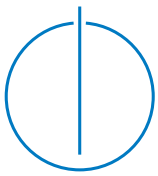
TECHNICAL UNIVERSITY OF MUNICH

DEPARTMENT OF INFORMATICS

Master's Thesis in Informatics

Exploring Cooperation and Competition in P2P Energy Markets through Contract-Augmented Multi-Agent Reinforcement Learning

Saini Rohan Rao



TECHNICAL UNIVERSITY OF MUNICH

DEPARTMENT OF INFORMATICS

Master's Thesis in Informatics

Exploring Cooperation and Competition in P2P Energy Markets through Contract-Augmented Multi-Agent Reinforcement Learning

Erforschung von Kooperation und Wettbewerb in P2P-Energiemärkten durch vertragserweitertes Multi-Agent Reinforcement Learning

Author:	Saini Rohan Rao
Supervisor:	Prof. Dr.-Ing. Matthias Althoff
Advisor:	Michael Eichelbeck, M.Sc.
Submission Date:	29.10.2025

I confirm that this master's thesis is my own work and that I have documented all sources and materials used.

Ich versichere, dass ich diese Master's Thesis selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

Munich, 29.10.2025 Saini Rohan Rao

Acknowledgments

I would like to take a moment to thank all the wonderful people in my life. First and foremost, my amma, Guddu. Her sacrifices have allowed me to dream big. Jatin, your creative genius never fails to motivate me. And papa, who has supported all my career choices along the way. Prost to all my friends in Munich, especially Ashish and Utkarsh—you guys are my family away from home. Elena, danke for having my back over the last year, pretty sure you're the most cheerful PhD I'll ever meet :-)

I would also like to express my gratitude to the CPS Group. It was the awesome "Fundamentals of AI" course by Dr. Althoff that got me into RL, and there has been no looking back. Shout-out to my advisor Michael, thanks for letting me engage in independent research. Freely working on a topic of my liking, without worrying much about consequential results. It did get messy towards the end, but all's well that ends well.

So how does it feel to finally complete a master's at TUM? Phew, perhaps it will take some time to sink in. But one thing's for sure, the experience has been great, exactly what I set out to achieve. And not just the academics, but the life in Munich, or should I say Europe, has been amazing. From writing an application essay on AI right at the last minute, to this acknowledgment, soon to be holding a degree in Advanced AI—I've come a long way. Not sure what the future holds, but the last three years would always be a cherished memory!

Abstract

The increasing penetration of Distributed Energy Resources is transforming power systems into decentralized local networks where individual households act as smart prosumers—producing, consuming, and actively trading energy. Enabling such autonomy calls for market designs and learning frameworks that can coordinate many small actors under uncertainty. Local Energy Markets (LEMs) offer one such economic platform for Peer-to-Peer (P2P) trading, but their design presents a central challenge: balancing the strategic freedom of competitive markets with the fairness and collective benefit of cooperative pooling. This thesis employs Multi-Agent Reinforcement Learning (MARL) to develop autonomous trading agents, allowing for a systematic investigation of this tension. Specifically, we compare a competitive Double Auction (DA) market with a devised Shapley Pooling mechanism (purely cooperative), and introduce a novel "Hybrid Contracting Game". This mixed market is a two-stage MARL framework where agents learn to propose contracts that pre-commit a fraction of their net supply or demand to the Shapley Pool while reserving the rest for DA trading.

Experiments were conducted on a community of eight agents from the real-world Ausgrid dataset, subject to representative grid pricing (ToU and FiT). All three P2P markets significantly reduce community-wide costs (by roughly 21%) and external grid reliance (by up to 25%) compared to a grid-only baseline, clearly demonstrating the merits of LEMs. In each mechanism, agents learn to effectively leverage the flexibility provided by the Energy Storage System throughout the day, and the resulting performance differences are subtle. However, a closer look reveals important distinctions: Shapley Pooling achieves optimal community savings and the most equitable distribution of benefits, whereas the DA market suffers from matching inefficiencies and exploitable strategies. By eliminating price actions, Shapley Pooling simplifies learning and emerges as the more robust and desirable design. The hybrid market, though, yields a surprising insight: when given the choice, agents prefer a cautious compromise between cooperation and competition, converging on balanced 50-50 contracts rather than expected dominant cooperation. The rudimentary nature of this setup, however, precludes drawing concrete conclusions and warrants further exploration into more sophisticated yet tractable contracting augmentations to MARL.

Contents

1. Introduction

The global energy landscape is undergoing a profound transformation, shifting from fossil fuel-based generation toward decarbonized systems increasingly reliant on renewable energy sources [1]. A key driver of this transition is the rapid proliferation of Distributed Energy Resources (DERs) such as rooftop Solar Photovoltaic (PV) systems and Energy Storage (ES) devices [2]. Their widespread adoption marks a shift from centralized generation toward decentralized, prosumer-driven networks, where traditional consumers evolve into prosumers—entities capable of producing, consuming, and storing energy, thus actively participating in the grid [3]. The growing autonomy of these numerous small-scale actors introduces coordination complexities that necessitate new economic platforms to orchestrate their local decisions and interactions. Local Energy Markets (LEMs) [4] fulfill this role, operating within specific geographical boundaries to manage local DERs and mediate the community's interaction with the external utility grid. Within these LEMs, Peer-to-Peer (P2P) energy trading [5] has emerged as a key mechanism enabling prosumers and consumers to directly exchange energy, often bypassing traditional intermediaries. This operational model contrasts sharply with conventional market structures where agents only interact with the grid individually, as illustrated in Figure 1.1. P2P markets aim to balance local supply and demand internally, often at mutually beneficial prices compared to grid tariffs. Through P2P trading, communities can realize significant benefits, including reduced energy costs, greater utilization of local renewable generation, enhanced market transparency, and lower grid dependence [6].

P2P energy trading markets can be broadly categorized based on their coordination architecture into centralized, decentralized, and distributed types (illustrated in Figure Figure 1.2). Centralized markets rely on a central coordinator managing trades, while decentralized markets allow direct negotiation between peers without a central authority. Distributed markets often blend these features, potentially using an auctioneer to facilitate trades indirectly. While recent research has focused more on blockchain-enabled decentralized P2P approaches, this thesis investigates mechanisms falling within the distributed and centralized paradigms.

Realizing the full potential of P2P trading in LEMs involves significant challenges, including the need for automated decision-making strategies for participants navigating dynamic market environments with inherent demand, PV, and price uncertainties. Multi-Agent Reinforcement Learning (MARL), a prominent subfield of AI, is particularly well-suited to address these challenges [6]. MARL enables multiple autonomous

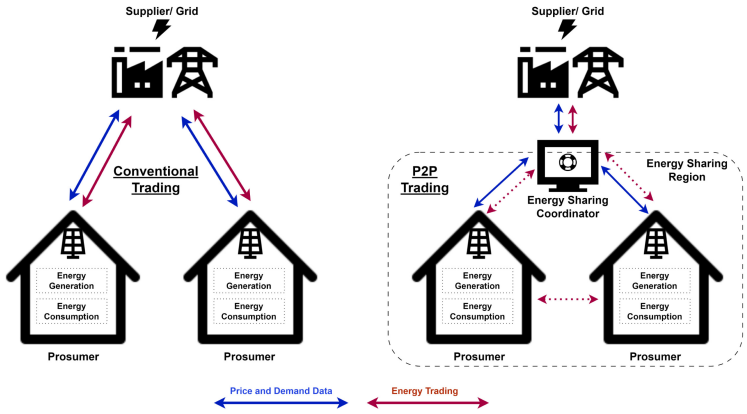
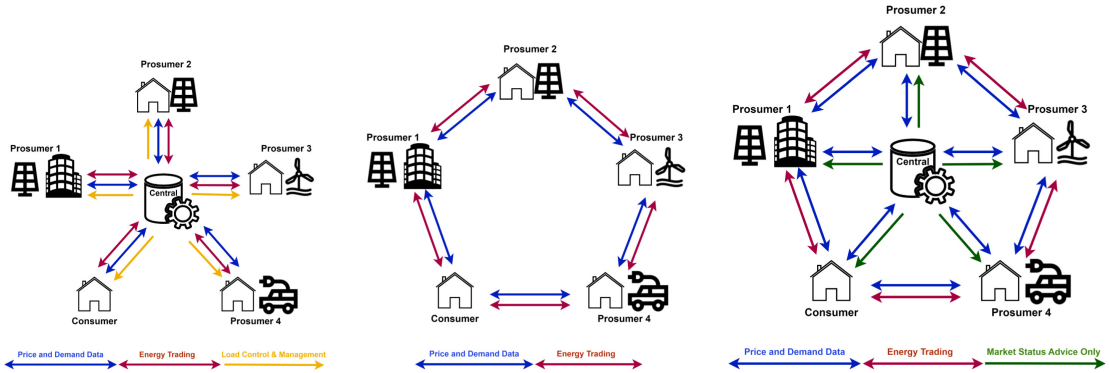


Figure 1.1. Conventional vs P2P Energy Trading Paradigm [6]



(a) Centralized Market (b) Decentralized Market (c) Distributed Market

Figure 1.2. Types of P2P Energy Trading Markets [6]

agents—representing prosumers or consumers—to learn optimal strategies through interaction with the P2P market, aiming to maximize their individual or collective objectives. It offers a powerful framework for developing adaptive strategies in modern energy networks, and has proven effective in applications ranging from load balancing and active voltage control, to coordinated demand-side management [7]. Of these diverse applications, market-level coordination struck us as particularly intriguing, presenting a natural testbed—blending economic behavior with algorithmic multi-agent learning.

Against this backdrop, this thesis leverages MARL to develop and evaluate autonomous agents for P2P trading in LEMs. We primarily aim to answer: How do different P2P market mechanisms compare to the grid-only baseline and to each other in terms of cost reduction, grid reliance, and fairness?

A central challenge in designing P2P markets lies in structuring agent interactions: should they be purely competitive or cooperative? Standard approaches often utilize a Double Auction (DA) Market [8], recognized for its efficiency but inherently foster-

ing competition that can lead to individualistic strategies and potentially inequitable outcomes. Conversely, cooperative pooling focuses on maximizing community-wide benefits, yet requires a principled method, such as Shapley value allocation [9], to fairly distribute the collective gains or costs among participants. This thesis explores this fundamental tension between competition and cooperation using MARL, comparing agent incentives under the two extremes, while also introducing a novel hybrid P2P market.

This Hybrid Contracting Game uses a two-stage MARL design inspired by [10]. By introducing contracts that allow agents to pre-commit a fraction of their exchange to the cooperative pool, this mechanism seeks to answer: Can a hybrid, contract-augmented market effectively navigate the spectrum between purely competitive and cooperative outcomes? The agents first learn robust operational policies and then optimal contract proposals. Through this second stage, we investigate: What market structure do agents prefer when given an explicit choice via contracts? Existing MARL-based studies on P2P energy trading typically assume fixed market rules [6]; here, we instead allow the market structure itself to emerge through learning. While our approach does not necessarily aim to resolve strict social dilemmas as defined by Haupt et al. [10], exploring formal contracts in P2P trading may offer distinct advantages. Contracts could enhance robustness against inherent system uncertainties, such as unexpected fluctuations in local energy volumes or prices; agents might prefer the stability offered by a pre-agreed contract over potentially more volatile, albeit sometimes cheaper, open market interactions. Furthermore, such formal agreements might facilitate coordination in related challenges like demand-side management, where individual comfort preferences can conflict with collective grid stability objectives (representing a dilemma; someone has to compromise). Contract-based MARL thus represents a promising research direction, offering valuable applications for the energy domain.

This thesis is structured as follows. Chapter 2 provides the theoretical background, detailing the DA market, MARL fundamentals, and the Shapley Value. It also includes a literature review of MARL applied to P2P energy trading. Chapter 3 then introduces the experimental setup, including the real-world Ausgrid dataset, grid pricing schemes, and community agent profiles. The core methodological framework is defined in Chapter 4, which specifies the MARL setup and formally outlines the dynamics of the three P2P markets under consideration: DA, Shapley Pooling, and the Hybrid Contracting Game. It also describes the key evaluation metrics used to assess performance. Chapter 5 presents the empirical results, including training stability and performance benchmarks on the held-out test set. These results are then thoroughly analyzed in Chapter 6 through detailed discussions on agent ES scheduling, DA market efficiency, fairness, and the interpretation of the learned optimal contracts. Finally, Chapter 7 concludes the thesis by summarizing the key findings, and Chapter 8 addresses the limitations of the current study while proposing avenues for future work.

2. Background and Related Work

The shift towards LEMs, driven by prosumers with DERs, requires efficient market mechanisms for P2P trading. This chapter provides the necessary background by reviewing three core concepts alongside the relevant literature. We first introduce the DA market as a framework for P2P trading. We then discuss how MARL is used to develop autonomous bidding strategies for agents within this market. Finally, we cover Shapley Value as a method for analyzing agent contributions and fairness.

2.1. Double Auction

The Double Auction (DA) is a market mechanism that facilitates competitive trading among multiple buyers and sellers, and is widely recognized for its high efficiency in resource allocation [8]. A DA market operates in discrete *auction periods*, and within each period, consists of the following core components:

- A set of buyers $\mathcal{B}$, where each buyer $i \in \mathcal{B}$ submits a bid with their desired price p_i^b and quantity q_i^b.
- A set of sellers $\mathcal{S}$, where each seller $j \in \mathcal{S}$ submits an ask with their desired price p_j^s and quantity q_j^s.
- A public *order book*, managed by an *auctioneer*, where all bids are sorted in decreasing order of price, and all asks are sorted in increasing order of price.

A key distinction from a continuous double auction is that a DA clears the market periodically [8]. The auctioneer collects all bids and asks submitted during an entire auction period. The market is then cleared once at the end of the period to determine all matched trades. During clearing, the matching algorithm iterates down the sorted order books. It attempts to match buy orders with sell orders, continuing until the bid price is less than the ask price, or no unmatched sell or buy order exists anymore. For each successful match, the transaction quantity is the minimum of the two orders. The clearing price is found using the traditional mid-pricing method [8].

This process is designed to maximize social welfare [8]. However, it does not guarantee a match for every participant, as the market remains inherently competitive. The full clearing procedure is formally outlined in Algorithm 1. A numerical example of this algorithm, applied to our specific problem setting, is provided in Section 4.2.1.

Algorithm 1 DA Market Clearing Algorithm

Input: For auction period t, buy order book $k_t^b = \{(i, p_{i,t}^b, q_{i,t}^b)\}_{i \in \mathcal{B}}$ and sell order book $k_t^s = \{(j, p_{j,t}^s, q_{j,t}^s)\}_{j \in \mathcal{S}}$
Sort k_t^b in descending order of price $p_{i,t}^b$
Sort k_t^s in ascending order of price $p_{j,t}^s$
Initialize matched trades for period t, $M_t \leftarrow \emptyset$
Initialize buyer index $i \leftarrow 1$, seller index $j \leftarrow 1$
while $i \leq |\mathcal{B}|$ **and** $j \leq |\mathcal{S}|$ **and** $p_{i,t}^b \geq p_{j,t}^s$ **do**
    Matched quantity $q_{\text{trade}} \leftarrow \min(q_{i,t}^b, q_{j,t}^s)$
    Clearing price $p_{\text{trade}} \leftarrow (p_{i,t}^b + p_{j,t}^s)/2$ ▷ Mid-pricing rule
    Add $(id_i, id_j, p_{\text{trade}}, q_{\text{trade}})$ to M_t ▷ Store agent IDs, not sorted indices
    Update buy order: $q_{i,t}^b \leftarrow q_{i,t}^b - q_{\text{trade}}$
    Update sell order: $q_{j,t}^s \leftarrow q_{j,t}^s - q_{\text{trade}}$
    if $q_{i,t}^b = 0$ **then**
        $i \leftarrow i + 1$
    end if
    if $q_{j,t}^s = 0$ **then**
        $j \leftarrow j + 1$
    end if
end while
Return: M_t

2.2. Multi-agent Reinforcement Learning

While the DA market provides the rules for interaction, it does not dictate how agents should learn optimal bidding strategies over time. This sequential decision-making problem is a natural fit for Reinforcement Learning (RL) [11], where agents improve their strategies through trial-and-error. As a data-driven approach, RL excels in dynamic and stochastic environments because it does not require an explicit model of the underlying system dynamics.

To formally structure this learning problem, the multi-agent environment is modeled as a finite Partially Observable Markov Game (POMG) [12, 13] with discrete time steps Δt. A POMG for N agents is defined by a tuple $(\mathcal{S}, \{\mathcal{O}_{1:N}\}, \{\mathcal{A}_{1:N}\}, \mathcal{T}, \{\mathcal{R}_{1:N}\})$. Here, $\mathcal{S}$ is the global state space, $\mathcal{T}$ is the state transition function, while $\{\mathcal{O}_{1:N}\}$, $\{\mathcal{A}_{1:N}\}$, and $\{\mathcal{R}_{1:N}\}$ are respectively, the collections of observation spaces, action spaces, and reward functions for each individual agent. At each time step t, which corresponds to the start of an auction period, every agent n receives a private observation $o_{n,t} \in \mathcal{O}_n$ of the global state. Based on this, it selects an action $a_{n,t} \in \mathcal{A}_n$ according to its policy $\pi_n(o_{n,t})$. In response to the joint action of all agents, the environment transitions to a new state (governed by $\mathcal{T}$) and provides each agent with an individual reward $r_{n,t}$

from its reward function $\mathcal{R}_n$. The objective of each agent is to learn a policy π_n^* that maximizes its expected cumulative discounted reward, defined as $R_n = \sum_{t=0}^{T} \gamma^t r_{n,t}$, over a time horizon T, where $\gamma \in [0, 1)$ is the discount factor. The exact POMG formulation for our problem is detailed in Section 4.1.

2.2.1. Proximal Policy Optimization

To approach the POMG, it is instructive to first consider its single-agent simplification, the Markov Decision Process (MDP), for which RL offers the solution framework [11]. Within this framework, a popular paradigm is the *actor-critic* method. This approach maintains two distinct components: an *actor* (π), which represents the *policy* and decides what actions to take, and a *critic* (V), which evaluates how advantageous those actions are by estimating a *value function*. This function predicts the expected cumulative reward (*return*) from a given state, thereby quantifying how "good" that state is. The actor is trained via *policy gradients* to maximize expected returns, guided by feedback from the critic. The critic, in turn, learns by minimizing the prediction error between its value estimates and the observed returns. For complex problems with large or continuous state-action spaces where tabular methods are infeasible, modern *Deep RL* utilizes neural networks as function approximators. Consequently, both the actor and critic are represented by networks parameterized by θ and ϕ, respectively (π_θ and V_ϕ).

A state-of-the-art algorithm in the actor-critic family is Proximal Policy Optimization (PPO) [14], which is known for its robust performance. The core of PPO's effectiveness is its *clipped surrogate objective function*. This mechanism ensures that policy updates do not deviate too drastically from the previous policy, optimizing the actor within a trusted region to stabilize training. The objective is formally expressed as:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{2.1}$$

The *advantage estimate* $\hat{A}_t$, provided by the critic, quantifies how profitable an action was compared to the average action in that state, serving as the directional signal for policy improvement. This signal is applied to the *probability ratio*, $r_t(\theta) = \frac{\pi_\theta(a_t|o_t)}{\pi_{\theta_{\text{old}}}(a_t|o_t)}$, which measures how likely an action is under the new policy relative to the old one. A positive advantage pushes to increase this ratio, making the action more likely, while a negative advantage pushes to decrease it. The clipping mechanism, controlled by the hyperparameter ϵ, limits how much this ratio can change, preventing destructively large updates and ensuring stable learning.

To ensure sufficient exploration, PPO augments its final objective with an *entropy bonus*. This bonus is weighted by a coefficient and combined with the actor's loss (Equation 2.1) and the critic's value loss to form the complete loss function. The entropy of a policy, $S[\pi_\theta]$, is defined as:

$$S[\pi_\theta](o_t) = -\sum_{a \in \mathcal{A}} \pi_\theta(a|o_t) \log \pi_\theta(a|o_t) \tag{2.2}$$

Maximizing this term encourages the policy to remain stochastic, preventing it from collapsing into a deterministic and potentially suboptimal strategy too early in training. For the continuous action space problem addressed in this thesis, the policy is modeled as a Tanh Normal distribution, and the entropy is therefore calculated using its analytical form [15].

However, directly applying a single-agent algorithm like PPO to a multi-agent setting is problematic due to the challenge of *non-stationarity* [16]. From the perspective of any individual agent n, the environment appears non-stationary because the policies of all other agents, $\{\pi_i\}_{i \neq n}$, are changing simultaneously during training, which violates the Markov assumption. To address this, Multi-Agent PPO (MAPPO) [17] adapts the PPO algorithm by employing the Centralized Training with Decentralized Execution (CTDE) paradigm [16]. Under this framework, agents utilize a *centralized critic* $V_n(s)$ during the training phase to learn a stable value function. It has access to global information, such as the true state of the system s or a concatenation of all agents' observations. This centralized perspective provides a consistent learning signal that resolves the non-stationarity issue. During execution, each agent's *decentralized actor*, $\pi_n(a_n|o_n)$, relies solely on its own local observation o_n to select an action. This approach effectively balances stable learning with scalable, independent execution. The general CTDE structure, often depicted using a centralized Q-function, is shown in Figure 2.1. Although MAPPO utilizes a centralized state-value function (V_n) instead.

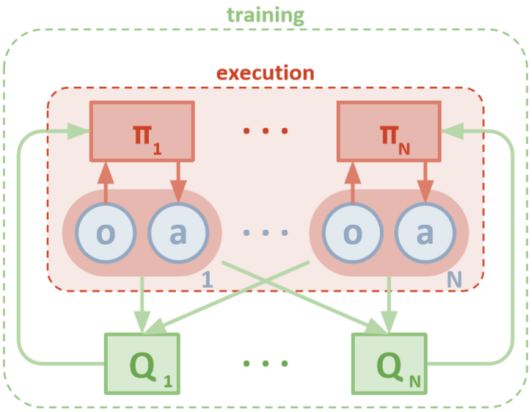


Figure 2.1. Centralized Training with Decentralized Execution in MARL [16]

2.2.2. MARL for P2P Energy Trading

A foundational study by Qiu et al. explored the potential of MARL for P2P energy trading within a DA market framework [18]. In their work, the authors proposed DA-MADDPG, an adaptation of the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [16] algorithm, where each agent's centralized critic was enhanced with

public information from the DA order book. They evaluated this approach in a simulated environment with four prosumers and four consumers, utilizing real-world load and generation data [19]. The results demonstrated the efficacy of the method, with DA-MADDPG achieving up to 13.22% and 16.69% lower total energy bills compared to a Zero Intelligence (ZI) strategy [8] and direct grid trading, respectively. Despite these promising results, the authors acknowledged limitations, including the need to test scalability with a larger number of agents and to incorporate physical network constraints for greater realism.

While the CTDE framework addresses non-stationarity, the centralized critic's reliance on global information creates a bottleneck in large-scale systems, leading to scalability and privacy challenges [20, 21, 22, 23]. To mitigate this, subsequent research has focused on abstracting the collective behavior of other agents into a concise representation for the critic. One prominent approach is *mean-field approximation*, which approximates the influence of a large population by a single average effect, allowing the critic's input to be independent of the number of agents. This was shown to be effective for 100 prosumers by Qiu et al. [20], who used the system's demand-supply ratio as the mean-field input, and by Zheng et al. [21], who used public market clearing results. Alternative structures have also been explored. Qiu et al. introduced a multi-cluster approach with *parameter sharing* within homogeneous groups [22], demonstrating its capabilities in a case study of 300 households. To further enhance privacy, Ye et al. proposed a *federated learning framework* [23], where agents train models locally and only share model parameters—not their private data—with a central aggregator, showing strong results in a system of 250 households.

Beyond addressing scalability, a further challenge for practical deployment is the incorporation of physical distribution network constraints. A common approach is to translate network limits into the economic incentives of the agents via the reward function. For instance, Samende et al. proposed using *Distribution Network Tariffs* [24], derived from locational marginal pricing, to penalize or reward trades. Their MADDPG-based agents learned to respond to these price signals, successfully reducing peak line congestion by over 50%. Similarly, Feng et al. introduced a fictitious penalty to the reward for any voltage or line capacity violations [25], demonstrating that their consensus-based MARL algorithm could effectively learn to operate within these limits. Beyond individual constraints, Chen et al. tackled the broader challenge of coordinating heterogeneous agents by modeling interconnected residential, commercial, and industrial multi-energy microgrids, proposing a MATD3 approach to manage their diverse trading and energy conversion objectives [26].

Building on these foundations, recent research has begun to explore more sophisticated applications of MARL that expand its scope beyond simple energy trading. Several studies now integrate P2P trading with granular demand side management, where agents learn not only when to trade but also how to optimally schedule their internal resources. This includes co-optimizing the scheduling of electric vehicles and air

conditioners [27], coordinating a wide range of residential flexibilities like space heating and deferrable loads [28], and even creating hierarchical structures where a central agent learns to set community prices to guide the local scheduling decisions of other agents [29]. Furthermore, researchers are investigating novel algorithmic paradigms, such as model-based MARL that leverages forecasting to improve policy learning [30] and multi-step bidding procedures designed for aggregator-led markets [31]. The growing breadth of these applications highlights the immense potential of MARL to create truly autonomous, efficient, and intelligent LEMs.

It is noteworthy that much of the reviewed literature utilizes some form of MADDPG or similar *off-policy*, deterministic actor-critic algorithms. This preference is driven by two main factors. First, these methods are naturally suited for the continuous action spaces found in energy management problems, overcoming the limitations of earlier, discretization-based approaches [18, 23]. Second, as off-policy algorithms, their ability to reuse data from a *replay buffer* leads to a perception of greater sample efficiency [17]. However, in our preliminary investigations, MADDPG exhibited high sensitivity to hyperparameter tuning and suffered from poor exploration. This often caused the policy to collapse into greedy, suboptimal strategies, such as persistently discharging the battery over the day without ever learning to charge. In contrast, MAPPO, a stochastic *on-policy* algorithm, was found to be a powerful alternative despite its relative underutilization in the P2P energy trading literature. These findings, detailed in Appendix B, motivate our choice of MAPPO for this work.

2.3. Shapley Value

In many scenarios, agents can achieve a greater collective outcome by forming cooperative coalitions rather than competing. The *Shapley value* [9] is a solution concept from cooperative game theory that provides a principled method for fairly distributing the total gains generated by their cooperation. Consider a set of N agents. The outcome of any coalition, $C \subseteq N$, is quantified by a value function $v(C)$, with $v(\emptyset) = 0$. The additional value an agent i brings when joining this coalition is its *marginal contribution*, calculated as $(v(C \cup \{i\}) - v(C))$. The Shapley value captures the average marginal contribution of agent i with respect to all possible orderings of agents joining the empty coalition to form the grand coalition N. The formal definition $Sh_i(v, N)$, provides an efficient way to calculate this without needing to explicitly enumerate all permutations:

$$Sh_i(v, N) = \sum_{C \subseteq N \setminus \{i\}} \frac{|C|!(|N| - |C| - 1)!}{|N|!} \left(v(C \cup \{i\}) - v(C)\right) \tag{2.3}$$

While Shapley values have long been recognized for their fairness in cooperative game-theoretic settings [32], their application in LEMs remains limited. In a prominent

study, Raja et al. introduced coalition-based P2P trading frameworks where Shapley allocations ensured equitable benefit distribution among participants [33]. Addressing the computational burden of exact calculations, Han et al. proposed scalable methods to approximate Shapley values for P2P energy sharing in larger communities [34]. Extending these concepts into the RL domain, Wang et al. developed Shapley-based *credit assignment* mechanisms within multi-agent systems [35], including applications to energy networks. However, these approaches typically employ Shapley as an external settlement mechanism or an auxiliary tool, rather than embedding it directly into the agents' learning and interaction process. This creates a gap that motivates approaches, such as the one developed in this work, where Shapley-based pooling is integrated more tightly with MARL dynamics. Yet, the practical application of Shapley allocation is not without challenges: under forecast uncertainty, it can disadvantage certain agents, in particular flexibility-providing prosumers (e.g., with storage or shiftable loads), whose uncertain contributions may be undervalued compared to more predictable participants [36]. To avoid this complication, in our work, we assume perfect knowledge, where each agent knows its exact demand and PV generation when acting, keeping the focus on methodological integration of Shapley values into MARL; the detailed mechanism, along with a numerical example, is provided in Section 4.2.2.

3. Ausgrid Dataset

The experiments in this thesis are conducted using a real-world dataset of residential load and rooftop solar PV generation, originally collected and made public by the Australian distribution utility, Ausgrid[1]. The dataset[2] contains separately metered load and PV generation profiles for 300 de-identified households in New South Wales, Australia. These profiles were recorded at a 30-minute resolution over a three-year period from 1 July 2010 to 30 June 2013. In their foundational work introducing this dataset, Ratnam et al. [19] performed a conservative cleaning process to identify and remove customers with anomalous records, such as those arising from inverter failures or prolonged absences. This resulted in a refined "clean dataset" consisting of 54 customers with reliable data for the entire three-year span. Utilizing this real-world data is crucial for our study, as it provides a realistic and challenging environment that captures the inherent variability and stochasticity of residential energy profiles—features that are essential for training and validating robust MARL agents.

The remainder of this chapter details the experimental context for the study. We first describe the ToU[3] and FiT[4] pricing schemes employed by Ausgrid. This is followed by an overview of the preprocessing applied to the raw data to prepare it for the MARL environment. Finally, the chapter concludes with an exploratory data analysis of the specific agent profiles selected for this work.

3.1. ToU and FiT Pricing

The economic environment for the agents in this study is defined by two key components. First, the Time-of-Use (ToU) tariff [37] is the price at which households buy (import) electricity from the grid. This price varies throughout the day, typically being higher during peak demand periods. Second, the Feed-in Tariff (FiT) [37] is the price, often fixed, at which prosumers can sell (export) their surplus energy, such as from PV generation, back to the grid. In the context of a P2P market, these two grid prices establish the economic boundaries, or "price corridor", for local trading. A rational buyer in the P2P market will not pay more than the current ToU price, and a rational

[1] https://www.ausgrid.com.au/

[2] https://github.com/pierre-haessig/ausgrid-solar-data

[3] https://www.ausgrid.com.au/Your-Energy-Use/Understanding-tariffs/Time-of-use-pricing

[4] https://www.ausgrid.com.au/Your-Energy-Use/Getting-the-most-out-of-your-solar

seller will not accept less than the FiT price. Furthermore, any energy that remains unmatched at the end of a DA auction period is settled with the main grid at these respective ToU and FiT rates.

The exact historical ToU and FiT for the 2010–2013 period of the Ausgrid dataset, to the best of our knowledge, is not publicly available. Therefore, this study adopts the pricing scheme from the work of Qiu et al. [22], which is consistent with Ausgrid's Pricing Proposal for 2018-19[5]. The scheme is based on three daily demand periods: off-peak, shoulder, and peak hours. The time slots for these periods vary seasonally and by day of the week, with Table 3.1 providing distinct definitions for summer weekdays, winter weekdays, and all other times. The introduction of this third category for weekends and other months is a novel methodological choice for this study. It differs from the scheme in [22], which did not specify this separate pricing, and simplifies the official Ausgrid proposal, which also includes public holidays in this category. The corresponding ToU prices are summarized in Table 3.2, and the daily ToU price structure is visualized in Figure 3.1. The FiT is fixed at 0.04$ per kWh for all days and times in the dataset.

Table 3.1. ToU Demand Periods

Demand Period	**Summer Weekdays** (November to March)	**Winter Weekdays** (June to August)	**Weekends (Sat & Sun) Other Months**
Off-Peak	22:00-7:00 (next day)	22:00–7:00 (next day)	22:00–7:00 (next day)
Shoulder	07:00-14:00, 20:00-22:00	07:00-17:00, 21:00-22:00	07:00-22:00
Peak	14:00-20:00	17:00-21:00	—

Table 3.2. Grid Prices in $/kWh

Demand Period	**ToU**	**FiT**
Off-Peak	0.15	0.04
Shoulder	0.22	0.04
Peak	0.44	0.04

[5] https://www.aer.gov.au/industry/networks/pricing-proposals/ausgrid-annual-pricing-2018-19

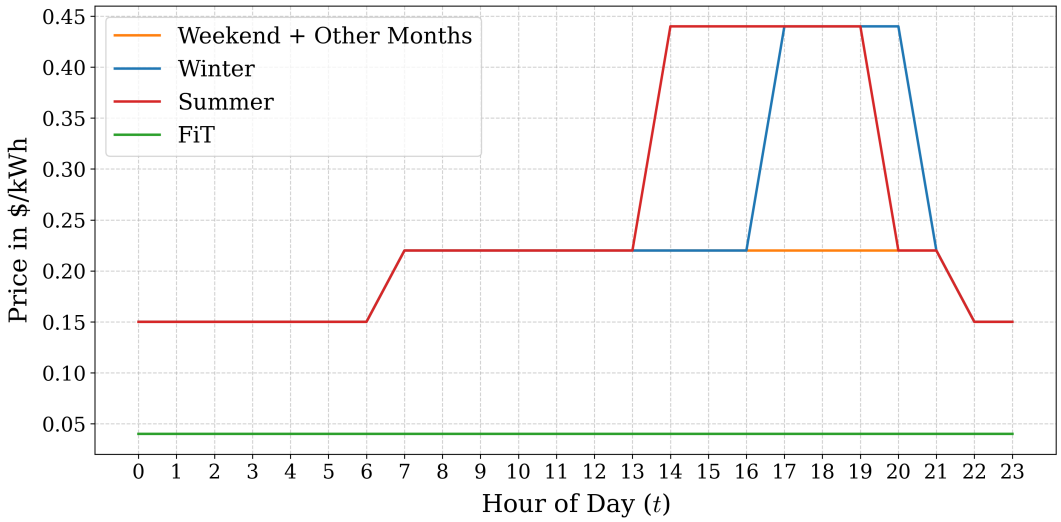


Figure 3.1. ToU Pricing

3.2. Data Preprocessing

To align with the hourly decision-making interval of our MARL environment, the raw 30-minute data is downsampled to an hourly resolution by summing the values of two consecutive half-hour periods. Furthermore, to ensure stable training of the actor and critic networks, the load and PV data for each agent are normalized independently using a Min-Max scaler. Since the load and PV values are non-negative, this scaling method effectively divides each agent's time-series data by their individual maximum observed value, mapping all inputs to a consistent range of [0, 1].

We use a stratified sampling approach to partition the data into train and test sets, ensuring that seasonal and weekly patterns are preserved. Specifically, one day is randomly sampled from each week in the three-year period to form the test set. This results in a test set of 157 days and a training set comprising the remaining 939 days. To ensure a balanced representation of weekdays, a random seed of 5 was used for this selection process. The resulting day-of-the-week distribution for the test set is summarized in Table 3.3.

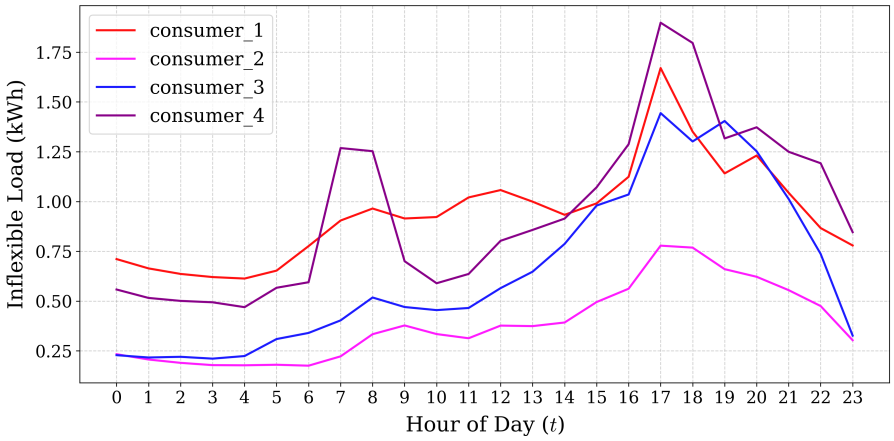
Table 3.3. Test Set Distribution

Day	# Test Days
Monday	25
Tuesday	21
Wednesday	19
Thursday	24
Friday	24
Saturday	25
Sunday	19

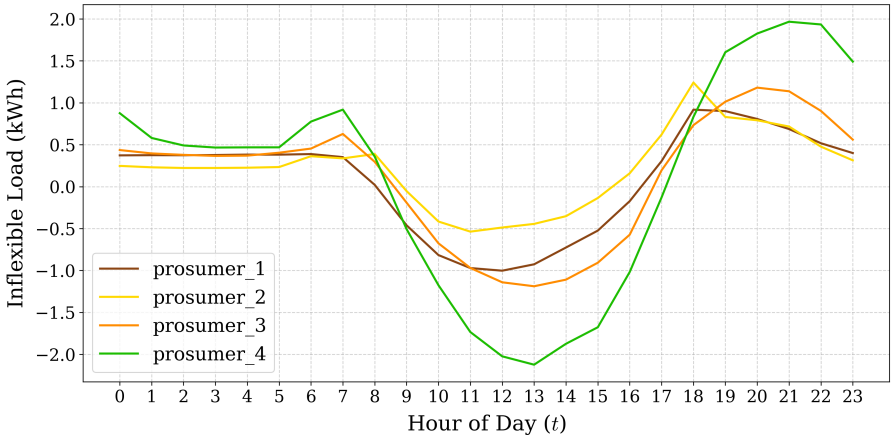
3.3. Agent Profiles

From the 54 households in the "clean" Ausgrid dataset, we manually select eight agents to form the experimental community. This small-scale LEM balances realism with tractability: it provides sufficient diversity in demand and PV generation patterns to reflect real residential behavior, while keeping MARL training manageable and avoiding the need for approximate Shapley computations. The selected set consists of four pure consumers and four prosumers with rooftop PV, chosen to represent a broad spread of load magnitudes, daily consumption patterns, and PV capacities. This configuration reflects a balanced mix of household types commonly found in residential energy communities.

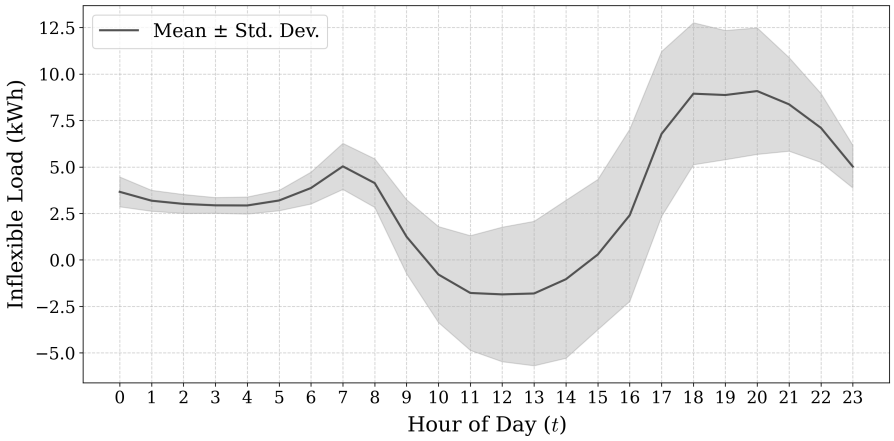
Table 3.4 reports the average daily load (inflexible demand) for all agents, alongside PV generation for the prosumers. To highlight seasonal variability in consumption behavior, Figure 3.2, Figure 3.3, and Figure 3.4 visualize representative hourly profiles for summer weekdays, winter weekdays, and weekends/other months, respectively. These profiles are computed as mean values across the full three-year dataset (1096 days). In the figures, we plot the inflexible load as formally defined later in Section 4.1. For consumers, this corresponds directly to their household load, while for prosumers, it is their load offset by PV generation. Accordingly, positive values indicate net demand, whereas negative values represent net supply. At the community level, profiles are obtained by aggregating the corresponding quantities across all eight agents. This characterization provides the empirical foundation for the MARL experiments in Chapter 5 and their evaluation in Chapter 6.



(a) Consumers

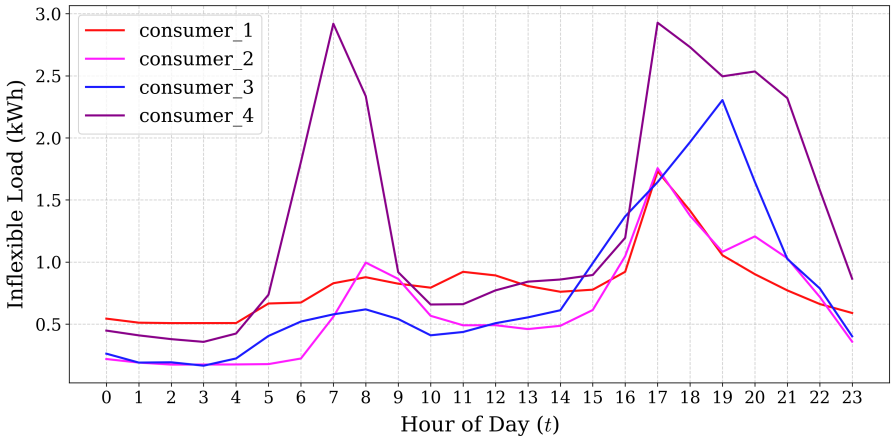


(b) Prosumers

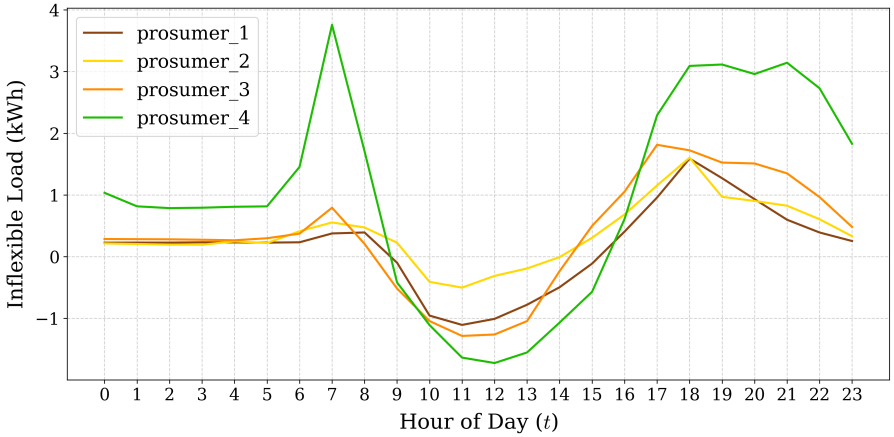


(c) Community

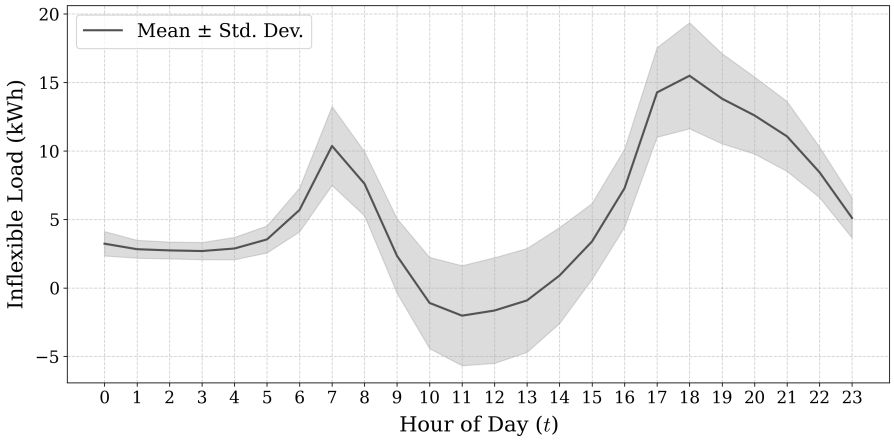
Figure 3.2. Inflexible Load Profiles for Summer Months



(a) Consumers

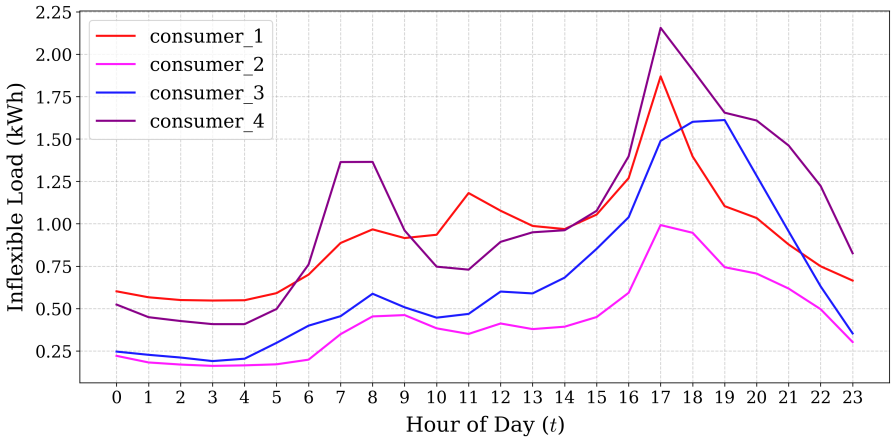


(b) Prosumers

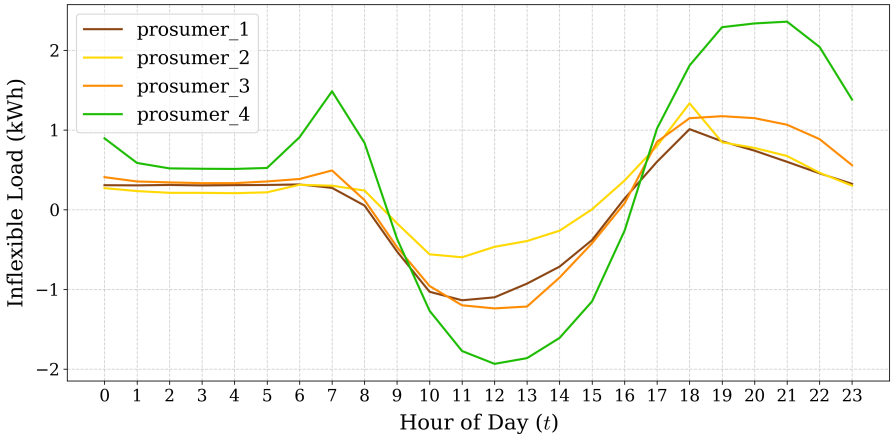


(c) Community

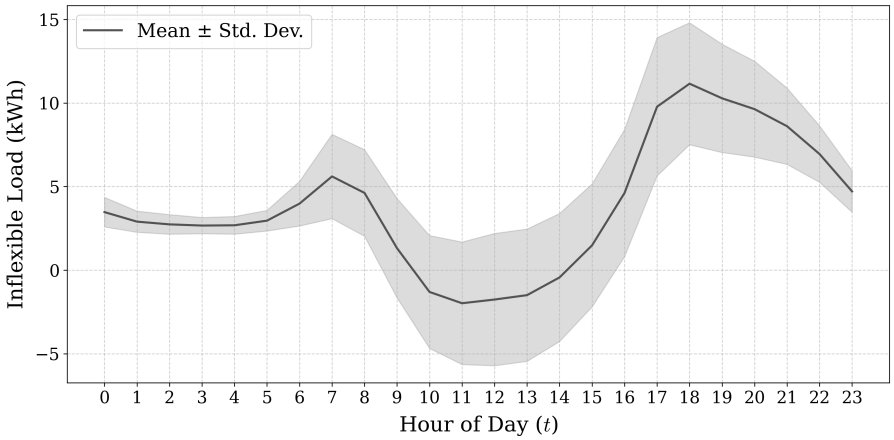
Figure 3.3. Inflexible Load Profiles for Winter Months



(a) Consumers



(b) Prosumers



(c) Community

Figure 3.4. Inflexible Load Profiles for Weekends and Other Months

Table 3.4. Average Daily Load and PV Generation

Agent	**Residential Load (kWh)**	**Rooftop PV (kWh)**
consumer_1	21.74 ± 5.87	—
consumer_2	10.92 ± 4.76	—
consumer_3	16.25 ± 3.87	—
consumer_4	25.47 ± 8.31	—
prosumer_1	12.24 ± 5.73	10.18 ± 4.47
prosumer_2	13.39 ± 4.62	7.48 ± 3.29
prosumer_3	17.20 ± 5.91	12.95 ± 5.27
prosumer_4	28.87 ± 10.39	18.69 ± 7.60
Community	146.09 ± 32.17	49.30 ± 19.68

4. Methodology

Having established the theoretical background and described the dataset, we now turn to the methodological framework of the study. We begin by defining the problem setting and formally specifying the MARL formulation. Next, we outline the three market mechanisms under investigation: a competitive DA market, a cooperative pooling mechanism with Shapley value allocation, and a hybrid game that combines both through contracts. The chapter concludes with a discussion of the evaluation metrics employed to assess learning performance and P2P market outcomes.

4.1. MARL Setup

We consider an LEM composed of $N = 8$ agents, four of which are consumers (demand-only) and four prosumers (with both demand and PV generation). This community is connected to a central utility that offers electricity at ToU prices and compensates surplus generation via a fixed FiT. We assume an unlimited grid supply and feed-in capacity from this external utility. The ToU and FiT tariffs, and the agent profiles used are presented in Section 3.1 and Section 3.3, respectively. Alongside the grid, agents can also trade energy locally via a P2P market, with the goal of reducing energy costs and external grid reliance. Each agent is equipped with an ES system, providing temporal flexibility; the identical operational parameters are adopted from [18] and are summarized in Table 4.1. The MARL environment is structured around a daily horizon of $T = 24$ hourly intervals, indexed as $t \in \{0, \ldots, 23\}$. At each timestep, agents must decide how much energy to charge or discharge from their ES and at what price to participate in the P2P market. The greedy objective for each agent is to minimize its own total energy costs over the day, with any cooperative behavior emerging implicitly from the decentralized learning dynamics. We now formally describe this setup within the framework of a MARL problem.

Observation Space

At each time step t, agent n receives a private observation $o_{n,t} \in \mathcal{O}_n$, defined as:

$$o_{n,t} = \left\{ P^{\mathrm{d}}_{n,t},\ P^{\mathrm{pv}}_{n,t},\ E^{\mathrm{es}}_{n,t},\ \lambda^{b}_{t},\ \lambda^{s}_{t},\ \sin\left(\frac{2\pi t}{T}\right),\ \cos\left(\frac{2\pi t}{T}\right),\ \mathrm{weekend}_t,\ \mathrm{summer}_t,\ \mathrm{winter}_t \right\} \tag{4.1}$$

where each component is described below:

- $P^{\mathrm{d}}_{n,t}$: Inflexible electricity demand
- $P^{\mathrm{pv}}_{n,t}$: PV generation; zero for consumers
- $E^{\mathrm{es}}_{n,t}$: ES energy level, or State of Charge (SoC)
- λ^{b}_{t}: Grid buy price (ToU)
- λ^{s}_{t}: Grid sell price (FiT)
- $\sin\left(\frac{2\pi t}{T}\right)$, $\cos\left(\frac{2\pi t}{T}\right)$: Positional encodings that help agents infer the time of day, where $T = 24$ denotes the episode length.
- $\mathrm{weekend}_t \in \{0, 1\}$: Boolean indicating whether the current day is a weekend (Saturday or Sunday).
- $\mathrm{summer}_t \in \{0, 1\}$, $\mathrm{winter}_t \in \{0, 1\}$: Booleans indicating whether the current day falls in summer or winter months, respectively.

The inflexible demand $P^{\mathrm{d}}_{n,t}$ and PV generation $P^{\mathrm{pv}}_{n,t}$ are derived from the Ausgrid dataset and are normalized prior to input into both actor and critic networks, as described in Section 3.2. Similarly, the SoC observation $E^{\mathrm{es}}_{n,t}$ is normalized to the range $[0, 1]$. And while seasonal and weekly patterns are indirectly encoded in the ToU price λ^{b}_{t}, these cannot be reliably inferred by agents until later in the day (post 2 PM), as pricing overlaps across ToU categories (see Section 3.1). Therefore, we explicitly provide these context flags to improve temporal awareness.

The global state s_t represents the full observable environment configuration and is used as input to the centralized critic in MAPPO. While a standard formulation would concatenate the private observations $o_{1:N,t}$ of all agents, we avoid repetition of global features—such as price signals λ^{b}_{t} and λ^{s}_{t}, time encoding, and seasonal context flags —by including them only once in the shared state. The state vector thus focuses on the agent-level components: $P^{\mathrm{d}}_{n,t}$, $P^{\mathrm{pv}}_{n,t}$, and $E^{\mathrm{es}}_{n,t}$ for all $n \in N$.

Table 4.1. ES Operating Parameters

Parameter	Description	Value
$\overline{P}^{\mathrm{es}}$	Power capacity (kW)	2
$\underline{E}^{\mathrm{es}}$	Minimum energy level (kWh)	0
$\overline{E}^{\mathrm{es}}$	Maximum energy level (kWh)	8
η^{esc}	Charging efficiency	0.95
η^{esd}	Discharging efficiency	0.95
E^{es}_{0}	Initial energy level (kWh)	$\mathcal{N}(4, 1.33^2, 0, 8)$

Action Space

At each time step t, agent n selects an action $a_{n,t} \in \mathcal{A}_n$, defined as:

$$a_{n,t} = \{a^p_{n,t},\ a^q_{n,t}\} \tag{4.2}$$

where:

- $a^p_{n,t} \in [0,1]$ denotes the agent's price decision: the price at which it is willing to trade in the DA market, expressed as a normalized position between λ^s_t and λ^b_t. The actual submitted price is calculated as:

$$p_{n,t} = \lambda^s_t + a^p_{n,t} \cdot \left(\lambda^b_t - \lambda^s_t\right) \tag{4.3}$$

- $a^q_{n,t} \in [-1,1]$ denotes the agent's ES scheduling decision. It controls the magnitude of intended charge (positive) or discharge (negative) as a ratio of ES power capacity $\overline{P}^{\text{es}}_n$.

State Transition

The environment evolves according to the stochastic transition function $\mathcal{T}$, which maps the current state s_t and joint action vector $a_{1:N,t}$ to the next state s_{t+1}. The dynamics are influenced both by agent actions and exogenous signals $\omega_t = \{P^{\text{d}}_{n,t}, P^{\text{pv}}_{n,t}, \lambda^b_t, \lambda^s_t\}$, which evolve independently of the agents' behavior. These exogenous variables capture real-world uncertainty arising from household consumption patterns, solar variability, and external tariff schemes. The resulting dynamics are well-suited to MARL, which learns directly from environment interaction, without relying on explicit system models.

On the other hand, the transition for endogenous features $E^{\text{es}}_{n,t}$ is determined by the agents' ES scheduling actions $a^q_{n,t}$. Let $C^{\text{es}}_{n,t}$ and $D^{\text{es}}_{n,t}$ denote the charging and discharging power of the ES, respectively. Since the ES cannot charge and discharge simultaneously at a given step t, these values are mutually exclusive, with one of them set to zero depending on the sign of $a^q_{n,t}$. Limited by SoC bounds $[\underline{E}^{\text{es}}_n, \overline{E}^{\text{es}}_n]$ and charge/discharge efficiencies η^{esc} and η^{esd}, they are calculated as:

$$C^{\text{es}}_{n,t} = \begin{cases} \min\left(a^q_{n,t}\,\overline{P}^{\text{es}}_n,\ \dfrac{\overline{E}^{\text{es}}_n - E^{\text{es}}_{n,t}}{\eta^{\text{esc}}\,\Delta t}\right), & a^q_{n,t} \geq 0 \\ 0, & a^q_{n,t} < 0 \end{cases} \tag{4.4}$$

$$D^{\text{es}}_{n,t} = \begin{cases} \max\left(a^q_{n,t}\,\overline{P}^{\text{es}}_n,\ \dfrac{(\underline{E}^{\text{es}}_n - E^{\text{es}}_{n,t})\cdot\eta^{\text{esd}}}{\Delta t}\right), & a^q_{n,t} < 0 \\ 0, & a^q_{n,t} \geq 0 \end{cases} \tag{4.5}$$

with $\Delta t = 1$ hour. The SoC then evolves as:

$$E^{\text{es}}_{n,t+1} = E^{\text{es}}_{n,t} + C^{\text{es}}_{n,t}\,\Delta t\,\eta^{\text{esc}} + D^{\text{es}}_{n,t}\,\Delta t\,/\,\eta^{\text{esd}} \tag{4.6}$$

Consequently, the net quantity submitted to the P2P market $q_{n,t}$ is given by:

$$q_{n,t} = (P^{\text{inf}}_{n,t} + C^{\text{es}}_{n,t} + D^{\text{es}}_{n,t}) \cdot \Delta t \tag{4.7}$$

This quantity (energy) reflects the agent's attempt to satisfy both its desired charging or discharging action and its residential load at time t. Here $P^{\text{inf}}_{n,t}$ denotes the inflexible load, defined as:

$$P^{\text{inf}}_{n,t} = \begin{cases} P^{\text{d}}_{n,t}, & n \text{ is consumer} \\ P^{\text{d}}_{n,t} - P^{\text{pv}}_{n,t}, & n \text{ is prosumer} \end{cases}$$

For prosumers, the demand is first offset against their PV generation $P^{\text{pv}}_{n,t}$, and only the residual demand (or surplus, if positive) is carried to the market. This design choice reflects the behavior of rational prosumers, who would first prioritize self-consumption.

Reward Function

At each time step t, agent n receives a reward $r_{n,t}$. In this formulation, negative values indicate expenditures (costs), while positive values indicate revenues (gains). This setup models the agent's money flow, which directly reflects its economic objectives and ensures the interpretability of the results. Since the reward calculation depends on the specific market mechanism, its detailed design is presented in the subsequent sections. For clarity, we use superscripts to distinguish between mechanisms: $r^{\text{da}}_{n,t}$ for DA trading, $r^{\text{pool}}_{n,t}$ for Shapley pooling, and $r^{\text{mix}}_{n,t}$ for the hybrid mechanism introduced later.

4.2. P2P Market Dynamics

The outcome of market clearing at each time step t is determined by the pair $(q_{n,t}, p_{n,t})$, capturing the net quantity and price submitted by each agent. To reiterate the convention, positive and negative values of $q_{n,t}$ are interpreted as follows:

$$q_{n,t} \geq 0 \;\Rightarrow\; \text{buy (net demand)}, \qquad q_{n,t} < 0 \;\Rightarrow\; \text{sell (net supply)}$$

4.2.1. Competition: DA Trading

In the competitive setting, the DA market is employed for local energy trading. At each time step t, the auctioneer collects all submitted orders $(q_{n,t}, p_{n,t})$ from the agents and clears the market using Algorithm 1. The outcome of the clearing process is given by:

$$\left(\lambda_{n,t}^{\text{da}},\ q_{n,t}^{\text{da}},\ q_{n,t}^{\text{grid}}\right)$$

where $\lambda_{n,t}^{\text{da}}$ denotes the local trading price, $q_{n,t}^{\text{da}}$ the quantity matched in the DA market, and $q_{n,t}^{\text{grid}}$ the unmatched residual. This residual energy is settled with the grid at price λ_t^b or λ_t^s, depending on whether the agent is a buyer or seller. Accordingly, the instantaneous reward of agent n at time step t is computed as:

$$r_{n,t}^{\text{da}} = -(\lambda_{n,t}^{\text{da}} \cdot q_{n,t}^{\text{da}} + \lambda_t^b \cdot \max(0, q_{n,t}^{\text{grid}}) + \lambda_t^s \cdot \min(0, q_{n,t}^{\text{grid}})) \tag{4.8}$$

where the second and third terms represent, respectively, expenditures from residual grid purchases and revenues from residual grid sales. It is worth noting that this DA market design follows the distributed P2P paradigm illustrated in Figure 1.2, involving auctioneer-mediated bilateral trades and individual residual settlements with the grid.

Numerical Example

For a simple illustration, let us consider a community with three agents. At time step t, the FiT is $\lambda_t^s = 1$ \$/kWh and the ToU price is $\lambda_t^b = 5$ \$/kWh. The submitted quantities and corresponding prices calculated using Equation (4.3) are:

- **Alice (A)**: $q_{A,t} = -3$ kWh (seller); $\quad a_{A,t}^p = 0.5 \ \Rightarrow \ p_{A,t} = 3.0$ \$/kWh
- **Bob (B)**: $q_{B,t} = 2$ kWh (buyer); $\quad a_{B,t}^p = 0.7 \ \Rightarrow \ p_{B,t} = 3.8$ \$/kWh
- **Dave (D)**: $q_{D,t} = 4$ kWh (buyer); $\quad a_{D,t}^p = 0.6 \ \Rightarrow \ p_{D,t} = 3.4$ \$/kWh

The market is cleared with the following transactions:

1. **Trade 1:** A sells 2 kWh to B at 3.4 \$/kWh, as B offers the higher bid price.
2. **Trade 2:** A sells her remaining 1 kWh to D at 3.2 \$/kWh.
3. **Trade 3:** D buys the remaining 3 kWh from the grid at $\lambda_t^b = 5$ \$/kWh.

Thus, A's full supply is matched, while B's demand is fully met locally and D partly resorts to the grid. The corresponding market outcome for each agent, expressed as the tuple $(\lambda_{n,t}^{\text{da}}, q_{n,t}^{\text{da}}, q_{n,t}^{\text{grid}})$, is:

$$\text{A: } (3.33, -3, 0), \quad \text{B: } (3.4, 2, 0), \quad \text{D: } (3.2, 1, 3)$$

The reward vector, calculated using Equation (4.8) is therefore:

$$(r_{A,t}^{\text{da}}, r_{B,t}^{\text{da}}, r_{D,t}^{\text{da}}) = (10.0\ \$, -6.8\ \$, -18.2\ \$)$$

4.2.2. Cooperation: Shapley Pooling

The DA market, though efficient in principle, remains highly sensitive to the bidding strategies ($a^p_{n,t}$) of participating agents. Even more so in a MARL context, where suboptimal bids may lead to missed trades and, consequently, missed opportunities for cost reduction. To address this limitation, we introduce a pooling-based mechanism in which the collective benefit of an LEM is fairly allocated among participants using Shapley values. This approach moves in the direction of a more centralized P2P market [6] (see Figure 1.2), where a coordinating entity facilitates transactions and dispatch control (note that, unlike the figure, agents control their own ES and load). Since the focus of this thesis lies on the economic dimension, we abstract away from network-level considerations and assume that such centralized coordination is technically feasible.

The core idea is to consider coalitions of collaborative agents. Within such a coalition, agents pool their individual net quantities $q_{n,t}$, and only the aggregate surplus or deficit is traded with the external grid. In effect, the coalition acts as a single unified agent with respect to the grid. Formally, the pooled quantity for a coalition $C \subseteq N$ at time t is defined as:

$$q_{C,t} = \sum_{n \in C} q_{n,t} \tag{4.9}$$

The total reward of coalition C at time t is then obtained as a simplification of Equation (4.8), namely:

$$r^{\text{pool}}_{C,t} = -(\lambda^b_t \cdot \max(0, q_{C,t}) + \lambda^s_t \cdot \min(0, q_{C,t})) \tag{4.10}$$

In particular, we focus on the grand coalition $C = N$, where all agents of the community cooperate. The resulting pooled quantity is denoted $q_{ec,t}$ (stylized "ec" for energy community), with total reward $r^{\text{pool}}_{ec,t}$. To distribute this communal cost (or benefit) among agents, we employ the Shapley value, as described in Section 2.3. The value function v of a coalition C is defined by its reward, i.e.,

$$v(C) = r^{\text{pool}}_{C,t} \tag{4.11}$$

The payoff allocated to agent n under this mechanism is thus given by:

$$r^{\text{pool}}_{n,t} = \text{Sh}_n(v, N) \tag{4.12}$$

where $\text{Sh}_n(v, N)$ denotes the Shapley value of agent n in the grand coalition N.

The benefits of adopting this pooling-based mechanism are threefold:

- It eliminates the need for individual pricing strategies, such that the agent's action space in Equation (4.2) reduces to $a_{n,t} = a^q_{n,t}$, thereby simplifying the MARL learning problem.
- By internally balancing surpluses and deficits before trading with the grid, the mechanism yields outcomes that are optimal from a community-wide perspective, though the results remain contingent on effective ES scheduling.
- It promotes fairness by allocating rewards in proportion to the provided liquidity, ensuring that prosumers are appropriately compensated.

Numerical Example

We now revisit the same example introduced in Section 4.2.1. The pooled net quantity of the grand coalition at time t is:

$$q_{ec,t} = q_{A,t} + q_{B,t} + q_{D,t} = -3 + 2 + 4 = 3 \text{ kWh}$$

indicating a net demand of 3 kWh. The corresponding coalition reward, using Equation (4.10), is:

$$r^{\text{pool}}_{ec,t} = -\lambda^b_t \cdot q_{ec,t} = -5 \cdot 3 = -15 \text{ \$}$$

This collective reward is then distributed fairly among the agents according to the Shapley value, as defined in Equation (4.12). The resulting reward vector is:

$$(r^{\text{pool}}_{A,t},\ r^{\text{pool}}_{B,t},\ r^{\text{pool}}_{D,t}) = (10.33 \text{ \$}, -8.67 \text{ \$}, -16.67 \text{ \$})$$

where the exact allocations follow from the marginal contributions over all possible coalitions. For clarity, this calculation is detailed through permutations in Table 4.2. The averages sum to the coalition reward $r^{\text{pool}}_{ec,t} = -15\,\$$, as required by Shapley efficiency.

It is also worth noting the following limiting cases of the pooling mechanism:

- If $q_{n,t} > 0$ for all agents (all buyers) or $q_{n,t} < 0$ for all agents (all sellers), the mechanism reduces to direct trading with the grid.
- If the pool is perfectly balanced, i.e. $q^{\text{pool}}_{ec,t} = 0$, the outcome is equivalent to all agents trading in the DA market at $a^p_{n,t} = 0.5$.

In either case, the pooling mechanism remains centralized; all energy flows through the coordinating entity.

Table 4.2. Marginal Contributions by Permutation for Shapley Pooling

Permutation	A	B	D
ABD	3	−2	−16
ADB	3	−10	−8
BAD	11	−10	−16
BDA	15	−10	−20
DAB	15	−10	−20
DBA	15	−10	−20
Average (Shapley)	10.33 $	−8.67 $	−16.67 $

4.2.3. Hybrid Contracting Game

While pooling ensures fairness at the community level, individual agents may still prefer competitive trading in certain situations. Agents with the ability to place extreme bids—for instance, a buyer bidding close to λ_t^b (e.g., $a_t^p = 0.9$) or a seller asking near λ_t^s (e.g., $a_t^p = 0.1$)—may secure matches at prices only marginally more favorable than ToU or FiT. Yet these bids remain profitable and can strategically ensure that the agent captures all available trades for itself under DA, instead of sharing them through pooling. In the earlier example, Bob would benefit more from DA, even though pooling distributes Alice's surplus more evenly across Bob and Dave, which is fairer from a community perspective. This tension between collective fairness and individual self-interest motivates the exploration of mixed cooperative–competitive games, where agents not only trade energy but also negotiate which market mechanism governs their P2P interactions.

We achieve this hybrid mechanism through the introduction of contracts, which allow agents to pre-commit a fraction of their net position to pooling while reserving the remainder for competitive DA trading. Formally, each agent's contract at time t is denoted by $\kappa_{n,t} \in [0, 1]$, representing the share of its net quantity $q_{n,t}$ allocated to the cooperative pool. The residual fraction $(1 - \kappa_{n,t})$ is then submitted to the DA market along with a price strategy $a_{n,t}^p$. While $\kappa_{n,t}$ does not depend on stochastic variables such as demand, PV generation, or ES state, it is implicitly shaped by the prevailing tariff regime (ToU and FiT) through seasonal and temporal indicators:

$$\kappa_{n,t} = f(\text{summer}_t, \text{winter}_t, \text{weekend}_t, t) \tag{4.13}$$

These contracts behave analogously to economic agreements: they are fixed and publicly known beforehand, allowing all agents to plan their strategies accordingly.

Determining the optimal contracts (and thereby selecting the collective market mechanism) constitutes an additional layer of learning, which is introduced later. For the present formulation, however, we assume that contracts $\kappa_{n,t}$ are exogenously given. The inclusion of contracts in the MARL environment (creating what we refer to as the "Hybrid Base Environment") leads to the following modifications:

- **Observation Space:** The private observation $o_{n,t}$ as defined in Equation (4.1) is extended as:

$$o_{n,t} \leftarrow o_{n,t} \cup \{\kappa_{n,t},\ \bar{\kappa}_t,\ \sigma_{\kappa,t}\}$$

  where $\bar{\kappa}_t$ and $\sigma_{\kappa,t}$ denote the mean and standard deviation of the contracts across all agents, respectively. In addition to its own contract $\kappa_{n,t}$, each agent observes these aggregate statistics rather than the full vector $\kappa_{1:N,t}$, which may introduce spurious variance. This design summarizes the overall community cooperation level and dispersion while keeping the observation space compact.

- **State Space:** The global state s_t is expanded to include the full set of contracts:

$$s_t \leftarrow s_t \cup \{\kappa_{1:N,t}\}$$

  ensuring the centralized critic accesses the full contract profile, unlike the aggregated view in private observations.

- **Action Space:** The action definition remains identical to DA trading, i.e. Equation (4.2).

- **Reward Function:** Each agent's net position is partitioned as:

$$q_{n,t}^{\text{pool}} = \kappa_{n,t}\, q_{n,t}, \qquad q_{n,t}^{\text{trade}} = (1 - \kappa_{n,t})\, q_{n,t} \tag{4.14}$$

  The pooled part $q_{n,t}^{\text{pool}}$ enters the cooperative mechanism and determines $r_{n,t}^{\text{pool}}$ via Equation (4.12), while the residual $q_{n,t}^{\text{trade}}$ enters the DA clearing and determines $r_{n,t}^{\text{da}}$ via Equation (4.8). The total reward is simply:

$$r_{n,t}^{\text{mix}} = r_{n,t}^{\text{pool}} + r_{n,t}^{\text{da}} \tag{4.15}$$

  Intuitively, $\kappa_{n,t}$ controls the degree of cooperation: $\kappa_{n,t} = 1$ recovers full pooling, while $\kappa_{n,t} = 0$ reduces to pure DA trading.

- During training, contracts $\boldsymbol{\kappa}$ are randomly initialized at the beginning of each episode as:

$$\kappa_{n,t} \sim \mathcal{U}(0,1), \qquad \forall n \in N,\ \forall t \in \{0,\ldots,23\} \tag{4.16}$$

Contract Proposal Subgame

Since contracts are fixed, the problem of selecting optimal contracts may at first glance resemble a multi-agent variant of the classical multi-armed bandit problem [11]. Unlike stochastic signals such as demand, PV generation, or ES state, which the Hybrid Base Environment is already trained to handle, contracts act as a stable design knob. Agents must therefore identify a "one-size-fits-all" solution by tuning this knob based on the rewards they obtain from this pre-trained environment (optimized across a range of randomly sampled contracts). However, a closer inspection reveals several complications. First, proposing an entire contract profile—spanning 24 hours and multiple ToU categories—leads to an explosion in dimensionality and fails to capture temporal dependencies, since a contract at time t inevitably affects outcomes later in the horizon. Second, rewards are inherently non-stationary: they depend not only on an agent's own contract but also on the contracts of others, which are being optimized in parallel. While this could be partially mitigated by collecting trajectories under fixed contracts and updating in an on-policy style, the environment still exhibits drifting incentives as contracts evolve. Third, contract selection is not merely a matter of choosing from a fixed action set: agents require contextual information to guide their proposals, leading to a circular dependency where contracts both define and depend on context. These factors render the contract proposal problem closer to a MARL setup, which is inherently more suited to handling temporal structure, strategic interaction, and non-stationarity.

We follow a setup inspired by Multi-Objective Contract Augmentation Learning (MOCA), as proposed by Haupt et al. [10]. The key idea is to decouple operational policy learning from contract selection by introducing a secondary MARL environment layered on top of the hybrid market formulation. The procedure unfolds in two stages:

Phase 1 (Hybrid Base Environment): Contracts are treated as exogenous and sampled randomly. Training across this distribution of contracts is akin to *multi-objective learning*, since each contract effectively defines a different market mechanism. This results in a base MARL environment that is robust across a wide range of regimes.

Phase 2 (Contract Proposal Environment): The second-stage MARL environment differs from the base environment in two key aspects:

- **Observation Space:** The contracting observation is derived from Equation (4.1) by excluding the stochastic components:

$$o^{\kappa}_{n,t} = o_{n,t} \setminus \{P^{\mathrm{d}}_{n,t},\, P^{\mathrm{pv}}_{n,t},\, E^{\mathrm{es}}_{n,t}\} \tag{4.17}$$

- **Action Space:** The action consists of proposing a contract share at time t:

$$a^{\kappa}_{n,t} = \kappa_{n,t} \tag{4.18}$$

Once proposed, these contracts are passed to the pre-trained and frozen base environment, which returns the resulting rewards. This stage thus serves as a form of *meta-learning* [38]: agents learn policies for contract selection by observing how their proposals influence outcomes in the underlying trading game. We refer to this two-stage design—comprising the Hybrid Base and Contract Proposal environments—as the Hybrid Contracting Game (abbreviated as Mix in subsequent figures and results). This represents the hybrid P2P mechanism that we compare to the DA and Shapley Pooling markets.

It is important to note that the contracts employed in this thesis differ from the zero-sum contracts introduced by Haupt et al. [10]. As a result, the theoretical guarantees established in their work do not directly apply here. Instead, we draw inspiration primarily from their two-stage learning framework and the use of contracts as a means to structure agent interaction within a MARL setting.

4.3. Evaluation Metrics

We keep the evaluation straightforward, focusing on metrics that directly capture the economic objectives of an LEM. The primary goals are to minimize energy costs and reduce external grid reliance over the course of a day. The key metrics considered are:

- **Daily Reward:** Obtained by aggregating hourly rewards across the day,

$$r_n = \sum_{t=0}^{23} r_{n,t}, \qquad \forall n \in N \tag{4.19}$$

  and can be generalized to the energy community as a whole,

$$r_{ec} = \sum_{n \in N} r_n \tag{4.20}$$

  This is the most important metric, as it directly reflects the MARL reward structure and captures the core economic outcome for both agents and the community. Higher values indicate better performance, corresponding to lower costs and greater revenues.

- **Grid Reliance:** Grid interactions are predominantly in the form of costly imports, as most local surpluses are either absorbed within the P2P market or stored in ES. To measure reliance, we compute the fraction of an agent's net position (considering only positive demand) that must be settled with the external grid. For DA, it is defined as:

$$G_n^{\text{da}} = \frac{\sum_t \max(0, q_{n,t}^{\text{grid}})}{\sum_t \max(0, q_{n,t})} \tag{4.21}$$

This scaling makes the metric comparable across agents and highlights how effectively P2P trading reduces external dependence. Lower values indicate stronger local integration and, implicitly, higher volumes of P2P trades. For Shapley pooling, however, the situation is different. Since the community interacts with the grid as a single entity, individual grid imports are not explicitly tracked—only the aggregate community import is observed. Accordingly, reliance can only be computed at the community level as:

$$G_{ec}^{\text{pool}} = \frac{\sum_t \max(0, q_{ec,t})}{\sum_t \sum_n \max(0, q_{n,t})} \tag{4.22}$$

Similarly, community-wide reliance for DA:

$$G_{ec}^{\text{da}} = \frac{\sum_t \sum_n \max(0, q_{n,t}^{\text{grid}})}{\sum_t \sum_n \max(0, q_{n,t})} \tag{4.23}$$

- **Fairness:** To assess whether the benefits of P2P trading are distributed evenly among participants, we consider a baseline MARL environment without a P2P market, trained using Independent PPO (IPPO) [14]. The baseline reward for agent n at time t reduces to pure grid interaction:

$$r_{n,t}^{\text{grid}} = -\left(\lambda_t^b \cdot \max(0, q_{n,t}) + \lambda_t^s \cdot \min(0, q_{n,t})\right) \tag{4.24}$$

We measure each agent's absolute improvement as:

$$\Delta r_n = r_n - r_n^{\text{grid}} \tag{4.25}$$

where higher values indicate greater savings relative to the grid-only baseline. Although rare, negative values indicate worse performance compared to the baseline. The Δr_n values thus not only assess the overall economic benefit of introducing the P2P market, but their spread across agents also captures fairness in benefit allocation. A more detailed discussion follows in Section 6.3.

5. Results

This chapter presents the core results of our experiments. We begin with a brief description of the experimental setup, including the framework, implementation details, and evaluation protocol. We then report on the training performance, focusing on learning stability, convergence behavior, and robustness across random seeds. Finally, we evaluate the trained policies on the held-out test set, analyzing their effectiveness in reducing energy costs, before proceeding to a more detailed discussion in Chapter 6.

5.1. Experimental Setup

All experiments were implemented using BenchMARL [39], which builds on TorchRL and PettingZoo to provide a standardized framework for MARL. The full codebase, including environments, training scripts, and configuration files, is available at a public GitHub Repository[1]. Since much of the existing literature is not open-source, reproducing published results is often difficult. Moreover, the outcomes we observed during this work differ noticeably from the expectations set in prior studies. To provide a realistic view of the performance achieved under the stated conditions, we present honest results without cherry-picking, ensuring that our findings remain transparent and reproducible.

All experiments were conducted on local hardware consisting of an AMD Ryzen 7 5800H CPU, an NVIDIA GeForce RTX 3060 Laptop GPU, and 16 GB of RAM. Given the relatively modest runtime of our experiments (at most a few hours per run), this hardware proved sufficient for training and evaluation, making large-scale computing resources unnecessary. We use the MARL setup as defined in Section 4.1, with the train–test split introduced earlier in Section 3.2. To assess robustness, all experiments were repeated with multiple random seeds, while the hyperparameters used for training are detailed in Appendix A.

[1] `https://github.com/rohanrao619/multi-agent_power-systems`

5.2. Training Performance

We trained each market mechanism with 8 random seeds (0–7), where seeds were used to initialize the ES states (E_0^{es}) and to sample training days. Contracts were initialized separately using an independent random generator. This setup ensures that performance differences arise solely from the market mechanism itself. The training curves for the four mechanisms—Grid Only Baseline (Grid), DA Market (DA), Shapley Pooling (Pool), and Hybrid Contracting Game (Mix) (first-stage only)—are shown in Figure 5.1. We report the episodic community-wide cost over the course of training, averaged across seeds with standard deviation as a shaded region. This cost is defined as the negative episodic reward ($-r_{ec}$), a transformation applied for readability. The x-axis denotes MAPPO collection steps, with each step corresponding to 8,192 environment interactions. Training proceeds for approximately 0.52 million steps (64 collections), with the full set of MAPPO hyperparameters listed in Table A.1. The results show that introducing a P2P market significantly reduces costs relative to the Grid baseline. All models learn rapidly in the early phase and then gradually converge, with Pool exhibiting marginally better performance, as expected given its guarantee of full internal matching.

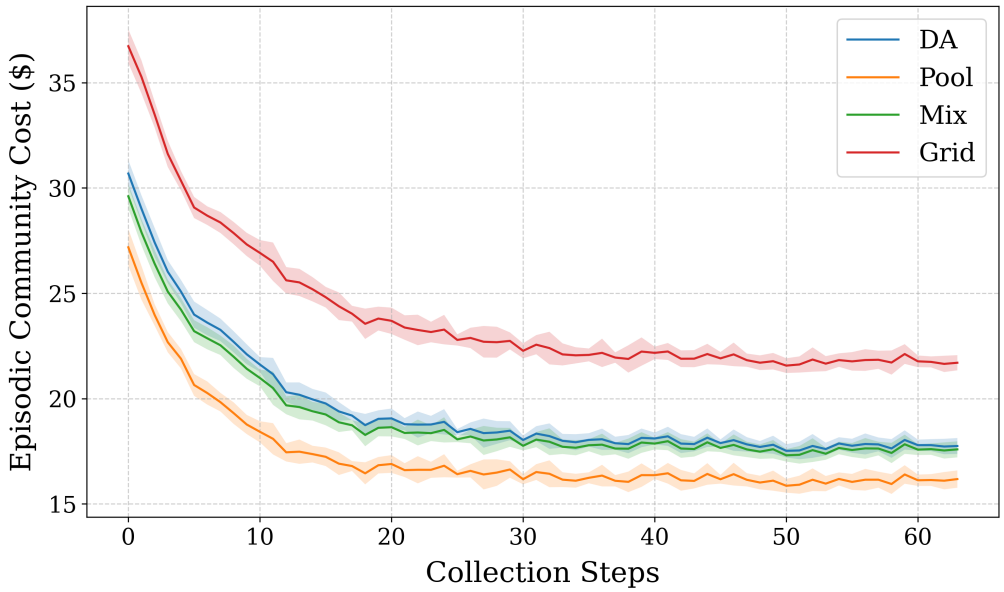


Figure 5.1. Training Curves for different Mechanisms

In parallel with training, we evaluate the models at the end of each collection step using deterministic actions. This provides a consistent measure of policy quality during training, independent of exploration noise. The evaluation spans 256 episodes, covering a representative portion of the training set, and is fixed by setting the random seed to 42. This ensures that evaluation is performed on the same set of days and initial ES states across all models, steps, and runs, providing a consistent basis for comparison. The

evaluation curves, shown in Figure 5.2, indicate that variance decreases toward the end of training, reflecting increasing confidence in the learned policies. However, differences between the three P2P markets are less pronounced. Mix shows signs of lagging behind toward the end of training, yet it still performs surprisingly well. This is noteworthy given that its contracts κ are randomly initialized, highlighting its ability to learn robust strategies in the presence of additional mechanism uncertainty.

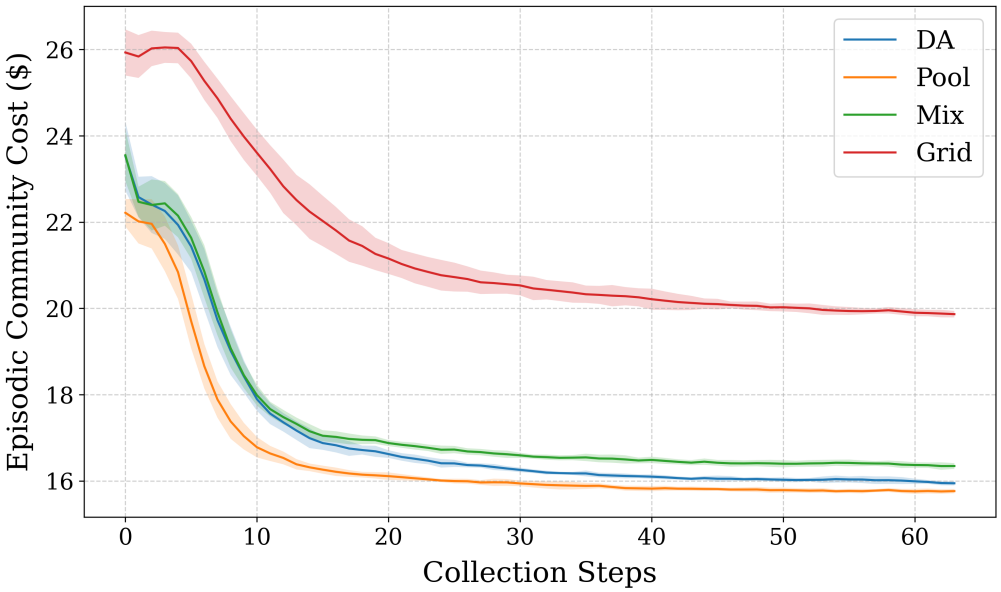


Figure 5.2. Evaluation Curves for different Mechanisms

The learning curves for the second-stage Contract Proposal Environment are shown in Figure 5.3. We build on the frozen policies from the base environment (Mix), selecting the best evaluation step for each of the eight training runs. The second-stage experiments are then run on top of these policies, using seeds 8–15 to ensure independent random trajectories from the first-stage runs. Training is carried out with a slightly relaxed set of hyperparameters (Table A.2), reflecting the relative simplicity of this environment compared to the base stage. Most of the optimization burden is already handled in the first stage, while the second stage primarily fine-tunes contract decisions. This explains the somewhat erratic appearance of the training curve, as convergence is reached very quickly and further updates mainly serve to refine the critic rather than drive substantive policy improvements. Nonetheless, the evaluation confirms that the learned policies achieve strong performance, surpassing Mix from the first stage when operating under optimal contracts.

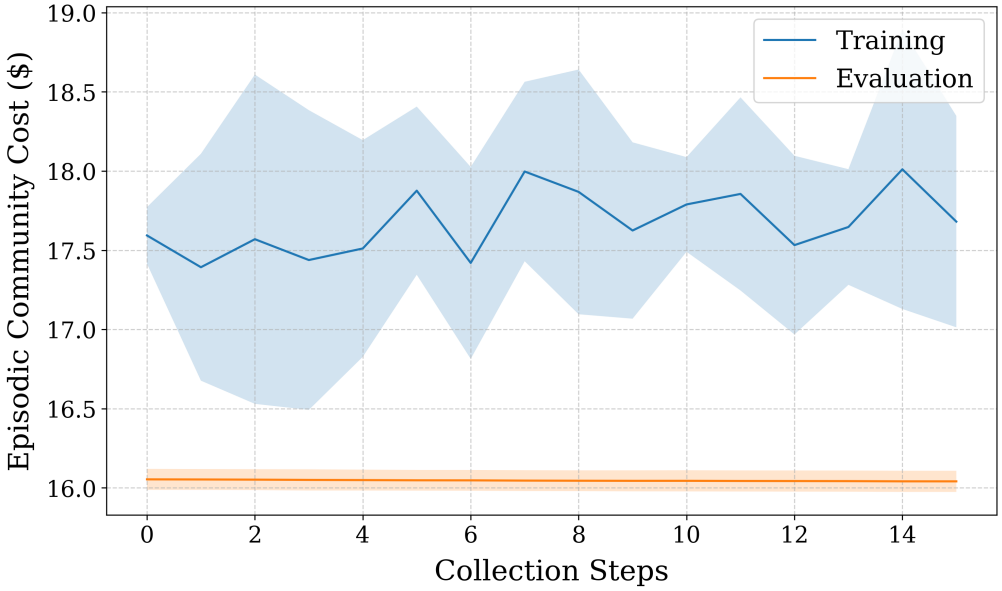


Figure 5.3. Learning Curves for second-stage Contract Proposal Environment

5.3. Test Performance

To benchmark performance under unseen conditions, we evaluate the trained models on the held-out test set. For each mechanism, we select the best epoch from all eight training seeds and run them across 157 test days, resulting in 157×8 evaluation scenarios with varied ES initializations (E_0^{es}). Randomness is controlled with a fixed seed (2025) to ensure consistency across mechanisms. For the Mix mechanism, testing is conducted using the optimal contracts produced by the second-stage policies, which will be analyzed in detail in Section 6.4. Results are reported as mean $\pm$ standard deviation across all scenarios, with the 95th percentile across scenarios (5% worst cases) shown in brackets to capture tail performance and robustness on unfavorable days.

Daily rewards for all agents and the community are presented in Table 5.1, while community-level grid reliance is reported in Table 5.2. Across the test set, all P2P markets substantially outperform the Grid baseline, reducing costs by roughly 21% on average and demonstrating improved robustness in worst-case outcomes. The differences between the three P2P markets are relatively small and warrant further investigation, but the results already suggest that each provides a significant improvement over relying solely on the grid. In terms of grid reliance, Pool achieves the lowest values, though again only marginally. Measured as a share of demand, this corresponds to about 25% being met internally through P2P trading rather than external purchases, yielding substantial savings at the system level.

Table 5.1. Daily Rewards under different Mechanisms (in $)

Agent	**Grid** (r^{grid})	**DA** (r^{da})	**Pool** (r^{pool})	**Mix** (r^{mix})
consumer_1	-3.53 ± 1.18 [-5.67]	-3.00 ± 1.10 [-4.94]	-3.03 ± 1.16 [-5.14]	-3.03 ± 1.12 [-5.07]
consumer_2	-1.75 ± 0.89 [-3.44]	-1.39 ± 0.88 [-3.15]	-1.36 ± 0.90 [-3.16]	-1.40 ± 0.90 [-3.18]
consumer_3	-2.59 ± 0.94 [-3.89]	-2.22 ± 0.86 [-3.51]	-2.24 ± 0.92 [-3.53]	-2.25 ± 0.90 [-3.55]
consumer_4	-4.71 ± 2.05 [-9.00]	-4.11 ± 1.90 [-7.82]	-4.12 ± 1.97 [-7.77]	-4.13 ± 1.91 [-7.71]
prosumer_1	-0.73 ± 1.10 [-2.50]	-0.28 ± 1.17 [-2.13]	-0.14 ± 1.14 [-1.95]	-0.22 ± 1.18 [-2.10]
prosumer_2	-1.17 ± 0.90 [-2.68]	-0.78 ± 0.93 [-2.41]	-0.68 ± 0.98 [-2.35]	-0.74 ± 0.93 [-2.35]
prosumer_3	-1.27 ± 1.60 [-4.69]	-0.85 ± 1.58 [-3.90]	-0.76 ± 1.58 [-3.79]	-0.86 ± 1.58 [-3.96]
prosumer_4	-2.76 ± 2.59 [-7.90]	-2.01 ± 2.60 [-6.87]	-2.09 ± 2.62 [-7.30]	-2.12 ± 2.61 [-7.19]
Community (r_{ec})	-18.53 ± 7.87 [-33.11]	-14.64 ± 7.85 [-28.09]	-14.43 ± 7.91 [-28.28]	-14.74 ± 7.86 [-28.30]

Table 5.2. Community-level Grid Reliance under different Mechanisms

Setup	**Grid Reliance** (G_{ec})
Grid	1 ± 0 [1]
DA	0.77 ± 0.10 [0.93]
Pool	0.75 ± 0.11 [0.92]
Mix	0.77 ± 0.10 [0.92]

6. Discussion

Having established the main results, we now turn to a more detailed discussion aimed at disentangling the behavior of the three P2P markets, which appeared broadly similar at first glance. To better understand their differences, we examine several salient aspects. First, we analyze how agents utilize their ES, providing insight into the role of flexibility in reducing costs. Second, we investigate the matching dynamics in DA to assess whether pricing strategies introduce inefficiencies. Third, we consider fairness, evaluating how evenly the benefits of P2P trading are distributed across participants. Finally, we take a closer look at the optimal contracts identified in the proposal stage of the Mix mechanism to shed light on the preferences agents express when given the ability to formalize agreements.

All analyses in this chapter are conducted on the test dataset, using the evaluation scenarios defined in Section 5.3. Unless stated otherwise, we report mean values across all these scenarios. As a reminder, results for Mix are generated by its first-stage policies operating under the optimal contracts learned in the second stage.

6.1. ES Scheduling Behavior

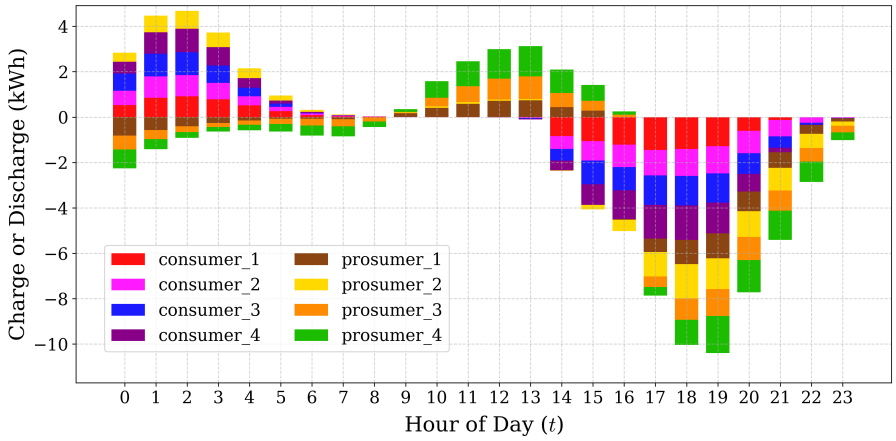
We analyze the ES scheduling behavior of all agents using bar plots, which show for each hour of the day how much the ES is charged (positive values) or discharged (negative values). Figure 6.1 shows the patterns under the Grid-only mechanism across different ToU categories, while Figure 6.2 illustrates the hybrid mechanism. Since the overall behavior of the three P2P markets was similar, we restrict the illustration to Mix, which represents a balanced case between the two extremes, resembling a convex combination of DA and Pool.

Several consistent patterns emerge. Agents learn to charge their ES early in the day and discharge during peak hours, as defined by the ToU tariffs (to reduce costly grid imports). This is a notable achievement, since a single policy adapts optimally to both winter and summer price profiles. On weekends, where no peak period exists, agents instead discharge in response to higher demand hours. Consumers typically charge overnight, taking advantage of cheap off-peak grid prices, whereas prosumers prefer midday charging (roughly between 10 AM to 4 PM), coinciding with peak PV generation. The most striking difference between Grid and Mix lies in this midday

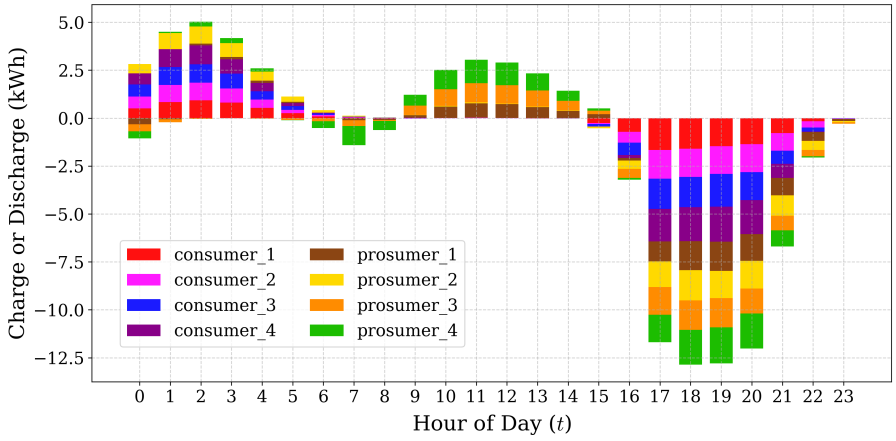
charging behavior: under Mix, prosumers charge notably less, suggesting that excess PV at these hours is instead matched directly to consumers to satisfy their demand via P2P trades. This reallocation is a major source of savings in P2P markets. To compensate, prosumers join consumers in overnight charging, albeit to a smaller extent.

For consumers under Mix, one might also expect midday charging. After all, P2P prices in the shoulder period are typically lower than off-peak grid imports. However, as shown in Figure 3.2c, Figure 3.3c, and Figure 3.4c, the community net load (supply) during these hours is insufficient to cover the charging needs of all consumers—especially since prosumers must also satisfy their own household demand and charge their ES. Consumers, therefore, rely on the safer, guaranteed overnight option. Another difference, which contributes to additional savings in Mix, is the earlier discharge observed for prosumers in summer and weekend profiles. With abundant midday PV supply, prosumers under Grid-only can strategically discharge earlier in the day to meet their own demand or sell to the grid to capture modest profits, and then fully recharge their ES before the peak period. This strategy works well when PV generation later replenishes the ES but can backfire on low-production days, potentially leading to costly grid imports. In Mix, this early discharge behavior is less pronounced, but when it does occur, it benefits consumers by providing extra low-priced supply to either charge their ES or meet direct demand, at prices favorable to both parties.

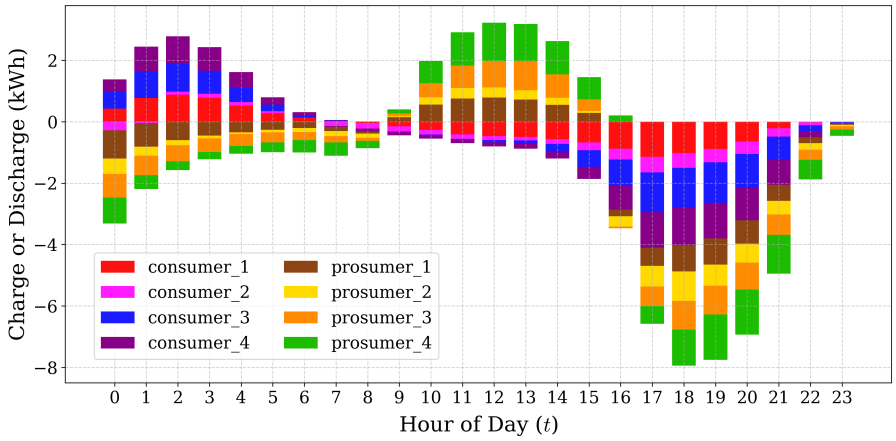
It is important to note that the observed behavior directly results from the ES scheduling action $a^q_{n,t}$ subject to the SoC constraints, which we also refer to as SoC actions for clarity, as this terminology is used in the plots. These actions under different mechanisms are presented in Figure 6.3. For this comparison, we report the mean across all pricing profiles, since the early-day behavior—which is our focus here—is largely consistent across categories; the main differences appear in peak-hour discharging, which depends on the specific tariff. While the overall patterns look similar, it is noteworthy that consumers also exhibit a willingness to charge at midday (most evident in Pool). However, in practice, their ES is typically already full by this time. Another observation is that agents act in greater harmony under Pool, which stems from the mechanism's inherent fairness. This fosters more stable scheduling decisions, as agents rely less on precautionary imports and instead place greater trust in the system and in each other.



(a) Summer Weekdays

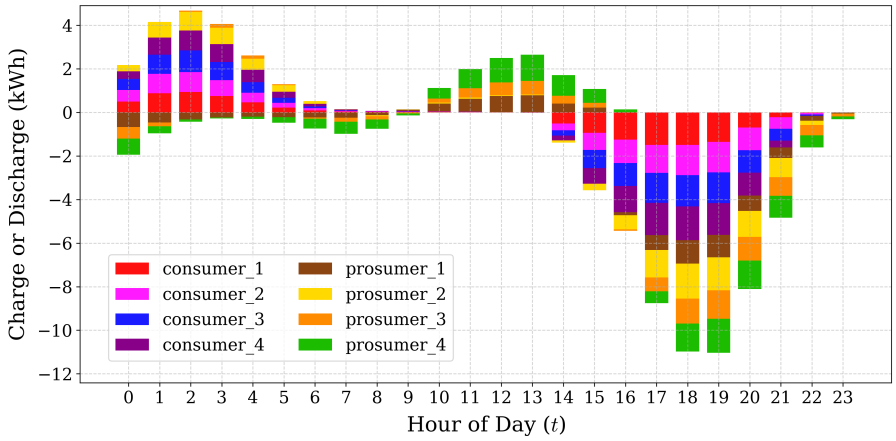


(b) Winter Weekdays

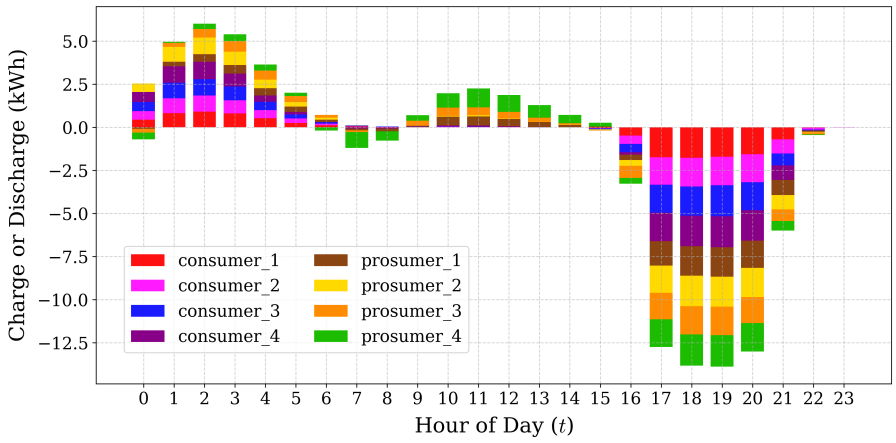


(c) Weekends and Other Months

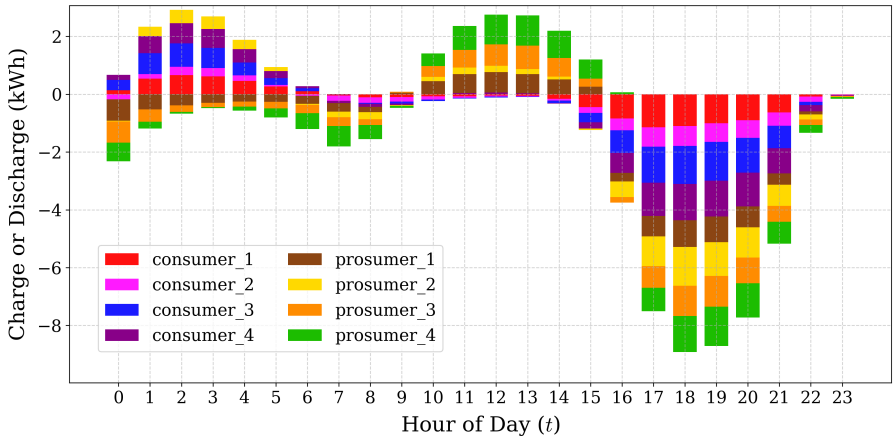
Figure 6.1. ES Scheduling under Grid-Only Mechanism



(a) Summer Weekdays

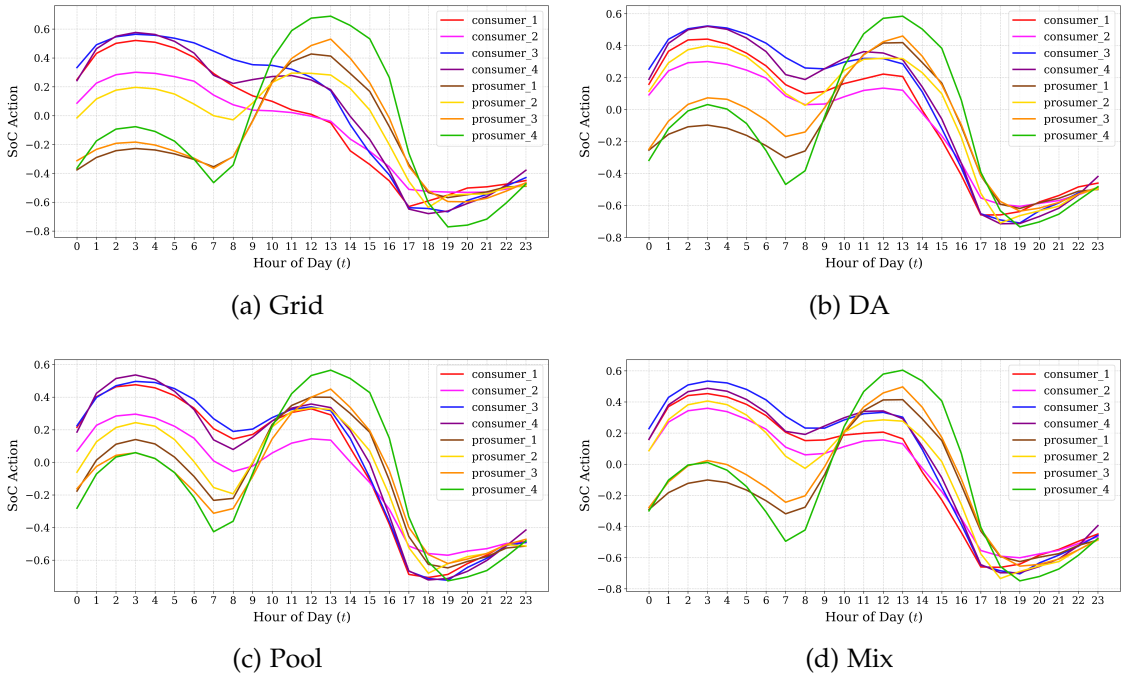


(b) Winter Weekdays



(c) Weekends and Other Months

Figure 6.2. ES Scheduling under Hybrid Contracting Game



(a) Grid (b) DA

(c) Pool (d) Mix

Figure 6.3. SoC Actions ($a^q_{n,t}$) under different Mechanisms

6.2. Matching Efficiency in DA

Theoretically, the marginally better performance of Pooling can be attributed to the structural differences in how trades are matched. Unlike Pool, the DA market relies on the agents' ability to learn sensible price actions ($a^p_{n,t}$) in addition to optimally scheduling their ES. This dual learning task introduces complexity, and convergence to a competitive equilibrium is not guaranteed. Consequently, as noted earlier, some bids might remain unmatched due to incompatible price–quantity pairs (i.e., non-overlapping bid and ask prices rather than volume limitations). Upon analysis, we find that approximately 7.7% of DA rounds terminate with such partial matching, as a subset of potential trades fail to clear once buy and sell offers stop overlapping. To illustrate this effect, Figure 6.4 shows the hourly trade volumes for the three P2P markets. Pool consistently achieves higher volumes, benefiting from its guaranteed full internal matching; resulting in greater cost reductions. Reinforcing our earlier discussion, most P2P activity can be seen to occur around midday hours for all markets, aligning with peak PV generation and serving as the primary driver of cost savings compared to the Grid-only baseline. Interestingly, substantial trading volume is also observed during peak hours, when local surplus is typically scarce. This indicates that agents use their ES not only to meet their own demand but also to trade stored energy (excess, if any) with other community members through the market—a mutually beneficial outcome where sellers earn additional revenue while buyers reduce reliance on the grid during the most

expensive ToU period, highlighting the collective advantage of P2P trading.

To further examine the sensitivity of DA to pricing actions, we present the hourly price actions in Figure 6.5a. To complement this, agent-level grid reliance under DA is reported in Table 6.1, with results again summarized as mean $\pm$ standard deviation and the 95th-percentile (5% worst cases) shown in brackets. The results reveal that prosumers (primarily acting as sellers, though occasionally as buyers) and consumers (as buyers) tend to form distinct clusters around the clearing price—prosumers with lower asks and consumers with higher bids—thereby facilitating market matching, especially during midday hours. However, this also creates opportunities for individual agents to exploit the market: those able to deviate slightly from the cluster and position their bids or asks more aggressively (higher bids or lower asks) can secure a disproportionate share of the available volume. For instance, Consumer 1, who places higher bids around the afternoon, captures more energy and consequently achieves lower grid reliance, whereas Consumer 2, who bids closer to the fair mid-price (≈ 0.5), secures less and suffers significantly (high grid reliance). Although this effect is most evident among consumers—who compete for limited surplus energy—prosumers are not exempt. Prosumer 4, for example, smartly benefits by offering lower ask prices earlier in the day and later switching to higher bids (as buyer) during the evening peak to meet its own high demand. These dynamics underscore the inherently competitive nature of DA markets, where success depends on anticipating others' strategies and adapting bids accordingly.

Since the Hybrid Contracting Game also incorporates a DA market, we present the corresponding price actions in Figure 6.5b. However, its behavior differs notably from standard DA. The clusters of bids and asks appear less tightly knit, and the overall price spread is smaller. This suggests that, given agents are guaranteed some fair allocation through contracts, they can bid more freely without the same pressure to undercut or overbid competitors. The market remains competitive, but less aggressively so than in pure DA, where agents often compromise to stay ahead of others, leading to a wider price spread. This relaxation comes with a minor drawback, however—approximately 18.8% of trading rounds exhibit partial matching, which helps explain why Mix slightly lags in performance, as presented in Section 5.3. To further highlight these contrasting competitive dynamics, Figure 6.6 shows the community-wide average matching price (volume-weighted) for both DA and Mix (computed only over successfully matched trades). The price distribution in Mix is narrower and centered closer to 0.5, whereas DA exhibits a clear upward skew, favoring prosumers with higher selling prices. These pricing asymmetries hint at emerging fairness implications that warrant further investigation.

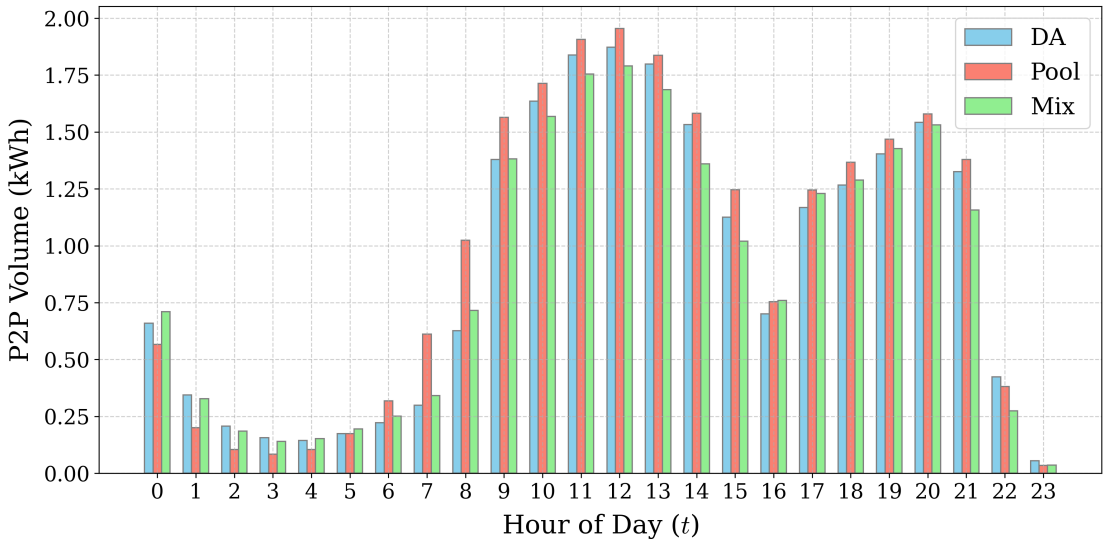
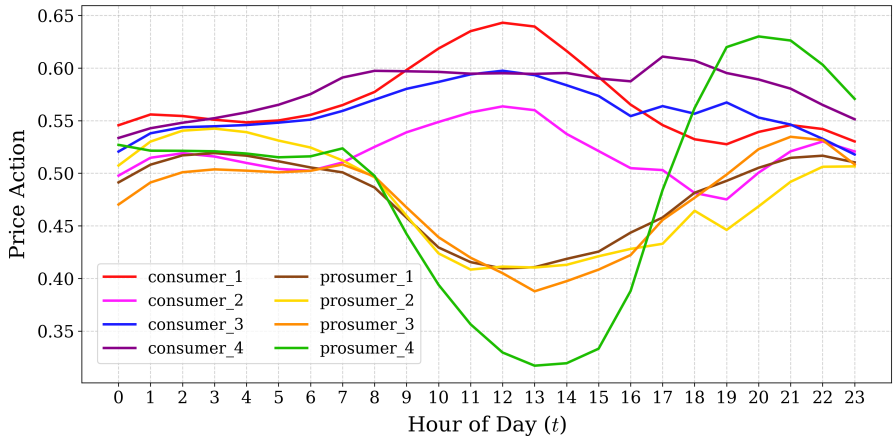


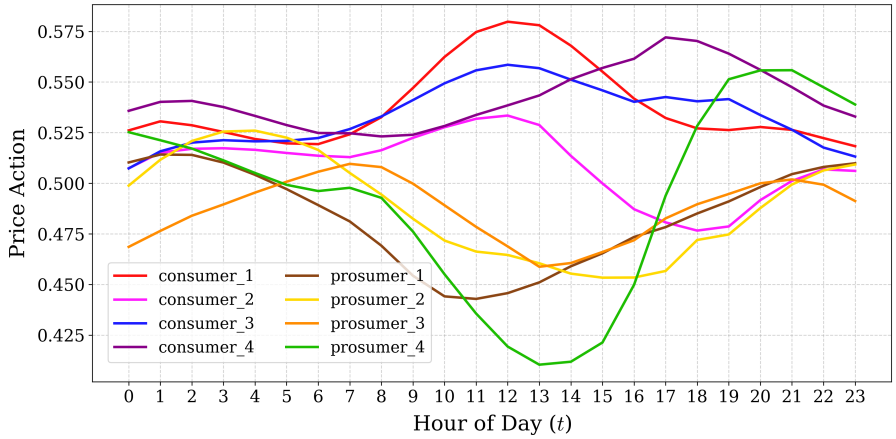
Figure 6.4. Local Trading Volume under different P2P Markets

Table 6.1. Grid Reliance under DA Market

Agent	**Grid Reliance** (G_n^{da})
consumer_1	0.68 ± 0.16 [0.94]
consumer_2	0.81 ± 0.18 [1.00]
consumer_3	0.70 ± 0.16 [0.96]
consumer_4	0.74 ± 0.15 [0.97]
prosumer_1	0.91 ± 0.17 [1.00]
prosumer_2	0.95 ± 0.10 [1.00]
prosumer_3	0.91 ± 0.14 [1.00]
prosumer_4	0.82 ± 0.15 [1.00]



(a) Double Auction (DA)



(b) Hybrid Contracting Game (Mix)

Figure 6.5. Hourly Price Actions ($a^{p}_{n,t}$) of the Agents

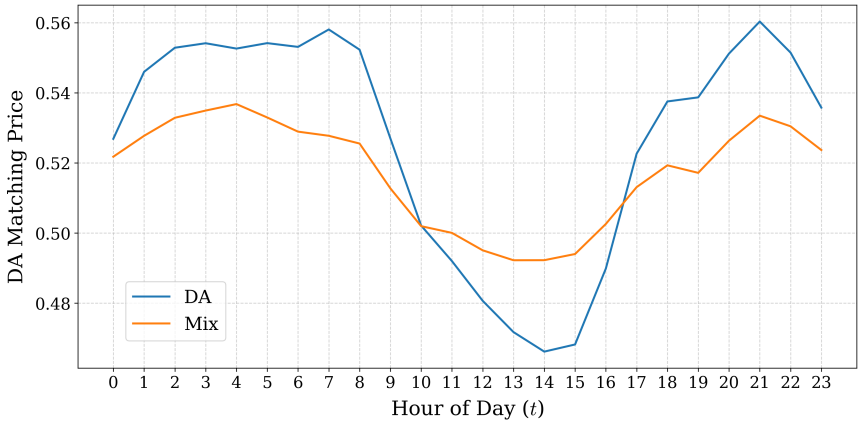


Figure 6.6. Community-level Average Matching Price in DA Markets

6.3. Fairness Analysis

While reducing costs at the community level is a key objective, an equally important question is how these savings are distributed among participants. From an aggregate perspective, different P2P markets may appear similar—since overall performance largely depends on how much energy is exchanged internally—but their fairness in benefit allocation can vary considerably. To examine this, we report each agent's savings relative to the Grid-only baseline (Δr_n, as described in Section 4.3) in Table 6.2, with tail performance shown in brackets. While ratio-based measures—such as percentage improvements over the baseline—would, in principle, account for varying agent sizes, they become unstable when r_n^{grid} is small or near zero; therefore, we use absolute savings for a more robust assessment. In addition, we compute the Gini coefficient [40] (over Δr_n across all eight agents) as a summary metric to quantify fairness, as shown in Table 6.3. A Gini value close to zero denotes perfect fairness, while values approaching one indicate increasing inequality in benefit distribution. Negative Δr_n are rare and indicate worse-than-baseline performance. To preserve interpretability, we exclude all scenarios from the Gini computation in which any agent has a negative Δr_n.

Pooling again stands out, yielding the highest community-level savings. These gains are also distributed more evenly across agents compared to the DA market. For instance, Prosumer 4—previously dominant in DA due to its strategic behavior (as discussed earlier)—appears more grounded in Pool, with other prosumers achieving comparable rewards. While Consumers 2 and 3 seem to benefit less, this is primarily a consequence of their smaller demand scales (see Table 3.4), rather than an indication of inequality. This balanced behavior is consistent with the marginally lower Gini values observed for Pool. Mix similarly exhibits a more even distribution than DA (closer to Pool), lagging slightly in savings only due to partial matching rounds. Overall, these subtle differences reinforce the contrast between the competitive nature of DA and the cooperative dynamics of Pool, further establishing Pool as an effective and fair market design.

Table 6.2. Relative Savings (Δr_n) for different P2P Markets (in $)

Agent	DA	Pool	Mix
consumer_1	0.53 ± 0.31 [0.05]	0.50 ± 0.31 [0.02]	0.50 ± 0.29 [0.04]
consumer_2	0.36 ± 0.28 [0.02]	0.39 ± 0.26 [0.09]	0.36 ± 0.25 [0.04]
consumer_3	0.37 ± 0.23 [0.03]	0.35 ± 0.24 [0.04]	0.34 ± 0.22 [0.03]
consumer_4	0.60 ± 0.45 [0.06]	0.59 ± 0.45 [0.08]	0.58 ± 0.44 [0.05]
prosumer_1	0.45 ± 0.32 [0.02]	0.59 ± 0.34 [0.10]	0.51 ± 0.32 [0.12]
prosumer_2	0.39 ± 0.27 [0.04]	0.49 ± 0.30 [0.13]	0.44 ± 0.28 [0.07]
prosumer_3	0.42 ± 0.36 [-0.06]	0.51 ± 0.34 [0.05]	0.41 ± 0.34 [-0.07]
prosumer_4	0.76 ± 0.41 [0.14]	0.68 ± 0.41 [0.14]	0.65 ± 0.39 [0.09]
Community	3.89 ± 1.51 [1.69]	4.10 ± 1.54 [1.76]	3.78 ± 1.51 [1.62]

Table 6.3. Gini Coefficient on Δr_n for different P2P Markets

P2P Market	Gini Coefficient
DA	0.280 ± 0.082 [0.427]
Pool	0.264 ± 0.083 [0.415]
Mix	0.267 ± 0.081 [0.417]

6.4. Optimal Contracts

A key advantage of excluding stochastic features from the observation space in the second-stage Contract Proposal Environment is that, after training, we can infer fixed optimal contracts ($\kappa^*_{n,t}$) for each ToU category (Summer Weekdays, Winter Weekdays, and Weekends + Other Months). These contracts are then applied in the base environment to enable joint **Mix** predictions (which is precisely how the reported results are obtained). Examining these optimal contracts provides insight into agents' preferences when given the freedom to choose between P2P market mechanisms. Since the learned contracts were found to be similar across ToU categories, we visualize their mean values in Figure 6.7. Based on earlier findings—where Pool performed best at the community level and for most agents—we might expect $\kappa^*_{n,t}$ to lean toward one. However, this is not what we observe: agents tend to remain cautious, converging around balanced 50–50 allocations. This behavior explains the hybrid nature of the Mix results and may reflect a strategic compromise between cooperative and competitive incentives. More notably, no consistent temporal pattern emerges across the day, suggesting that time-dependent contracting may add unnecessary complexity. While these findings offer valuable preliminary insights, further investigation is required to draw conclusive interpretations of contracting behavior. Nonetheless, this setup represents an early yet promising research direction, and potential improvements are discussed in Chapter 8.

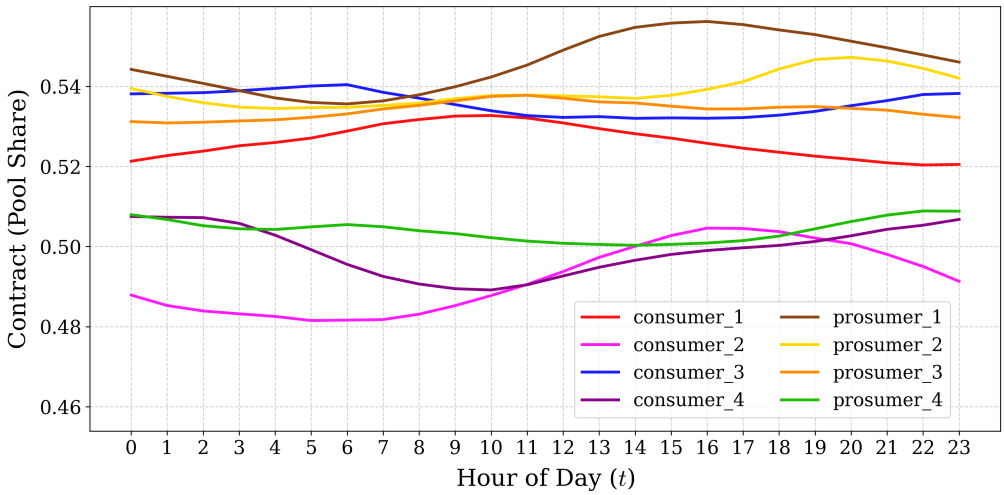


Figure 6.7. Optimal Contracts from second-stage Environment

7. Conclusion

This thesis developed autonomous agents for P2P trading in LEMs, addressing the need for efficient mechanisms to manage prosumer interactions. Our core objective was to leverage MARL to enable agents, each equipped with an ES system, to learn optimal scheduling and bidding strategies that minimize their daily energy costs and reduce external grid reliance, all validated on a real-world residential dataset (Ausgrid). To this end, we formulated the problem as a POMG and employed MAPPO as the core learning algorithm, which our preliminary investigation found to be significantly more robust than common off-policy alternatives. We designed and systematically compared a competitive DA market against two novel P2P markets developed in this work: (1) a cooperative mechanism based on Shapley Pooling (Pool) for fair cost allocation; and (2) a Hybrid Contracting Game (Mix). This hybrid market was realized through a two-stage contract-augmented MARL framework, where agents first learn robust operational policies under random contracts and subsequently learn to propose optimal contracts in a secondary meta-learning environment.

Our evaluation on the held-out test set confirms that agents in all setups learned intelligent ES scheduling strategies, rationally charging during cheap off-peak hours and discharging to shave peak demand during expensive ToU periods. The introduction of P2P markets, however, unlocked superior performance, reducing community-wide costs by roughly 21% and cutting external grid reliance by up to 25% compared to the Grid-only baseline. This was driven by key emergent behaviors: prosumers learned to sell their midday PV generation directly to consumers rather than storing it; and all agents learned to share their ES flexibility, trading excess stored energy during peak hours to mutually reduce the community's reliance on the grid. When comparing the P2P markets, the competitive DA market, while effective, suffered from structural inefficiencies. Its reliance on learned pricing actions led to partial matching rounds and created an exploitable environment where strategically aggressive agents could secure disproportionate gains. In contrast, the cooperative Pool mechanism consistently achieved higher trade volumes by guaranteeing full internal matching. This led to the highest relative community savings—approximately 5% more than DA —and critically, a more equitable distribution of benefits among the participants, as evidenced by its lower Gini coefficient. While the quantitative performance differences were subtle, Pool proved to be a more desirable market design, simplifying the learning task by removing price actions, ensuring community-optimal outcomes, and promoting a fairness that fosters greater trust in the system.

The hybrid market offered a more nuanced perspective; while it achieved good performance under its learned optimal contracts, it lagged marginally behind Pool and DA. This performance lag is explained by its unique market dynamics: the contract-guaranteed pool allocation reduced the pressure on agents to secure trades, leading them to bid less aggressively. This elicited a less competitive environment with a narrower price spread, but also resulted in a higher rate of partial matching rounds, impeding its overall efficiency. Despite this efficiency trade-off, the hybrid dynamic yielded a key benefit: a fairer distribution of savings than pure DA. This balanced outcome is a direct result of the agents' learned preferences, as the contact proposal subgame revealed they did not converge on expected full cooperation; instead, they learned cautious, balanced 50-50 contracts, reflecting a strategic compromise. While this contract-augmented framework is a promising avenue, the current setup is preliminary and requires further research to fully model complex hybrid P2P markets.

8. Limitations and Future Work

The limitations of this study define several key areas for improvement, which are essential for moving this work from a simulation-based study towards practical, real-world deployment. We hope that the methodologies and insights developed herein provide a valuable foundation for continued investigation in this domain.

8.1. Scalability, Privacy, and Realism

The primary limitations relate to scalability and real-world applicability. Our experiments were conducted on a small community of eight agents, and it remains unclear if the findings would scale to larger LEMs. The computational burden, in particular, could render these methods infeasible at a larger scale. Compounding this issue is the significant privacy challenge of the CTDE framework: its centralized critic requires access to all agents' private information (demand, PV, SoC), a setup that may not be acceptable to the community participants. Furthermore, the model's realism is limited by the assumption of perfect information, as agents operate with full knowledge of their immediate demand and generation rather than under the uncertainty of forecasts. Lastly, the model operates in a purely economic dimension, treating the LEM as an ideal market where any agent can trade with any other. This approach abstracts away from physical network constraints and presumes a central entity capable of managing not only fair cost allocation (as in the Shapley mechanism) but also the complex energy dispatch needed to respect network limits. Future work should aim to address these challenges. Techniques like mean-field approximation could alleviate the scalability and privacy bottlenecks of the centralized critic, while Graph Neural Networks (GNNs) could be explored to explicitly model the physical network topology and its constraints (like line congestion or voltage violations).

8.2. Contract Design and Learning

The proportional, community-wide contracts explored in this thesis are preliminary, and this design choice was a direct result of early, unsuccessful experiments with more complex contract forms. We initially attempted to implement traditional one-to-one contracts using a secondary DA market on top of the base DA market (for energy). In

this framework, agents would bid for multi-day contracts, with matching occurring (per ToU-period) to formalize the agreements. To ensure these proposals felt like binding agreements rather than dynamic hourly actions, this meta-environment operated over a longer time horizon. However, this approach fared poorly, as it drastically over-complicated the POMG formulation. This experience, which motivated our pivot to the simpler, pooling-based solutions, highlights a valuable insight. Future work must identify expressive, alternative contracting mechanisms that remain tractable for MARL agents to learn.

Beyond the relative (proportional) contracts, which reflect a commitment to the collective good, absolute (fixed-quantity) contracts may offer a more intuitive alternative, as they represent a firm commitment that is independent of an agent's real-time stochastic demand or PV generation. In this setup, an agent would commit to buying or selling a specific energy quantity at each timestep. This would, in theory, provide the actors with a clear a priori view of the exact volumes available in the pool, enabling more informed trading choices. However, our preliminary analysis revealed a significant flaw: this can create inefficient "market loops" due to over-commitment. An agent (buyer or seller) who over-commits could be forced to interact with the grid merely to honor the contract, such as buying a deficit from the grid to fulfill a sale obligation. If this contract-bound energy is successfully matched (in the contract pool), the agent incurs a financial loss, effectively subsidizing the trade for their counterpart's benefit. If, however, it remains unmatched, the agent must immediately resell it back to the grid, creating an unnecessary and loss-making transaction. While the agent's ES can help counteract these imbalances, it is not guaranteed to solve the problem in all scenarios.

The root of this "market loop" problem—and a key area for future work—lies in developing methods to analyze marginal quantity contributions, not just costs. Our current Shapley Pooling equitably allocates the final community reward, but it abstracts the internal market into a black box; it does not disaggregate the underlying energy flows or transactions. Consequently, it is impossible to determine how much of an agent's position was met by the P2P market versus the external grid. While this level of decomposition may seem counterintuitive to the pooling mechanism's core philosophy of acting as a single unified agent, a method to analyze these marginal energies would be a significant breakthrough. It would primarily allow for a more efficient **Mix** market (under relative contracts). Agents could pool from their contracted share ($q_{n,t}^{\text{pool}}$) only the net-zero, internally matchable component, leaving the residual energy for competitive DA trading—unlike the current setup, where any imbalance in the pool is settled directly with the grid, leading to potentially missed trades. This same analysis would also, by validating commitments against actual internal need, unlock the use of absolute contracts as a viable alternative. Critically, this granular view would also permit evaluation of agent-level grid reliance in pooling mechanisms for the first time—a metric currently only computable at the community level—revealing far deeper insights into the true dynamics of cooperative P2P markets.

Finally, the two-stage MARL framework for the hybrid market can be remarkably advanced. The current setup, relying on random contract exploration, is quite naive. A more integrated approach would be a joint-learning framework where agents concurrently learn trading policies and contract proposals, which would enable guided, intelligent exploration of the contract space. This could be implemented using "shadow agents" dedicated to proposing contracts. Such a unified model would also be far more efficient, as an agent and its shadow proposer could share a single centralized critic—fostering tighter coordination—unlike our current decoupled setup, which requires retraining a separate critic for the second stage. Further algorithmic improvements could also enhance agent learning. Exploring action quantization (discretizing the continuous action space), for example, would enable action masking. Prohibiting invalid actions—like discharging an empty ES—would prune the exploration space, accelerating convergence and helping agents learn a critical "do-nothing" SoC action. The robustness of all proposed mechanisms could also be rigorously tested against adversarial agents who intentionally deviate from rational, cost-minimizing behavior.

These future directions—from scalable, physics-aware models to more sophisticated contract designs and robust algorithms—are all intriguing and necessary avenues of research. We are confident that continued investigation in this domain, building on the foundations presented here, will lead to the development of truly autonomous, efficient, and fair LEMs.

A. MAPPO Hyperparameters

Table A.1 reports the key hyperparameters used in the MARL setup. Each agent is equipped with its own actor and critic, without parameter sharing. The actor's final layer parameterizes a Tanh-Normal distribution to enable stochastic action sampling. Hyperparameters were selected after coarse tuning, with the aim of encouraging exploration beyond short-sighted strategies and avoiding premature convergence to greedy behaviors. In particular, the relatively high values of γ and λ, the elevated entropy coefficient, and the use of larger batch collections (corresponding to roughly 340 days of experience) were chosen to expose agents to a richer variety of scenarios during training.

As for training the second-stage Contract Proposal Environment, defined in Section 4.2.3, we employ a slightly relaxed set of hyperparameters, as summarized in Table A.2. This adjustment reflects the relative ease of the second-stage training, which converges rapidly to near-optimal solutions. Most of the heavy lifting is carried out in the first stage, while this stage serves as a knob for fine-tuning contract decisions.

Table A.1. MAPPO Hyperparameters

Parameter	**Value**
Total Frames	524,288
Collected Frames per Batch	8,192
Discount Factor γ	0.99
GAE λ	0.96
Learning Rate	5×10^{-4}
Adam ϵ	1×10^{-6}
Minibatch Size	256
Number of Minibatch Iterations	4
Clip ϵ	0.2
Entropy Coefficient	0.04
Polyak τ (Soft Target Update)	0.005
Actor NN	[128, 64] + ReLU
Critic NN	[256, 128] + ReLU

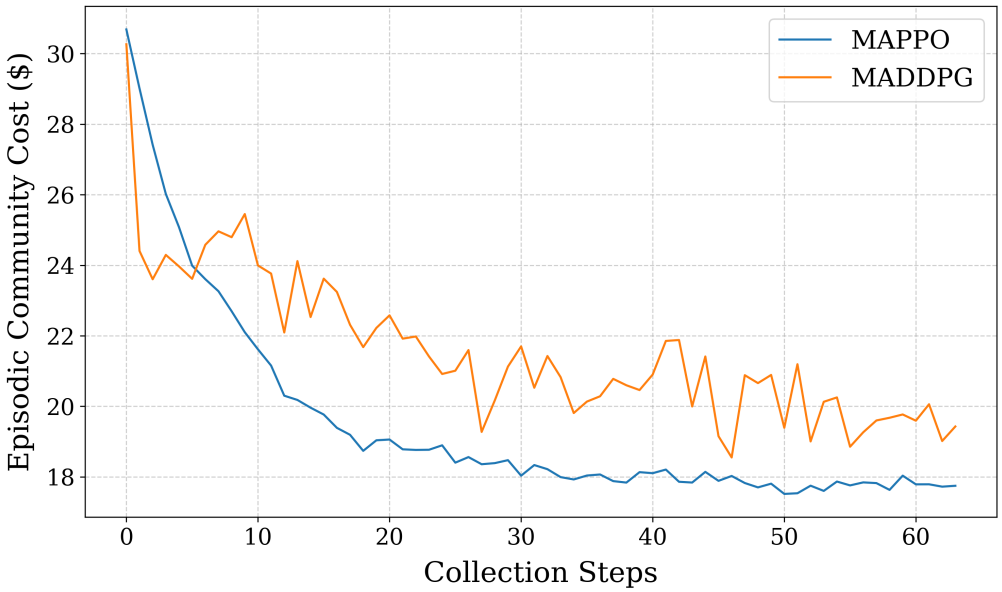
Table A.2. Modified Hyperparameters for second-stage Contract Proposal Environment

Parameter	**Value**
Total Frames	65,536
Collected Frames per Batch	4,096
Learning Rate	2.5×10^{-4}
Number of Minibatch Iterations	2
Entropy Coefficient	0.02
Actor NN	[64, 32] + ReLU

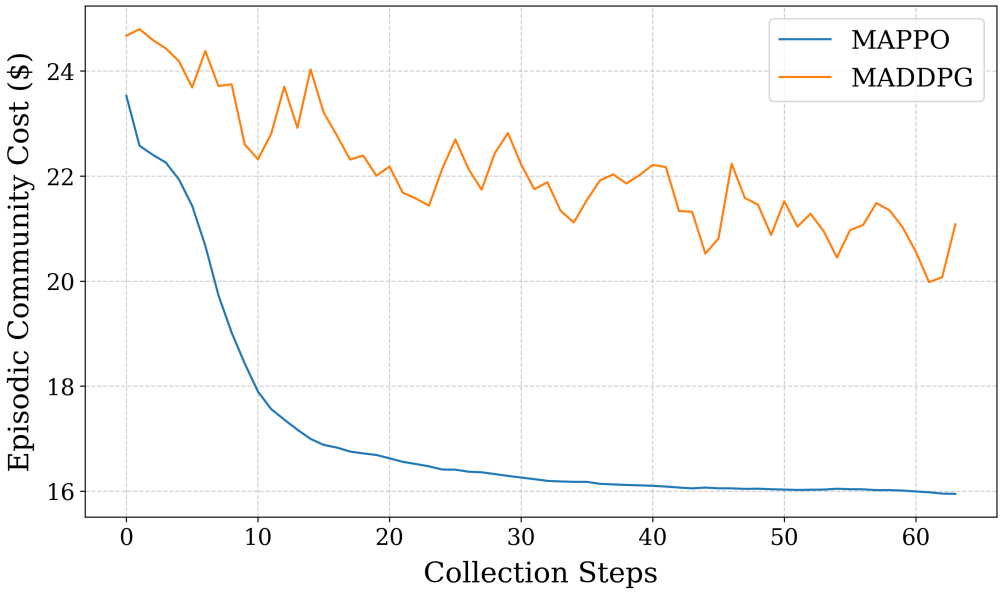
B. Greedy Policy Collapse in MADDPG

In the early stages, we experimented with MADDPG, but were quickly met with unfavorable results. By greedy policy collapse, we refer to the algorithm's tendency to converge to myopic discharge strategies due to poor exploration. In a large state space, relying on a suboptimal policy combined with random noise as the exploration strategy leaves many regions unexplored, causing premature convergence to short-sighted behaviors. To briefly illustrate, we present a training run for DA under MADDPG (using a single seed) and compare this to results under MAPPO, as shown in Section 5.2. Figure B.1 depicts the respective learning curves: while MAPPO converges stably to lower community costs, MADDPG remains volatile and fails to achieve comparable improvements. To dissect this further, Figure B.2 shows the emergent charge and discharge behavior under MADDPG, where agents converge toward persistently discharging across most hours of the day. They never learn to charge their ES cheaply during off-peak or shoulder hours and later shave off costs during peak periods. Instead, they act rather foolishly by discharging precious stored energy even at night hours when both demand and ToU prices are low. In effect, the flexibility of ES is never fully utilized. The erratic nature of this collapse is further evident in the hourly SoC action patterns in Figure B.3, where the learned policy appears unstable and unconverged.

We attempted extensive hyperparameter tuning to address this issue, but to no avail. Even DA-MADDPG [18] was tested and exhibited the same tendency toward greedy policy collapse. A similar collapse was also observed under the Shapley pooling extension, which further suggests that the issue is inherent to the algorithm rather than the P2P market employed. In our view, the root cause lies in the exploration strategy of MADDPG—we experimented with both Gaussian additive noise and ϵ-greedy exploration, yet neither proved effective. That said, the precise source of the problem remains unclear and may also relate to specifics of the BenchMARL implementation. For the purposes of this thesis, we therefore adopted MAPPO, which demonstrated greater robustness and lower sensitivity to hyperparameters. Nonetheless, we acknowledge that this diagnosis is preliminary, and it is possible that further investigation could reveal shortcomings in our setup.



(a) Training Performance



(b) Evaluation Performance

Figure B.1. Learning Curves comparing MADDPG with MAPPO

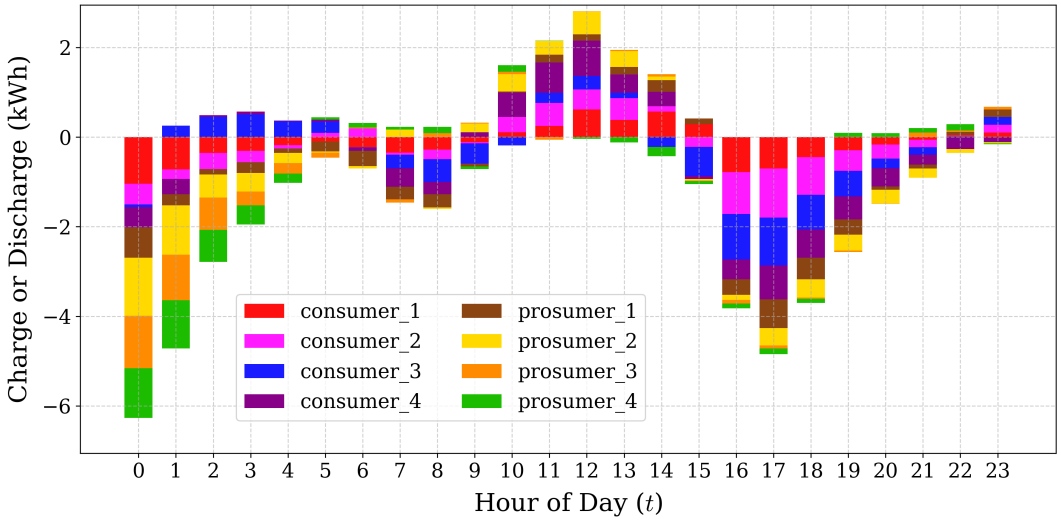


Figure B.2. ES Charging Schedule under MADDPG

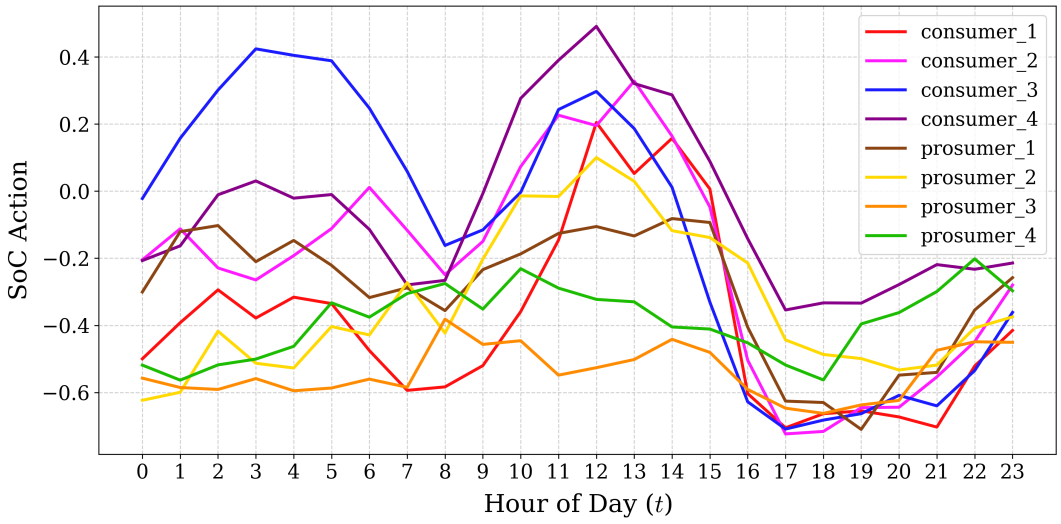


Figure B.3. Hourly SoC Action under MADDPG

List of Figures

List of Tables

Bibliography

[1] M. Haller, S. Ludig, and N. Bauer. "Decarbonization scenarios for the EU and MENA power system: Considering spatial distribution and short term dynamics of renewable generation." In: *Energy Policy* 47 (Aug. 2012), pp. 282–290. DOI: 10.1016/j.enpol.2012.04.069.

[2] H. Jiayi, J. Chuanwen, and X. Rong. "A review on distributed energy resources and MicroGrid." In: *Renewable and Sustainable Energy Reviews* 12.9 (Dec. 2008), pp. 2472–2483. DOI: 10.1016/j.rser.2007.06.004.

[3] S. Howell, Y. Rezgui, J.-L. Hippolyte, B. Jayan, and H. Li. "Towards the next generation of smart grids: Semantic and holonic multi-agent management of distributed energy resources." In: *Renewable and Sustainable Energy Reviews* 77 (Sept. 2017), pp. 193–214. DOI: 10.1016/j.rser.2017.03.107.

[4] S. Khaskheli and A. Anvari-Moghaddam. "Energy Trading in Local Energy Markets: A Comprehensive Review of Models, Solution Strategies, and Machine Learning Approaches." In: *Applied Sciences* 14.24 (Dec. 10, 2024), p. 11510. DOI: 10.3390/app142411510.

[5] A. Shrestha et al. "Peer-to-Peer Energy Trading in Micro/Mini-Grids for Local Energy Communities: A Review and Case Study of Nepal." In: *IEEE Access* 7 (2019), pp. 131911–131928. DOI: 10.1109/ACCESS.2019.2940751.

[6] M. I. A. Shah, A. Wahid, E. Barrett, and K. Mason. "Multi-agent systems in Peer-to-Peer energy trading: A comprehensive survey." In: *Engineering Applications of Artificial Intelligence* 132 (June 2024), p. 107847. DOI: 10.1016/j.engappai.2024.107847.

[7] S. Keren, C. Essayeh, S. V. Albrecht, and T. Morstyn. *Multi-Agent Reinforcement Learning for Energy Networks: Computational Challenges, Progress and Open Problems*. May 25, 2024. DOI: 10.48550/arXiv.2404.15583.

[8] D. Friedman and J. Rust. *The Double Auction Market: Institutions, Theories, and Evidence*. Santa Fe Institute. Boulder: Routledge, 2018. 1 p. DOI: 10.4324/9780429492532.

[9] L. S. Shapley. "17. A Value for n-Person Games." In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by H. W. Kuhn and A. W. Tucker. Princeton University Press, Dec. 31, 1953, pp. 307–318. DOI: 10.1515/9781400881970-018.

[10] A. Haupt, P. Christoffersen, M. Damani, and D. Hadfield-Menell. "Formal contracts mitigate social dilemmas in multi-agent reinforcement learning." In: *Autonomous Agents and Multi-Agent Systems* 38.2 (Dec. 2024), p. 51. DOI: 10.1007/s10458-024-09682-5.

[11] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. Second edition. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2018. 526 pp.

[12] Y. Shoham and K. Leyton-Brown. *Multiagent systems: algorithmic, game-theoretic, and logical foundations*. Cambridge: Cambridge University Press, 2009. 1 p. DOI: 10.1017/CBO9780511811654.

[13] F. A. Oliehoek and C. Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Cham: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-28929-8.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. *Proximal Policy Optimization Algorithms*. Aug. 28, 2017. DOI: 10.48550/arXiv.1707.06347.

[15] T. Haarnoja et al. *Soft Actor-Critic Algorithms and Applications*. Jan. 29, 2019. DOI: 10.48550/arXiv.1812.05905.

[16] R. Lowe, Y. WU, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. DOI: 10.48550/arXiv.1706.02275.

[17] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. WU. "The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games." In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 24611–24624. DOI: 10.48550/arXiv.2103.01955.

[18] 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: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Z.-H. Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 2913–2920. DOI: 10.24963/ijcai.2021/401.

[19] E. L. Ratnam, S. R. Weller, C. M. Kellett, and A. T. Murray. "Residential load and rooftop PV generation: an Australian distribution network dataset." In: *International Journal of Sustainable Energy* 36.8 (Sept. 14, 2017), pp. 787–806. DOI: 10.1080/14786451.2015.1100196.

[20] D. Qiu, J. Wang, Z. Dong, Y. Wang, and G. Strbac. "Mean-Field Multi-Agent Reinforcement Learning for Peer-to-Peer Multi-Energy Trading." In: *IEEE Transactions on Power Systems* 38.5 (Sept. 2023), pp. 4853–4866. DOI: 10.1109/TPWRS.2022.3217922.

[21] J. Zheng, Z.-T. Liang, Y. Li, Z. Li, and Q.-H. Wu. "Multi-Agent Reinforcement Learning With Privacy Preservation for Continuous Double Auction-Based P2P Energy Trading." In: *IEEE Transactions on Industrial Informatics* 20.4 (Apr. 2024), pp. 6582–6590. DOI: 10.1109/TII.2023.3348823.

[22] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac. "Scalable coordinated management of peer-to-peer energy trading: A multi-cluster deep reinforcement learning approach." In: *Applied Energy* 292 (June 2021), p. 116940. DOI: 10.1016/j.apenergy.2021.116940.

[23] Z. Ye, D. Qiu, S. Li, Z. Fan, and G. Strbac. "Federated Reinforcement Learning for decentralized peer-to-peer energy trading." In: *Energy and AI* 20 (May 2025), p. 100500. DOI: 10.1016/j.egyai.2025.100500.

[24] C. Samende, J. Cao, and Z. Fan. "Multi-agent deep deterministic policy gradient algorithm for peer-to-peer energy trading considering distribution network constraints." In: *Applied Energy* 317 (July 2022), p. 119123. DOI: 10.1016/j.apenergy.2022.119123.

[25] C. Feng and A. L. Liu. "Peer-to-peer energy trading of solar and energy storage: A networked multiagent reinforcement learning approach." In: *Applied Energy* 383 (Apr. 2025), p. 125283. DOI: 10.1016/j.apenergy.2025.125283.

[26] T. Chen, S. Bu, X. Liu, J. Kang, F. R. Yu, and Z. Han. "Peer-to-Peer Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning." In: *IEEE Transactions on Smart Grid* 13.1 (Jan. 2022), pp. 715–727. DOI: 10.1109/TSG.2021.3124465.

[27] J. Wang, L. Li, and J. Zhang. "Deep reinforcement learning for energy trading and load scheduling in residential peer-to-peer energy trading market." In: *International Journal of Electrical Power & Energy Systems* 147 (May 2023), p. 108885. DOI: 10.1016/j.ijepes.2022.108885.

[28] F. Charbonnier, T. Morstyn, and M. D. McCulloch. "Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility." In: *Applied Energy* 314 (May 2022), p. 118825. DOI: 10.1016/j.apenergy.2022.118825.

[29] L. Yan, X. Chen, Y. Chen, and J. Wen. "A Hierarchical Deep Reinforcement Learning-Based Community Energy Trading Scheme for a Neighborhood of Smart Households." In: *IEEE Transactions on Smart Grid* 13.6 (Nov. 2022), pp. 4747–4758. DOI: 10.1109/TSG.2022.3181329.

[30] M. Sanayha and P. Vateekul. "Model-Based Approach on Multi-Agent Deep Reinforcement Learning With Multiple Clusters for Peer-To-Peer Energy Trading." In: *IEEE Access* 10 (2022), pp. 127882–127893. DOI: 10.1109/ACCESS.2022.3224460.

[31] W.-Y. Chiu, C.-W. Hu, and K.-Y. Chiu. "Renewable Energy Bidding Strategies Using Multiagent Q-Learning in Double-Sided Auctions." In: *IEEE Systems Journal* 16.1 (Mar. 2022), pp. 985–996. DOI: 10.1109/JSYST.2021.3059000.

[32] A. Kulmala, M. Baranauskas, A. Safdarian, J. Valta, P. Järventausta, and T. Björkqvist. "Comparing Value Sharing Methods for Different Types of Energy Communities." In: *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*. 2021, pp. 1–6. DOI: 10.1109/ISGTEurope52324.2021.9640205.

[33] A. A. Raja and S. Grammatico. "Bilateral Peer-to-Peer Energy Trading via Coalitional Games." In: *IEEE Transactions on Industrial Informatics* 19.5 (May 2023), pp. 6814–6824. DOI: 10.1109/TII.2022.3196339.

[34] L. Han, T. Morstyn, and M. McCulloch. "Estimation of the Shapley Value of a Peer-to-peer Energy Sharing Game Using Multi-Step Coalitional Stratified Sampling." In: *International Journal of Control, Automation and Systems* 19.5 (May 2021), pp. 1863–1872. DOI: 10.1007/s12555-019-0535-1.

[35] J. Wang. "Shapley value based multi-agent reinforcement learning: theory, method and its application to energy network." In: (Dec. 2022). In collab. with Y. Gu, T. Green, T.-K. Kim, and Engineering And Physical Sciences Research Council. Publisher: Imperial College London. DOI: 10.25560/109306.

[36] M. Eichelbeck and M. Althoff. "Fair Cost Allocation in Energy Communities Under Forecast Uncertainty." In: *IEEE Open Access Journal of Power and Energy* 12 (2025), pp. 2–11. DOI: 10.1109/OAJPE.2024.3520418.

[37] D. Qiu, Y. Ye, and D. Papadaskalopoulos. "Exploring the effects of local energy markets on electricity retailers and customers." In: *Electric Power Systems Research* 189 (Dec. 2020), p. 106761. DOI: 10.1016/j.epsr.2020.106761.

[38] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey. "Meta-Learning in Neural Networks: A Survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. DOI: 10.1109/TPAMI.2021.3079209.

[39] M. Bettini, A. Prorok, and V. Moens. "BenchMARL: Benchmarking Multi-Agent Reinforcement Learning." In: *Journal of Machine Learning Research* 25.217 (2024), pp. 1–10.

[40] L. Ceriani and P. Verme. "The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini." In: *The Journal of Economic Inequality* 10.3 (Sept. 2012), pp. 421–443. DOI: 10.1007/s10888-011-9188-x.