPG	479	515	163	1157	2816	1310	1143	5269
ZI	464	504	91	1059	2857	1290	1164	5311
IDDPG	448	490	41	979	2915	1269	1190	5374
MADDPG	440	471	−12	899	3048	1237	1064	5349
MATD3	418	450	−11	857	2997	1112	1144	5253
DA-MATD3	415	456	−2	869	2926	939	1222	5087

B. Performance Evaluation

We compare the training performance of five examined MARL methods and the conventional ZI strategy for the test set. Specifically, Fig. 3 illustrates the convergence curve of episodic reward of three MEMGs for different control methods, where the solid lines and the shaded areas, respectively, depict the moving average over 50 episodes and the oscillations of the reward during the training process. The converged performance of mean and std of three MEMGs' aggregated reward are also compared in Fig. 4. Furthermore, their energy (electricity and gas) costs and carbon emissions for the test dateset are also presented in Table I for comparison.

Our first observation in Fig. 3 is that all five MARL methods show an upward trend, and their policies are being improved, even for the UDDPG method without considering the energy coordination benefits. On the other hand, IDDPG, the most straightforward MARL method, exhibits the highest oscillation

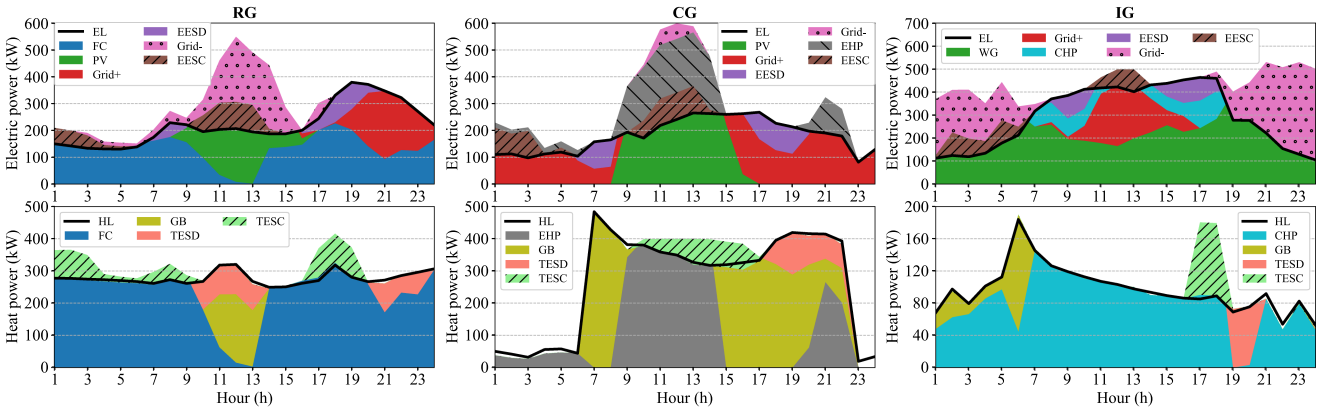

Fig. 5. Power supply and demand for three MEMGs under the DA-MATD3 method.

and unstable learning behavior, ultimately failing to reach an optimal policy (the highest carbon emission). As discussed in Section III-A, this is because IDDPG focuses on local information while ignoring the others' behaviors, rendering the environment dynamics nonstationary. As such, MADDPG and MATD3 with centralized training can effectively mitigate such nonstationarity issues and exhibit superior training performance. Furthermore, MATD3 owing to its double critic networks (more accurate Q-value estimation) can achieve a higher reward with regard to MADDPG. However, both the methods suffer from the privacy issue requiring all others' local observations and actions for the centralized critic. Our proposed DA-MATD3 method learns the DA market dynamics directly by extracting the others' observations and actions through the DA market public order books. In addition, the performance of the traditional ZI strategy during the training process is illustrated in Fig. 3. ZI as a static control method does not tend to go up but tends to flatten out over 5000 episodes.

The mean and std of the aggregated rewards of three MEMGs are quantified in Fig. 4. The figure shows that DA-MATD3 has the best performance, since it achieves the highest reward among all six control methods. DA-MATD3 also has lower std compared to MATD3, MADDPG, IDDPG, and ZI, so that it is more effective in stabilizing the training performance. UDDPG obtains much lower reward than DA-MATD3, even though its std is lower than DA-MATD3. The reason is that UDDPG does not consider a DA market; therefore, the economic benefits of energy coordination cannot be obtained. The test results presented in Table I obtain the similar performance as the training results in Fig. 3. The proposed DA-MATD3 achieves 7.31%, 6.50%, 6.25%, 4.67%, and 2.52% lower total energy costs and carbon emissions than UDDPG, ZI, IDDPG, MADDPG, and MATD3, respectively.

C. Analysis of Multienergy Management

To further validate the learned policies in DA-MATD3 for the test set, we provide the energy management schedules of three MEMGs for both the electric and heat supplies in Fig. 5. Residential MEMG features abundant PV production during mid-day hours and high EL peaks during night hours as well as a relatively flat HL profile. As its high combined electricity and heating generation efficiencies, the FC is learned to supply both EL and HL over the day, apart from the mid-day with PV sources. Furthermore, the MGCC learns to use the storage (EES and TES) flexibility to charge power when energy prices are low or PV is abundant and discharge power when the energy price is high or HL is at the peak. Finally, GB is a backup component to supply HL when the FC is not in use. Similar to the residential MEMG, the commercial MEMG also features abundant PV, but its HL is concentrated during the daytime. Without the converter from natural gas, the electricity grid and PV are major sources to supply EL. The EHP is used to supply HL during the mid-day hours by converting the free PV from electricity to heat power, while EES and TES also exhibit their flexibility to charge cheap and free energy and discharge them to the peak demand hours. Finally, GB in the heat sector is used to supply the left part of HL. Unlike residential and commercial ones, the industrial MEMG installs a WG and its energy usage mainly focuses on EL. It can be observed that there is abundant WG production supplying EL and is also used for EES charging power and surplus fed to the grid to obtain extra revenue. The electricity grid partly supplies EL during the mid-day hours with low wind sources. In the heating sector, CHP accounts for the major proportion of HL supply, while TES is learned to discharge to reduce CHP usage when energy prices are high. It can be concluded that the proposed DA-MATD3 is able to learn effective energy management policies for all three MEMGs to various price signals, demand patterns, and renewable output. In addition, the complementary effect among multienergy vectors (interaction between electric and heat supplies) can also be verified based on the above analysis.

D. Benefits of Energy Coordination

Having demonstrated the superiority of the DA-MATD3 method over the state-of-the-art MARL methods and analyzed the energy schedules of three MEMGs, this section aims to compare the trading strategies under the dynamic DA-MATD3 method with the statistic ZI policy and quantifying the benefits of energy coordination among three MEMGs. Fig. 6 shows the net load (positive for consumption and negative for generation)

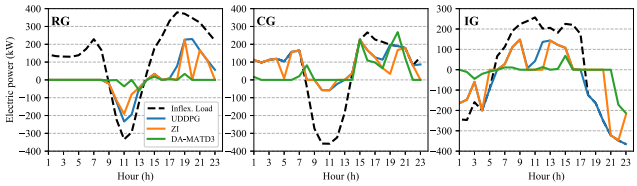

Fig. 6. Net loads for three MEMGs under UDDPG, ZI, and DA-MATD3 methods.

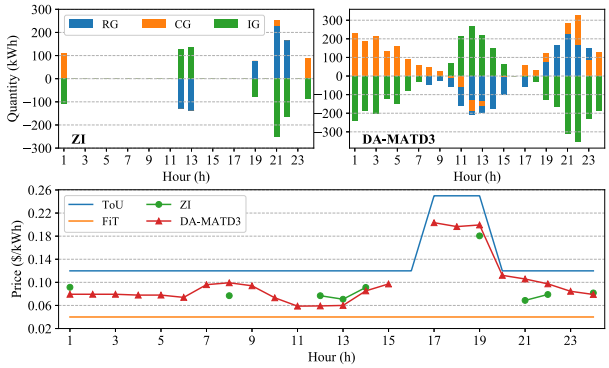

Fig. 7. Local trading quantities and prices under ZI and DA-MATD3 methods.

of three MEMGs under the methods of UDDPG without energy coordination and ZI and DA-MATD3 with energy coordination but in different trading strategies. Dash lines as the baselines represent the aggregated load of electric demand and renewable. Fig. 7 illustrates the local trading quantities and the averaged trading prices under ZI and DA-MATD3 methods.

When energy coordination is allowed in the DA market, MEMGs with energy surplus/deficiency are incentivized to trade locally. As a result, we can observe that compared with UDDPG, the generation and demand of three MEMGs in Fig. 6 are both reduced under ZI and DA-MATD3, since an amount of energy is balanced locally in the DA market, which can also be confirmed in Fig. 7. The figure shows that the DA-MATD3 method trades more frequently and in greater quantities than the ZI method due to the following reasons.

1) For the DA-MATD3 method, the agents are trained to select the suitable trading prices, so that the buyers and the sellers can achieve more trading deals. For the ZI method, the trading prices of the MEMGs are chosen randomly within the range of FiT and ToU, which affects how many times the trading deals are successful.
2) For the DA-MATD3 method, the agents are more likely to trade larger quantities in the DA market to reduce the costs, since each agent considers others' trading strategies. For the ZI method, each MEMG decides the energy trading quantity without considering the trading strategies of the other MEMGs.

More importantly, compared with the nonstrategically sampling behaviors in the ZI method, MGCC agents under DA-MATD3 learn to trade a large amount of energy locally, thereby reducing their dependence on the utility company. Such results can also be validated in Table II: 1) there is no internal trading

TABLE II
COMMUNITY DAILY INTERNAL, EXTERNAL TRADING QUANTITIES, AND ENERGY COSTS UNDER UDDPG, ZI, AND DA-MATD3 METHODS

Method	Internal (kWh)	External (kWh)	Energy Cost ($)
UDDPG	-	7382	1151
ZI	1929	5327	1062
DA-MATD3	7263	1933	881

under UDDPG, so the net demand and generation (7382 kWh in total) are all bought at high ToU and sold at low FiT; 2) ZI achieves $89 total cost saving by 1929 kWh internal trading within the DA market; and 3) DA-MATD3 achieves the lowest total energy cost by making the highest internal trading at 7263 kWh. In relative terms, DA-MATD3 achieves 2.82/1.76 times lower external trading with the utility company (higher balance of local demand-generation) and 30.65%/20.54% lower energy cost (more economic benefits of local trading) over UDDPG/ZI methods.

V. CONCLUSION

This article proposed a novel MARL method to address the energy coordination problem of MEMGs local trading in a highly efficient DA market, incentivizing MEMGs to participate in local trading with economic benefits. The examined MEMGs, featuring various demand and renewable characteristics, were categorized into residential MEMGs, commercial MEMGS, and industrial MEMGs. The proposed MARL method named DA-MATD3: 1) constructs the centralized critic by abstracting the others' observations and actions through the DA market public information, thereby preserving MEMGs' privacy and capturing the market dynamics and 2) uses a pair of critic networks to overcome the Q-value overestimation issue and stabilize the training performance. The effectiveness of the proposed DA-MATD3 method was evaluated through simulations using a real-world setting. Specifically, the proposed method achieved superior performance in reducing both energy costs and carbon emissions compared to the state-of-the-art ZI and MARL methods. Finally, the trading strategies and outcomes were also analyzed to show the significant economic benefits of the community by more internal energy trading among three MEMGs within the DA market.

Future work aims at enhancing the proposed work from two directions. First, the DA market introduced in this article focuses on electricity trading. Future work will explore a new market mechanism enabling multienergy trading within a local MEMG community. Second, although this article focuses on a local energy community, the proposed method can be extended to a larger and wider energy community with the following changes: 1) in the system model, the transmission losses need to be considered, as long-distance transmission tends to lose energy; 2) the matching algorithm in the DA market should take the distance into consideration when matching a buyer and a seller; and 3) distribution network constraints need to be considered, since different distribution networks often have different constraints such as transformer and line limitations, phase unbalance, and voltage stability.

REFERENCES

[1] G. Strbac *et al.*, "Cost-effective decarbonization in a decentralized market: The benefits of using flexible technologies and resources," *IEEE Power Energy Mag.*, vol. 17, no. 2, pp. 25–36, Mar./Apr. 2019.

[2] *Manag. Large-Scale Penetration of Intermittent Renewables*, Massachusetts Inst. Technol. Energy Initiative, Cambridge, MA, USA, Apr. 2011.

[3] P. Mancarella, "MES (multi-energy systems): An overview of concepts and evaluation models," *Energy*, vol. 65, pp. 1–17, Feb. 2014.

[4] N. Lidula and A. Rajapakse, "Microgrids research: A review of experimental microgrids and test systems," *Renew. Sustain. Energy Rev.*, vol. 15, no. 1, pp. 186–202, Jan. 2011.

[5] Z. Li and Y. Xu, "Optimal coordinated energy dispatch of a multi-energy microgrid in grid-connected and islanded modes," *Appl. Energy*, vol. 210, pp. 974–986, Jan. 2018.

[6] P. W. MacAvoy, *The Natural Gas Market*. New Haven, CT, USA: Yale Univ. Press, 2008.

[7] T. Couture and Y. Gagnon, "An analysis of feed-in tariff remuneration models: Implications for renewable energy investment," *Energy policy*, vol. 38, no. 2, pp. 955–965, Feb. 2010.

[8] J. Yang, J. Zhao, F. Luo, F. Wen, and Z. Y. Dong, "Decision-making for electricity retailers: A brief survey," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4140–4153, Sep. 2017.

[9] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nature Energy*, vol. 3, no. 2, pp. 94–101, Feb. 2018.

[10] E. A. M. Cesena, N. Good, A. L. Syrri, and P. Mancarella, "Techno-economic and business case assessment of multi-energy microgrids with co-optimization of energy, reserve and reliability services," *Appl. Energy*, vol. 210, pp. 896–913, Jan. 2018.

[11] C. Li, Y. Xu, X. Yu, C. Ryan, and T. Huang, "Risk-averse energy trading in multienergy microgrids: A two-stage stochastic game approach," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2620–2630, Oct. 2017.

[12] D. Xu *et al.*, "Peer-to-peer multienergy and communication resource trading for interconnected microgrids," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2522–2533, Apr. 2021.

[13] N. Liu, J. Wang, and L. Wang, "Hybrid energy sharing for multiple microgrids in an integrated heat-electricity energy system," *IEEE Trans. Sustain. Energy*, vol. 10, no. 3, pp. 1139–1151, Jul. 2019.

[14] D. Xu *et al.*, "Distributed multienergy coordination of multimicrogrids with biogas-solar-wind renewables," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3254–3266, Jun. 2019.

[15] D. Xu, Q. Wu, B. Zhou, C. Li, L. Bai, and S. Huang, "Distributed multi-energy operation of coupled electricity, heating, and natural gas networks," *IEEE Trans. Sustain. Energy*, vol. 11, no. 4, pp. 2457–2469, Oct. 2020.

[16] W. Gu *et al.*, "Residential CCHP microgrid with load aggregator: Operation mode, pricing strategy, and optimal dispatch," *Appl. Energy*, vol. 205, pp. 173–186, Nov. 2017.

[17] D. Friedman, *The Double Auction Market: Institutions, Theories, and Evidence*. Evanston, IL, USA: Routledge, Mar. 2018.

[18] Z. Li and T. Ma, "Peer-to-peer electricity trading in grid-connected residential communities with household distributed photovoltaic," *Appl. Energy*, vol. 278, Aug. 2020, Art. no. 115670.

[19] P. Vytelingum, D. Cliff, and N. R. Jennings, "Strategic bidding in continuous double auctions," *Artif. Intell.*, vol. 172, no. 14, pp. 1700–1729, Sep. 2008.

[20] D. Qiu, J. Wang, J. Wang, and G. Strbac, "Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 2913–2920.

[21] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized P2P energy trading under network constraints in a low-voltage network," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5163–5173, Sep. 2019.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[23] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3068–3082, Jul. 2020.

[24] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[25] D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac, "Safe reinforcement learning for real-time automatic control in a smart energy-hub," *Appl. Energy*, vol. 309, Mar. 2022, Art. no. 118403.

[26] T. Chen and S. Bu, "Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning," in *Proc. IEEE PES Innov. Smart Grid Technol.*, 2019, pp. 1–5.

[27] T. Zhang, D. Yue, and N. Zhao, "Energy optimization management of multi-microgrid using deep reinforcement learning," in *Proc. Chin. Autom. Congr.*, 2020, pp. 4049–4053.

[28] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6382–6393.

[29] Y. Xu, L. Yu, G. Bi, M. Zhang, and C. Shen, "Deep reinforcement learning and blockchain for peer-to-peer energy trading among microgrids," in *Proc. Int. Conf. Internet of Things*, 2020, pp. 360–365.

[30] J. Ackermann, V. Gabler, T. Osa, and M. Sugiyama, "Reducing overestimation bias in multi-agent domains using double centralized critics," Dec. 2019. [Online]. Available: https://arxiv.org/abs/1910.01465

[31] A. Kaur, J. Kaushal, and P. Basak, "A review on microgrid central controller," *Renewable Sustain. Energy Rev.*, vol. 55, pp. 338–345, 2016.

[32] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[33] H. V. Hasselt, "Double Q-learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 2613–2621.

[34] *Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States*, Office of Energy Efficiency and Renewable Energy, Washington, DC, USA. Accessed: Feb. 24, 2022. [Online]. Available: https://openei.org/datasets/dataset/

[35] *Smart Energy Data: Aachen/ Cologne Virtual Power Plant*, RWTH Aachen Univ., Aachen, Germany. Accessed: Feb. 24, 2022. [Online]. Available: https://data.lab.fiware.org/organization/rwth-aachen-university

[36] W. Huang, N. Zhang, J. Yang, Y. Wang, and C. Kang, "Optimal configuration planning of multi-energy systems considering distributed renewable energy," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1452–1464, Mar. 2019.

[37] *SRP Time-of-Use Price Plan*, Salt River Project, Tempe, AZ, USA. Accessed: Feb. 24, 2022. [Online]. Available: https://srpnet.com/prices/home/tou.aspx

[38] *Renewable Energy Feed-in Tariffs*, Organisation for Economic Co-operation and Development, Paris, France. Accessed: Feb. 24, 2022. [Online]. Available: https://stats.oecd.org/Index.aspx?DataSetCode=RE_FIT#

[39] *United States Natural Gas Ind. Price*, U.S. Energy Inf. Admin., Washington, DC, USA. Accessed: Feb. 24, 2022. [Online]. Available: https://www.eia.gov/dnav/ng/hist/n3035us3m.htm

[40] *U.S. Energy-Related Carbon Dioxide Emissions*, U.S. Energy Inf. Admin., Washington, DC, USA. Accessed: Feb. 24, 2022. [Online]. Available: https://www.eia.gov/environment/emissions/carbon/

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–15.

[42] S. M. LaValle, M. S. Branicky, and S. R. Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *Int. J. Robot. Res.*, vol. 23, nos. 7/8, pp. 673–692, Aug. 2004.

Dawei Qiu (Member, IEEE) received the Ph.D. degree in electrical engineering from the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., in 2020.

He is currently a Research Associate with the Department of Electrical and Electronic Engineering, Imperial College London. His research interests include the development and application of decentralized and market-based approaches for electricity market, peer-to-peer energy trading, multienergy system integration, and microgrid resilience control.

Tianyi Chen (Student Member, IEEE) received the M.Sc. degree in electronics and computer science from the University of Southampton, Southampton, U.K., in 2017. He is currently working toward the Ph.D. degree with the University of Glasgow, Glasgow, U.K.

His current research interests include peer-to-peer energy trading, multienergy systems, machine learning optimization, and deep reinforcement learning application.

Goran Strbac (Member, IEEE) is a Professor of Energy Systems with Imperial College London, London, U.K. He led the development of novel advanced analysis approaches and methodologies that have been extensively used to inform industry, governments, and regulatory bodies about the role and value of emerging new technologies and systems in supporting cost-effective evolution to smart low-carbon future. He is the Director of the joint Imperial-Tsinghua Research Centre on Intelligent Power and Energy Systems, a leading author with Intergovernmental Panel on Climate Change Working Group III, a Member of the European Technology and Innovation Platform for Smart Networks for the Energy Transition, and a Member of the Joint European Union Program in Energy Systems Integration of the European Energy Research Alliance.

Shengrong Bu (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Carleton University, Ottawa, ON, Canada, in 2012.

From 2012 to 2014, she held a research position with Huawei Technologies Canada Inc., Ottawa, as a Natural Sciences and Engineering Research Council (NSERC) Industrial R&D Fellow. From 2014 to 2021, she was a Lecturer with the James Watt School of Engineering, University of Glasgow, Glasgow, U.K. She is currently an Associate Professor with Brock University, St. Catharines, ON, Canada. Her research interests include multivector energy microgrids, smart grids, future wireless networks, cyber security, deep reinforcement learning, and big data analytics.

Dr. Bu was a recipient of three best paper awards at IEEE International conferences. Her work has been supported by Engineering and Physical Sciences Research Council (EPSRC) (U.K.) and NSERC (Canada). Highlights of her professional activities include duties as a peer reviewer for EPSRC and Carnegie Trust, an Associate Editor for *Wireless Networks*, a Topic Editor for *Energies*, and the Technical Program Committee Co-Chair for six international workshops or conference symposiums. Promoting engineering to younger girls and supporting junior female researchers are her passion, and she has been involved with Ontario Go Eng Girl in Canada, Monster Confidence in U.K., and N2Women as a Mentoring Co-Chair.