Energy and AI 25 (2026) 100809

Contents lists available at ScienceDirect

Energy and AI

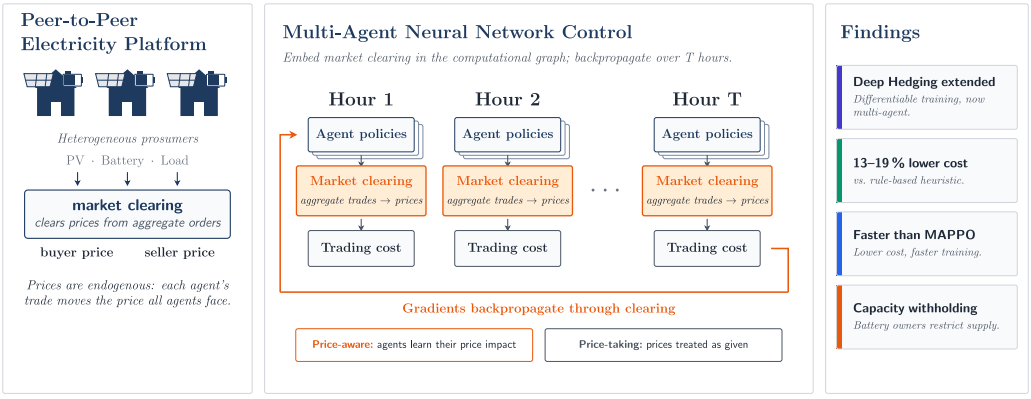
journal homepage: www.elsevier.com/locate/egyai

Peer-to-peer electricity platforms with endogenous prices: A Multi-Agent Neural Network Control approach

Nicolas Greber [a,b,*], Nicolas Eschenbaum [b], Oleg Szehr [c]

[a] University of Zurich, Switzerland
[b] Swiss Economics SE AG, Zurich, Switzerland
[c] Dalle Molle Institute for Artificial Intelligence (IDSIA) - SUPSI/USI, Lugano, Switzerland

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Deep Hedging extended to multi-agent electricity markets with endogenous prices.
- Reduces community cost by 13%–19% relative to a rule-based dispatch heuristic.
- Lower cost and faster training than a model-free reinforcement learning baseline.
- Capacity withholding emerges under price-aware learning and quantity-sensitive pricing.
- Findings reproduce on real Swiss smart-meter demand across twelve communities.

ARTICLE INFO

Keywords:
Electricity platforms
Peer-to-peer energy trading
Prosumer policy optimisation
Deep hedging
Multi-agent reinforcement learning
Multi-Agent Neural Network Control

ABSTRACT

The transition towards decentralised energy systems creates a need for local electricity platforms on which prosumers trade energy at endogenously determined prices. We develop a differentiable multi-agent control framework for such platforms, drawing on neural stochastic control and the pathwise optimisation principle used in Deep Hedging, whose appeal stems from its simplicity, computational efficiency, industrial use, and scalability to large-scale optimisation problems. The proposed approach embeds physical storage dynamics, feasible prosumer actions, and the market-clearing rule in a single computational graph, allowing decentralised agents to train trading policies by backpropagating through aggregate supply, aggregate demand, and cleared prices. A stop-gradient construction at the clearing layer separates two economically distinct learning modes: price-aware learning, in which agents internalise their marginal price impact, and price-taking learning, in

* Corresponding author.
E-mail address: nicolasjoel.greber@uzh.ch (N. Greber).

https://doi.org/10.1016/j.egyai.2026.100809
Received 10 April 2026; Received in revised form 2 June 2026; Accepted 9 June 2026
Available online 23 June 2026

which prices are treated as exogenous. This yields a market-mechanism-aware alternative to reward-only multi-agent reinforcement learning methods such as Multi-Agent Proximal Policy Optimisation, which treat the market as a black-box environment component. We evaluate the framework on peer-to-peer electricity markets with heterogeneous prosumers, photovoltaic generation, battery storage, and multiple non-discriminatory pricing rules, using both large-scale synthetic simulations and Swiss smart-meter demand data. The experiments show that Multi-Agent Neural Network Control learns effective decentralised policies, reduces community cost relative to rule-based control, and outperforms Multi-Agent Proximal Policy Optimisation in the studied settings in both cost and training efficiency. The learned policies exhibit emergent, economically sensible behaviour: they recover standard storage arbitrage and, under price-aware learning, reveal capacity withholding under quantity-sensitive clearing, which benefits flexible agents at the expense of aggregate community cost. Embedding market clearing in the training graph therefore provides both an efficient learning signal and a diagnostic tool for strategic behaviour in local electricity-market design.

1. Introduction

The transition towards decentralised energy systems is reshaping electricity markets. Increasingly, households and small commercial consumers are no longer passive load points but *prosumers* equipped with photovoltaic generation, batteries, and controllable demand. This paradigm shift creates a need for local market platforms on which prosumers exchange energy within a community before settling residual imbalances with the external grid. Such peer-to-peer and local energy markets promise higher utilisation of distributed generation, reduced grid dependence, and new forms of community-level coordination. They also create a difficult algorithmic problem: the trading decisions of many self-interested agents jointly determine aggregate supply, aggregate demand, and hence the internal market-clearing prices faced by all participants.

This paper develops algorithms for learning trading policies in such platform-based electricity markets. We consider a local electricity trading platform populated by heterogeneous prosumers with stochastic demand, photovoltaic generation, and battery storage. In each trading period, agents choose storage and curtailment actions, submit the resulting net positions to the platform, and receive buy and sell prices determined by a known market-clearing rule. The resulting decision problem is a general-sum Markov game with endogenous prices: agents share a common opportunity to improve community efficiency through internal trade, but each individual agent may also have an incentive to influence prices strategically. This motivates multi-agent reinforcement learning (MARL) approaches that model prosumers as interacting learning agents.

Methodologically, our approach builds on the neural stochastic control perspective of Han and E [1], in which time-dependent controls are parameterised by deterministic neural networks and trained end-to-end through simulated dynamics. Deep Hedging [2] is a finance-specific realisation of this idea for derivative hedging under market frictions. We transfer this differentiable control principle to local electricity markets and extend it to a multi-agent setting in which prosumers are strategically coupled through an endogenous market-clearing mechanism. Deep Hedging has been successful in financial applications because it is direct, scalable to high-dimensional problems, and capable of exploiting known structure in the simulation model rather than relying only on observed reward signals. We use the same principle for local electricity markets: the physical dynamics, the agents' feasible action mappings, and the market-clearing rule are embedded in a differentiable computational graph, and prosumer policies are trained end-to-end through the resulting unrolled market simulation. We refer to this extension as *Multi-Agent Neural Network Control* (MANNC).

Our contributions are fourfold. First, we formulate local peer-to-peer electricity trading with heterogeneous prosumers, storage, curtailment, and endogenous platform prices as a differentiable multi-agent control problem. Second, we introduce MANNC, a neural stochastic control architecture in which decentralised prosumer policies are trained end-to-end through market dynamics and a known market-clearing rule. Third, we use a stop-gradient construction at the cleared prices to separate price-aware learning, in which agents internalise their marginal price impact, from price-taking learning, in which prices are treated as exogenous. This provides a direct way to diagnose strategic behaviour induced by the clearing mechanism. Fourth, we evaluate the framework across clearing rules, community sizes, synthetic simulations, and Swiss smart-meter demand data, benchmarking against a centralised differentiable planner, Multi-Agent Proximal Policy Optimisation (MAPPO), a rule-based heuristic, and a perfect-foresight LP lower bound.

The MANNC architecture is motivated by three considerations. First, local energy platforms may involve large numbers of autonomous participants, making scalability a central architectural requirement. Deep Hedging is a natural template for this setting because it was designed for high-dimensional portfolio optimisation; in our experiments, we mirror this scaling challenge by starting from a six-agent base community and constructing markets with up to 96 participants. Second, when the clearing mechanism is known to the platform designer and market participants, treating it as a black-box environment wastes information. By differentiating through the clearing rule, each agent receives gradients that quantify how its actions affect aggregate volumes, cleared prices, and ultimately its own cost. This produces a market-aware learning signal absent from standard reward-only MARL methods. Third, the resulting algorithm is comparatively simple: it does not require many of the engineering components of MARL systems (such as value-function learning, temporal-difference targets, replay buffers, clipped surrogate objectives, or centralised critics). In contrast, methods such as proximal policy optimisation (PPO) and MAPPO rely on stochastic policy-gradient estimates, advantage estimation, policy-update epochs, and value-function approximation; related MARL algorithms (such as MADDPG, QMIX, and actor–critic variants) introduce additional machinery for credit assignment, centralised training, or value decomposition. Whereas the original Deep Hedging architecture relies on stochastic market paths for implicit exploration, we add decaying ε-greedy noise to mitigate convergence to poor local optima.

We evaluate MANNC in an experimental framework that tests both algorithmic performance and economic behaviour. We use large-scale synthetic simulations for controlled and reproducible comparison across market designs, agent compositions, and community sizes, and real Swiss smart-meter demand data to validate empirical robustness under realistic household consumption patterns. MANNC allows us to separate two economically distinct learning modes: a price-aware mode, in which gradients pass through the clearing mechanism and agents internalise their marginal price impact, and a price-taking mode, in which a stop-gradient operator treats cleared prices as exogenous. We compare both variants with a centralised differentiable planner, which provides a full-information efficiency benchmark, and with MAPPO, a representative model-free MARL baseline that learns from reward signals without passing gradients through the market mechanism.

Our experiments show that MANNC can solve the resulting platform optimisation problems and outperforms MAPPO in this setting. For a six-agent community, MANNC reduces mean daily community cost by 12.6–18.9% relative to a rule-based heuristic, compared with 2.5% for

Nomenclature

Abbreviations

AR(1)	First-order autoregressive
CTCE	Cent. Train., Cent. Exec.
CTDE	Cent. Train., Decent. Exec.
DP	Dynamic programming
DRL	Deep reinforcement learning
DTDE	Decent. Train., Decent. Exec.
GAE	Generalised Advantage Estim.
LIN	Linear pricing rule
LP	Linear program
MANNC	Multi-Agent Neural Net. Control
MAPPO	Multi-Agent Proximal Policy Opt.
MARL	Multi-agent RL
MMR	Mid-market rate rule
P2P	Peer-to-peer
PPO	Proximal Policy Opt.
PV	Photovoltaic
SDR	Supply and demand ratio
SoC	State of charge

Sets and indices

$\mathcal{P}$	Prosumers $\{1, \dots, N\}$
$\mathcal{T}$	Periods $\{1, \dots, T\}$
N, T	No. of prosumers, horizon
i, t, b	Prosumer, time, trajectory idx
ℓ	Network-layer index

Spaces and mappings

$\mathcal{S}$	Global state space
$\mathcal{A}^i$	Action space of agent i
$\mathcal{O}^i$	Observation space of agent i
$\mathcal{B}^i$	Feasible battery action set
P	Transition kernel
f	Market clearing rule
$\Delta(\cdot)$	Probability simplex

Prosumer and battery

d_t^i	Demand (kWh)
g_t^i	PV generation (kWh)
$\bar{g}^i$	Installed PV capacity (kWp)
S_t^i	Battery state of charge (kWh)
$\bar{S}^i$	Battery capacity (kWh)
b_t^i	Charge/discharge action (kW)
$\bar{b}^i, \underline{b}^i$	Max charge/discharge rate (kW)
$\eta^{\text{ch}}, \eta^{\text{dis}}$	Charge/discharge efficiency
κ_t^i	PV curtailment ratio
Δt	Time-step length (h)

Market

x_t^i	Net order (kWh)
m_t^i, q_t^i	Buy/sell quantity (kWh)
M_t, Q_t	Aggregate demand/supply (kWh)
$\bar{p}$	Retail import price (EUR/kWh)
$\underline{p}$	Feed-in tariff (EUR/kWh)
Δp	Tariff spread (EUR/kWh)
p_t^D, p_t^S	Internal buyer/seller price
r_t^S, r_t^D	MMR matching ratios
β	LIN price-sensitivity param.
c_t^i	Per-period trading cost (EUR)
C^i	Total cost over horizon (EUR)

Game and learning

s_t	Global state
a_t^i	Action (battery, curtailment)
o_t^i	Observation of agent i
r^i	Per-step reward
π^{θ^i}	Policy network of agent i
θ^i	Trainable parameters
W_ℓ, b_ℓ	Layer weights/biases
h^ℓ	Hidden-layer activations
$\sigma(\cdot)$	Activation function
L	Number of hidden layers
$\text{sg}(\cdot)$	Stop-gradient operator
$\bar{C}^i$	Batch-averaged objective
$\bar{d}_{t-1}$	Community-avg. demand (kWh), $t-1$
$\bar{g}_{t-1}$	Community-avg. PV generation (kWh), $t-1$
$\mathcal{R}^i$	Deviation regret (EUR/day)
B	Batch size (trajectories)
K	Number of training episodes
α	Learning rate
γ	Discount factor
ρ_k	Exploration rate at episode k

MAPPO, while requiring less training time on a single CPU. The cost advantage persists in scalability experiments up to 96 agents, with MANNC delivering 7.8–13.9% lower community cost than the rule-based heuristic at all tested scales. These results indicate that passing gradients through a known market-clearing layer provides a more informative learning signal than reward-only policy-gradient training for the class of local electricity markets considered here.

Beyond aggregate performance, the framework reveals economically interpretable emergent behaviour. All learning-based methods discover the expected diurnal storage-arbitrage pattern: charging during midday hours, when photovoltaic generation depresses internal prices, and discharging in the evening, when demand exceeds local supply. However, under the supply and demand ratio (SDR) clearing rule, the price-aware agents learn a more strategic variant of this behaviour. In particular, the pure battery agent charges less aggressively late in the solar period and withholds part of its stored energy during the evening peak. This capacity-withholding strategy sustains higher internal prices, reduces the battery owner's own cost, and shifts costs onto net buyers such as passive consumers and prosumers without storage. The same qualitative pattern emerges under linear pricing rules with high price sensitivity. The behaviour is individually rational but socially costly: the price-aware decentralised policies exhibit lower unilateral-deviation regret, suggesting approximate strategic stability, while producing higher aggregate cost than the price-taking decentralised policies. The withholding signature persists when evaluated using Swiss smart-meter demand data.

These findings position MANNC as both a learning algorithm and an analysis tool for local energy-market design. By embedding the clearing mechanism directly in the training graph, the framework enables platform designers to exploit known market structure, compare alternative information regimes, and diagnose potential strategic prosumer behaviour in controlled simulations. The resulting insights can inform market-design choices before field validation, while deployment-level assessment requires additional operational constraints such as network limits, settlement rules, and communication frictions.

1.1. Literature review

P2P electricity trading platforms and local energy markets have attracted growing attention. Tushar et al. [3] survey market architectures, pricing rules, and deployment barriers, and Khorasany et al. [4] provide a complementary classification of market designs and clearing approaches for local energy trading. Within this design space, two rule-based pricing mechanisms have become canonical references in the literature: the mid-market rate [5], which sets the internal price as the midpoint between the wholesale buy and sell prices, and the supply–demand ratio (ratio-based) pricing of Liu et al. [6], which adjusts the

local price to reflect the community-level energy balance. Auction-based clearing has also been studied [7], but is outside the scope of this work. A central issue is that the economic properties of these mechanisms depend on participant behaviour, motivating the joint study of market design and learning dynamics rather than relying on stylised behavioural assumptions such as zero-intelligence bidding.

Agent-based computational economics has a long tradition in electricity-market research [8], and recent work increasingly combines agent-based models with deep reinforcement learning (DRL). The ASSUME framework [9,10] is a comprehensive open-source toolbox in this space, coupling agent-based market models with DRL for adaptive bidding at scales up to 145 simultaneously learning agents; extensions address storage-unit bidding [11], bi-level optimisation benchmarks [12], and explainability of learned strategies [13]. In the peer-to-peer (P2P) setting specifically, May et al. [14] train DRL agents to participate in a ratio-based local energy market and benchmark against dynamic programming; Ye et al. [15] formulate P2P trading as a partially observable Markov game and propose a federated MARL approach for scalability and privacy; and Ye et al. [16] combine local energy trading with flexibility services using multi-agent DRL, with a model-based centralised formulation serving as an optimality benchmark. For broader surveys of MARL methods and the taxonomy of centralised versus decentralised training paradigms (CTDE, DTDE, CTCE) adopted in this paper, we refer to Zhang et al. [17] and Gronauer and Diepold [18].

A common property of these approaches is that they treat the clearing mechanism as a black-box environment component: agents interact with the market through reward signals but receive no gradient information about how their actions affect prices. This creates a disconnect between market design – where the clearing rule is a known, parametric object – and learning, which ignores this structure.

Deep Hedging, introduced by Buehler et al. [2], addresses dynamic hedging of derivative portfolios under market frictions such as transaction costs, market impact, liquidity constraints, and risk limits. The method's appeal lies in combining modelling flexibility with computational scalability: it can incorporate realistic frictions and constraints while remaining tractable in high-dimensional portfolio problems. Importantly for our setting, Deep Hedging also provides a natural way to embed a known market model into the computational graph, allowing gradients to propagate through the simulated dynamics rather than treating the market as a black-box environment. Architecturally, Deep Hedging is a derivatives-specific realisation of neural stochastic control and continuous-action reinforcement learning: it follows [1] in parametrising time-dependent controls by neural networks trained end-to-end through simulated dynamics, and its direct optimisation of a deterministic trading policy links it to deterministic policy-gradient methods such as DDPG [19]; see also Szehr [20] and Cannelli et al. [21] for an explicit reinforcement-learning interpretation of Deep Hedging.

Subsequent work extends the deep hedging paradigm beyond derivatives' portfolios. Krabichler and Teichmann [22] apply it to asset–liability management, Jaisson [23] formalise it as differentiable reinforcement learning for known single-agent market models, and Tschora et al. [24] use a differentiable approximation of the EUPHEMIA electricity-market clearing algorithm to incorporate price formation into training.

These contributions remain single-agent or single-decision-maker frameworks, without strategic coupling through shared prices. Since Deep Hedging is a deterministic policy-gradient architecture, it can converge to suboptimal local optima when exploration is insufficient [25], a general concern in deterministic policy-gradient methods that motivates explicit exploration noise [19,26,27]; in our multi-agent setting, simultaneous adaptation and endogenous prices make this issue more pronounced, so we introduce decaying (ε)-greedy noise during training.

Beyond the model and learning perspective, broader work on energy communities considers system-level dimensions such as hybrid renewable integration and policy frameworks [28], demand-side machine-learning optimisation [29], metaheuristic dispatch [30], battery lifecycle considerations [31], and component-level battery state estimation [32–34]. These dimensions are complementary to but distinct from the strategic learning question we address.

2. Model setup

We consider a peer-to-peer (P2P) electricity trading platform operated by a market operator and populated by N heterogeneous prosumers indexed by $\mathcal{P} = \{1, \dots, N\}$. Time is discretised into trading periods $\mathcal{T} = \{1, \dots, T\}$. In each period every prosumer observes its own local state (demand, generation, and battery state), chooses physical controls (storage charge/discharge and curtailment), and submits a net buy or sell order to the platform. The platform aggregates all submitted orders across agents, determines internal prices as a deterministic function of these aggregates and settles any residual imbalances with the external grid at regulated tariffs. We refer to this price-determination step as *market clearing*.

2.1. Prosumers, storage, and feasibility constraints

Demand and PV generation. Prosumer i has demand $d_t^i \geq 0$ and PV generation $g_t^i \geq 0$. Their difference $g_t^i - d_t^i$ is positive when generation exceeds demand and negative otherwise.

Battery model. Each prosumer owns a battery with state of charge (SoC) $S_t^i \in [0, \bar{S}^i]$, where $\bar{S}^i$ is the storage capacity. Throughout, overbars denote upper bounds and underbars lower bounds. At time t, prosumer i selects an energy transfer b_t^i ($b_t^i > 0$ charging, $b_t^i < 0$ discharging). Conversion losses with efficiency $\eta^{\mathrm{ch}}, \eta^{\mathrm{dis}} \in (0, 1]$ are applied at the battery side, so that the SoC evolves as

$$S_{t+1}^i = S_t^i + \eta^{\mathrm{ch}} (b_t^i)^+ - \frac{(b_t^i)^-}{\eta^{\mathrm{dis}}}, \tag{1}$$

where $(\cdot)^+ := \max\{\cdot, 0\}$ and $(\cdot)^- := \max\{-\cdot, 0\}$. The action b_t^i is constrained by the power limits $\underline{b}^i$ (discharge) and $\bar{b}^i$ (charge) and by the current SoC:

$$b_t^i \in \mathcal{B}^i(S_t^i) := \left[\max\{-\underline{b}^i, -\eta^{\mathrm{dis}} S_t^i\},\ \min\{\bar{b}^i, (\bar{S}^i - S_t^i)/\eta^{\mathrm{ch}}\}\right]. \tag{2}$$

The lower bound ensures $S_{t+1}^i \geq 0$ and the upper bound $S_{t+1}^i \leq \bar{S}^i$. Throughout, b_t^i denotes the grid-side energy exchange; conversion losses appear only in the SoC update (1), while the submitted order (3) uses b_t^i directly.

Curtailment. Prosumer i may curtail PV generation by choosing $\kappa_t^i \in [0, 1]$; the effective injected generation is $(1 - \kappa_t^i)\, g_t^i$.

Submitted net order. Given actions (b_t^i, κ_t^i) and local realisations (d_t^i, g_t^i), the grid-side energy balance determines the order that prosumer i submits to the platform:

$$x_t^i := \left(d_t^i - (1 - \kappa_t^i)\, g_t^i\right) + (b_t^i)^+ - (b_t^i)^-, \tag{3}$$

where the first term is the net load after curtailment, $(b_t^i)^+$ is energy drawn for charging, and $(b_t^i)^-$ is energy delivered by discharging. A positive value $x_t^i > 0$ is an import (buy) and $x_t^i < 0$ is an export (sell).

2.2. Market clearing

Aggregate demand and supply. Each submitted order x_t^i is either a buy (import), $m_t^i := (x_t^i)^+$, or a sell (export), $q_t^i := (x_t^i)^-$. The platform aggregates across all agents:

$$M_t := \sum_{i=1}^{N} m_t^i, \qquad Q_t := \sum_{i=1}^{N} q_t^i.$$

External tariffs and internal price corridor. Residual imbalances are settled with the external grid at regulated tariffs: a retail import price $\bar{p}$ and a feed-in export tariff $\underline{p}$, with $0 \leq \underline{p} \leq \bar{p}$. Internal platform prices are constrained to

$$\underline{p} \leq p_t^S \leq p_t^D \leq \bar{p}, \tag{4}$$

where p_t^S is the price received by sellers and p_t^D the price paid by buyers.

Clearing rule. The market operator applies a clearing rule

$$f : (M_t, Q_t, \underline{p}, \bar{p}) \mapsto (p_t^D, p_t^S), \tag{5}$$

mapping aggregate orders into internal prices subject to (4). The rule f is a design choice of the platform operator and is available to all market participants. In particular, market participants might use this representation in their optimisation models, e.g. by embedding f into the computational graph of gradient-based optimisation schemes. We consider three non-discriminatory instantiations of f – all satisfying the corridor constraint (4) with a uniform seller price p_t^S and buyer price p_t^D – that span a range of price-formation structures.

1. *Supply and demand ratio (SDR).* Following Liu et al. [6], prices are set as a nonlinear function of the ratio of aggregate supply to demand. Define

   $$\mathrm{SDR}_t := \frac{Q_t}{M_t + \varepsilon},$$

   where $\varepsilon > 0$ avoids division by zero. Internal prices are

   $$p_t^S = \begin{cases} \dfrac{\underline{p}\,\bar{p}}{(\bar{p} - \underline{p})\,\mathrm{SDR}_t + \underline{p}}, & 0 \le \mathrm{SDR}_t \le 1, \\ \underline{p}, & \mathrm{SDR}_t > 1, \end{cases}$$

   and

   $$p_t^D = \begin{cases} p_t^S \cdot \mathrm{SDR}_t + \bar{p}\,(1 - \mathrm{SDR}_t), & 0 \le \mathrm{SDR}_t \le 1, \\ \underline{p}, & \mathrm{SDR}_t > 1. \end{cases}$$

   When SDR_t is high (supply abundant), both prices move towards $\underline{p}$; when it is low (supply scarce), prices move towards $\bar{p}$.
2. *Mid-market rate (MMR).*
   Adapted from Long et al. [5], this mechanism anchors the internal reference price at the midpoint of the tariff corridor,

   $$p_t^{\mathrm{MMR}} := \frac{\underline{p} + \bar{p}}{2},$$

   and blends each agent's effective price between this midpoint and the external tariff according to the fraction of trade that can be matched internally. Define the matching ratios

   $$r_t^S = \min\left(\frac{M_t}{Q_t}, 1\right), \qquad r_t^D = \min\left(\frac{Q_t}{M_t}, 1\right). \tag{6}$$

   The sell and buy prices are then

   $$p_t^S = r_t^S\, p_t^{\mathrm{MMR}} + (1 - r_t^S)\,\underline{p}, \tag{7}$$

   $$p_t^D = r_t^D\, p_t^{\mathrm{MMR}} + (1 - r_t^D)\,\bar{p}. \tag{8}$$

   When supply and demand are balanced ($r_t^S = r_t^D = 1$), both prices equal the midpoint. When one side dominates, the surplus is settled at the external tariff, pulling the effective price towards $\underline{p}$ (for sellers) or $\bar{p}$ (for buyers).
3. *Linear pricing rule (LIN).* This mechanism expresses the internal reference price as a linear function of the relative supply–demand imbalance:

   $$p_t^{\mathrm{LIN}} = \mathrm{clamp}\left(\frac{\underline{p} + \bar{p}}{2} - \beta\,\frac{\bar{p} - \underline{p}}{2} \cdot \frac{Q_t - M_t}{Q_t + M_t},\ \underline{p},\ \bar{p}\right),$$

   with sensitivity parameter $\beta = 0.5$. The normalisation by $Q_t + M_t$ makes the price response scale-invariant: it depends on the *relative* imbalance between supply and demand rather than on their absolute levels. Side prices p_t^S and p_t^D are computed from the same pro-rata blend (7)–(8) using the matching ratios (6).

Per-period cost. Given internal prices, prosumer i incurs one-period trading cost

$$c_t^i = p_t^D\, m_t^i - p_t^S\, q_t^i, \tag{9}$$

and total cost over the planning horizon

$$C^i := \sum_{t=1}^{T} c_t^i - \underline{p}\, S_{T+1}^i, \tag{10}$$

where the final term values the energy remaining in storage at the end of the horizon, S_{T+1}^i, at the feed-in tariff $\underline{p}$; this terminal credit prevents the policy from spuriously depleting the battery in the last period.

2.3. Markov game formulation

We cast the trading platform as a finite-horizon Markov game [17, 35]. A Markov game is defined by a tuple $(\mathcal{P}, \mathcal{S}, \{A^i\}_{i\in\mathcal{P}}, P, \{r^i\}_{i\in\mathcal{P}}, T)$ consisting of agents $\mathcal{P}$, state space $\mathcal{S}$, per-agent action spaces A^i, transition kernel P, reward functions r^i, and horizon T. Since each prosumer observes its own demand, generation, and battery level but not those of other prosumers, we consider per-agent observation spaces O^i and define policies as mappings $\pi^i : O^i \to \Delta(A^i)$, where $\Delta(A^i)$ denotes the probability simplex over A^i (e.g. Štrupl et al. [25],Oliehoek and Amato [36],Eschenbaum et al. [37]).

State. The state comprises time, all battery levels, and exogenous processes:

$$s_t := \left(t,\ \{S_t^i, d_t^i, g_t^i\}_{i=1}^N\right) \in \mathcal{S}. \tag{11}$$

Actions. Each prosumer selects a continuous action

$$a_t^i := (b_t^i,\ \kappa_t^i) \in \mathcal{A}^i(s_t) := \mathcal{B}^i(S_t^i) \times [0, 1], \tag{12}$$

and the submitted trade x_t^i follows from (3).

Observations. At each step, prosumer i observes

$$o_t^i := \left(t,\ S_t^i,\ d_t^i,\ g_t^i,\ p_{t-1}^D,\ p_{t-1}^S,\ \bar{d}_{t-1},\ \bar{g}_{t-1}\right) \in \mathcal{O}^i, \tag{13}$$

where (p_{t-1}^D, p_{t-1}^S) are the most recent cleared prices broadcast by the platform. The observation excludes other prosumers' individual states and actions but includes aggregate community statistics broadcast by the platform.

Transition rule. Let $a_t = (a_t^1, \ldots, a_t^N)$ denote the joint action. The transition comprises (i) battery updates via (1), (ii) stochastic evolution of demand and generation (d_{t+1}^i, g_{t+1}^i), and (iii) deterministic market clearing, in which f maps the aggregates induced by $\{x_t^i\}_{i=1}^N$ to prices.

Reward. We define the per-step reward as the negative of the trading cost:

$$r^i(s_t, a_t) := -c_t^i. \tag{14}$$

Each prosumer seeks to minimise its expected total cost $\mathbb{E}[C^i]$ over the planning horizon, with C^i as in (10) (the terminal credit enters as a terminal reward).

General-sum structure. The game is *general-sum* [17, Section 2.1]: it is neither zero-sum (fully competitive) nor cooperative (fully aligned). All prosumers benefit from internal trading, which avoids the external retail price $\bar{p}$ or the low feed-in tariff $\underline{p}$, yet each has an individual incentive to shift aggregate volumes to obtain favourable prices. A centralised planner that minimises the sum of all agents' costs treats the problem as cooperative; we use this as an efficiency benchmark in Section 3.4.

Real-world execution. Each prosumer computes a_t^i from its observation o_t^i and submits the implied trade x_t^i to the platform. The operator aggregates all trades, clears prices via (5), and broadcasts (p_t^D, p_t^S).

2.4. Gradient structure of the clearing rule

This work studies gradient-based policy-optimisation schemes, where the clearing rule f is represented as part of the computational graph. The pathway from actions to costs is

$$a_t \mapsto \{x_t^i\}_{i=1}^N \mapsto (M_t, Q_t) \mapsto f(M_t, Q_t) \mapsto (p_t^D, p_t^S) \mapsto \{c_t^i\}_{i=1}^N. \tag{15}$$

Gradients along this chain quantify *price impact*: how a marginal change in a prosumer's trade perturbs the cleared prices and the costs of all agents. Because all three clearing rules map (M_t, Q_t) to prices via closed-form expressions, the gradient $\nabla_{a^i} p_t$ is available in closed form for each mechanism, and the same training infrastructure applies without modification.

Non-smooth components. Non-differentiabilities arise from $(\cdot)^+$ and $(\cdot)^-$, the piecewise definitions, the $\min(\cdot, 1)$ operators, and the clamp in Section 2.2. In practice we rely on the subgradients supplied by automatic differentiation. When a fully smooth graph is needed, these operators can be replaced by softplus-type relaxations.

By the chain rule, agent i's gradient splits into a direct term and a price-impact term,

$$\nabla_{\theta^i} c_t^i = \underbrace{\frac{\partial c_t^i}{\partial x_t^i}\Big|_p \frac{\partial x_t^i}{\partial \theta^i}}_{g_{\text{direct}}} + \underbrace{\frac{\partial c_t^i}{\partial p_t}\frac{\partial p_t}{\partial x_t^i}\frac{\partial x_t^i}{\partial \theta^i}}_{g_{\text{price}}},$$

where $\partial p_t / \partial x_t^i$ is agent i's price impact through the shared aggregates. The stop-gradient sets g_{price}=**0**.

3. Multi-agent learning framework

We train prosumer policies within a shared simulation environment – the platform operator's digital twin – that models physical dynamics, stochastic demand and generation, and the market clearing rules of Section 2.2. The central idea is to *unroll* complete trajectories of this environment and optimise policies end-to-end via backpropagation through the resulting computational graph. This is analogous to the deep hedging architecture of Buehler et al. [2], which has found widespread adoption in industry applications [22]. The novelty of the present work lies in extending this pathwise approach to a multi-agent setting in which agents are strategically coupled through a shared price-formation mechanism: each agent's action affects the clearing prices faced by all other agents, turning the optimisation problem into a game. Following the taxonomy of Maggiolo et al. [38], end-to-end optimisation through a known simulator corresponds to supervised learning; we adopt this terminology throughout.

We consider two MARL frameworks, following the standard taxonomy of Zhang et al. [17]:

1. *Decentralised Training, Decentralised Execution* (DTDE): each agent maintains an independent policy parameterised by its own weights and acts from local observations only. The two gradient modes defined in Section 3.3 – price-aware (strategic) and price-taking (competitive) – apply within this setting.
2. *Centralised Training, Centralised Execution* (CTCE): a single global policy observes the full system state and outputs joint actions for all agents, minimising community cost.

We call the resulting framework *Multi-Agent Neural Network Control* (MANNC) and instantiate it in two settings: MANNC-DTDE for decentralised training and execution (Section 3.2), and MANNC-CTCE for the centralised planner benchmark (Section 3.4).

The CTCE planner serves as a cooperative efficiency bound: it represents the best a fully coordinated planner can achieve, but is not deployable at scale because the joint action space grows exponentially in N. As a model-free baseline we additionally compare against Multi-Agent Proximal Policy Optimisation [MAPPO;39], a Centralised-Training, Decentralised-Execution (CTDE) method in which a shared critic estimates state values. In contrast to the supervised approach, MAPPO treats the environment – including the clearing mechanism – as a black box and learns from reward signals alone.

Fig. 1 illustrates both architectures; Sections 3.1–3.4 provide the technical details.

Exploration. It is known that gradient-based optimisation in the context of deep hedging can get stuck in local optima [38]. To remedy this, we inject exploration noise via a decaying ε-greedy scheme: with probability ε_k (decreasing over episodes) each agent replaces its policy output with a uniformly random feasible action, promoting broad exploration early and a gradual shift towards exploitation. This technique plays a central role in achieving our performance results; Appendix C.2 (Fig. C.1(a)) compares convergence with and without noise injection.

3.1. Unrolled objective

For a batch of B sampled trajectories, each of horizon T, the unrolled objective for agent i is

$$\bar{C}^i = \frac{1}{B}\sum_{b=1}^{B} C_b^i, \tag{16}$$

where $C_b^i = \sum_{t=1}^{T} c_{t,b}^i - \underline{p}\, S_{T+1,b}^i$ is agent i's total cost (10) on trajectory b, including the terminal storage credit. The gradient $\nabla_{\theta^i}\bar{C}^i$ captures both the direct effect of actions on traded quantities and the indirect effect through endogenous prices.

3.2. Decentralised training and execution (DTDE)

Policy architecture. Each agent $i \in \mathcal{P}$ maintains its own policy $\pi^{\theta^i} : \mathcal{O}^i \to \Delta\mathcal{A}^i$, parameterised as a feedforward neural network with L hidden layers[1,2]

$$h^0 = o^i, \qquad h^\ell = \sigma(W_\ell\, h^{\ell-1} + b_\ell), \quad \ell = 1, \dots, L-1, \qquad \pi^{\theta^i}(o^i) = \tanh(W_L\, h^{L-1} + b_L), \tag{17}$$

where $\theta^i = \{W_\ell, b_\ell\}_{\ell=1}^{L}$ are agent-specific trainable parameters. The raw outputs in $[-1, 1]$ are scaled by the charge/discharge rate limits $\bar{b}^i, \underline{b}^i$ and applied to the battery; the resulting SoC is clamped to $[0, \bar{S}^i]$, which enforces the feasibility bounds (2), while the curtailment output is mapped affinely to $\kappa_t^i \in [0, 1]$. At each of the T time steps the same network is applied independently, so the full policy over one episode corresponds to a stack of T identical networks of depth L.

Each agent's parameter update depends only on its own cost gradient. In the price-taking mode, these gradients are fully decoupled; in the price-aware mode, they are coupled through the shared clearing rule. Because prices depend only on the two scalars (M_t, Q_t), the forward pass scales linearly with the number of agents and the horizon. The stop-gradient backward pass inherits the same linear cost; under price-aware learning, gradient coupling through the shared aggregates makes the backward pass scale quadratically in the number of agents.

3.3. Price-aware vs. price-taking learning

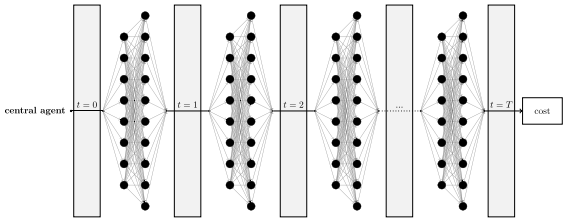
The clearing rule supports two learning modes that differ in a single operation: whether gradients pass through the cleared prices or are stopped. Let $\mathrm{sg}(\cdot)$ denote the stop-gradient operator (`.detach()` in PyTorch), which evaluates its argument in the forward pass but blocks gradient flow in the backward pass: $\nabla_\theta\, \mathrm{sg}(z(\theta)) = \mathbf{0}$.

Price-aware learning (strategic). Gradients flow through the full chain from costs to actions, including the clearing rule f. Agent i's objective is

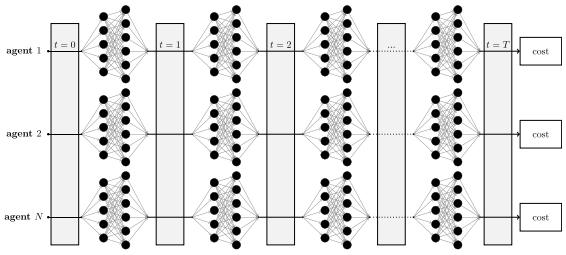
$$\bar{C}_{\text{PA}}^i(\theta) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T} c^i\big(a_{t,b}^i(\theta^i),\ p_{t,b}(\theta)\big), \tag{18}$$

[1] The tanh output bounds the two components – battery charge/discharge rate and PV curtailment ratio – to $[-1, 1]$.

[2] We also investigated recurrent architectures (RNN, LSTM) but did not observe substantial performance differences.



(a) Centralised architecture (CTCE). A single neural network receives the complete system state s_t and outputs actions for all N agents. Grey blocks denote the transition.



(b) Decentralised architecture (DTDE). Each agent i has an individual neural-network policy mapping its private observation o_t^i to actions.

Fig. 1. Comparison of centralised and decentralised policy architectures.

where $\theta = (\theta^1, \dots, \theta^N)$ collects all agents' parameters and $p_{t,b}(\theta) = f\big(\mathbf{x}_{t,b}(\theta), \underline{p}, \bar{p}\big)$. Because $\nabla_{\theta^i} p_{t,b} \neq \mathbf{0}$, the policy update

$$\theta^i \leftarrow \theta^i - \alpha \nabla_{\theta^i} \bar{C}^i_{\mathrm{PA}}(\theta) \tag{19}$$

accounts for both the direct effect of the trade and the indirect effect through prices. The agent learns, for instance, that increasing supply may lower the sell price.

Price-taking learning (competitive). The cleared prices are detached from the graph before entering the cost function. Defining $\tilde{p}_{t,b} = \mathrm{sg}\big(p_{t,b}(\theta)\big)$, the objective becomes

$$\bar{C}^i_{\mathrm{PT}}(\theta^i) = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{T} c^i\big(a^i_{t,b}(\theta^i),\ \tilde{p}_{t,b}\big). \tag{20}$$

Because $\nabla_{\theta^i} \tilde{p}_{t,b} = \mathbf{0}$, the update

$$\theta^i \leftarrow \theta^i - \alpha \nabla_{\theta^i} \bar{C}^i_{\mathrm{PT}}(\theta^i) \tag{21}$$

optimises only with respect to trade quantities, treating prices as given. This mode corresponds to the standard economic price-taking assumption.

Algorithm 1 summarises the DTDE training procedure; the evaluation protocol is described in Section 4.

Algorithm 1 MANNC-DTDE

Require: Agents $\mathcal{P} = \{1, \dots, N\}$, horizon T, batch size B, episodes K, learning rate α, mode $\in$ {price-aware, price-taking}

for episode $k = 1, \dots, K$ **do**
  Set exploration rate ρ_k
  for trajectories $b = 1, \dots, B$ **(in parallel) do**
    Initialise state s_0; set $C^i_b \leftarrow 0$ for all i
    for $t = 1, \dots, T$ **do**
      Each agent i computes $a^i_t = \pi^{\theta^i}(o^i_t)$
      With probability ρ_k, replace a^i_t with $\tilde{a}^i_t \sim \mathrm{Uniform}(\mathcal{A}^i)$
      Platform clears market: $(p^D_t, p^S_t) \leftarrow f(M_t, Q_t, \underline{p}, \bar{p})$
      Transition: $s_{t+1} \leftarrow$ (battery update, exogenous draw, price broadcast)
      $C^i_b \mathrel{+}= c^i_t$
    end for
    $C^i_b \mathrel{-}= \underline{p}\, S^i_{T+1,b}$ {terminal storage credit, Eq. (10)}
  end for
  Compute batch objective $\bar{C}^i \leftarrow \frac{1}{B} \sum_b C^i_b$
  if price-aware **then**
    $\theta^i \leftarrow \theta^i - \alpha \nabla_{\theta^i} \bar{C}^i_{\mathrm{PA}}(\theta)$ for all i
  else
    $\theta^i \leftarrow \theta^i - \alpha \nabla_{\theta^i} \bar{C}^i_{\mathrm{PT}}(\theta^i)$ for all i
  end if
end for

3.4. Centralised planner benchmark (CTCE)

A single policy $\pi^\theta : \mathcal{S} \to \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$ is trained to minimise the total system cost $\sum_{i=1}^N C^i$ via batch backpropagation.[3] The architecture mirrors that of the decentralised agents but receives the complete state s_t and outputs actions for all agents simultaneously; the same ϵ-greedy exploration schedule applies. Algorithm 2 summarises the procedure.

Algorithm 2 MANNC-CTCE

Require: Number of agents N, horizon T, batch size B, episodes K, learning rate α

for episode $k = 1, \dots, K$ **do**
  Set exploration rate ρ_k
  for trajectories $b = 1, \dots, B$ **(in parallel) do**
    Initialise global state s_0; set $C_b \leftarrow 0$
    for $t = 1, \dots, T$ **do**
      Compute joint action $a_t = \pi^\theta(s_t)$
      With probability ρ_k, replace a_t with $\tilde{a}_t \sim \mathrm{Uniform}(\prod_i \mathcal{A}^i)$
      Platform clears market; transition to s_{t+1}
      $C_b \mathrel{+}= \sum_{i=1}^N c^i_t$
    end for
    $C_b \mathrel{-}= \underline{p} \sum_{i=1}^N S^i_{T+1,b}$ {terminal storage credit}
  end for
  $\theta \leftarrow \theta - \alpha \nabla_\theta \big(\frac{1}{B} \sum_b C_b\big)$
end for

4. Experimental design

We evaluate the proposed framework on a peer-to-peer energy trading platform populated by heterogeneous residential prosumers. We present results for both synthetic data generated from calibrated physical models (Sections 4.1–4.2) and empirical Swiss smart-meter traces (Section 4.2; validation in Section 5.6).

4.1. Community composition

We consider the following experimental setup, where we optimise the energy consumption of an energy-trading community. This stylised community comprises $N = 6$ prosumers representing six distinct residential archetypes commonly found in central-European settings, see Table 1. We refer to this configuration as the base community throughout.

[3] This centralised optimiser differs from the standard *central planner* in economic theory, which selects resource allocations to maximise social welfare. Our optimiser instead selects players' *actions* to minimise total cost, subject to the same physical dynamics and market clearing rule.

Table 1
Prosumer archetypes in the base community (N=6). One-way battery efficiencies $\eta^{ch} = \eta^{dis} = 0.95$ for all storage-equipped agents (round-trip ≈ 0.90).

ID	Archetype	PV (kWp)	Battery (kWh)	Charge/discharge rate (kW)	Avg. demand (kWh/day)	Role
A	Full prosumer	8.0	8.0	2.5	13	Controllable
B	PV + small battery	5.0	4.0	2.0	11	Controllable
C	Flexibility provider	0.0	10.0	3.0	15	Controllable
D	PV generator	8.0	0.0	–	10	Passive
E	Small PV consumer	3.0	0.0	–	12	Passive
F	Pure consumer	0.0	0.0	–	14	Passive

The archetypes differ in installed PV capacity (0–8 kWp), battery storage (0–10 kWh), and average daily electricity consumption (10–15 kWh/day), reflecting the heterogeneity observed in field deployments [40]. Daily consumption values correspond to annual consumption of approximately 3 650–5 475 kWh, consistent with typical central-European single-family households [41].

Three agents (A–C) own battery storage and are *controllable*: their charge/discharge schedules and PV curtailment ratios are the decision variables optimised by the learning algorithms. The remaining three agents (D–F) are *passive* participants whose net trade is determined entirely by their demand and PV generation realisations. This 50/50 split between controllable and passive agents reflects a realistic market environment in which battery strategies must co-exist with exogenous supply and demand flows.

Agent C (flexibility provider) owns a large battery (10 kWh, C/3.3 rate) but no PV generation, and must therefore purchase all stored energy from the community market. Its optimal strategy is pure temporal arbitrage – buying when community prices are low and selling when they are high – making it sensitive to the price formation mechanism and the information available during training.

For the scalability analysis (Appendix A), we replicate the six-agent base unit to create communities of $N \in \{6, 12, 24, 48, 96\}$ agents, preserving the proportional composition of archetypes at every scale. As the community grows, the ratio of controllable to passive agents remains constant, but each individual battery owner's share of total market volume decreases, reducing per-agent price impact.

4.2. Data

We model a daily trading horizon of T=24 periods at hourly resolution ($\Delta t = 1$ h). We use two data sources: *synthetic* generator calibrated to central-European residential profiles (our primary setting, below), and a *real-world* instance built from Swiss smart-meter demand and PVGIS solar data, used to validate the findings in Section 5.6. Environmental uncertainty enters through two stochastic processes – photovoltaic (PV) generation and household demand – both generated synthetically from calibrated physical models.

PV generation. Solar output for prosumer i at hour t is

$$g_t^i = \bar{g}^i \cdot \hat{g}_t \cdot \xi_t^{PV} \cdot \Delta t, \tag{22}$$

where $\bar{g}^i$ is the installed PV capacity (kWp), $\hat{g}_t$ is a clear-sky irradiance curve (sinusoidal between sunrise at 06:00 and sunset at 18:00, zero otherwise), and $\xi_t^{PV} \in [0, 1]$ is a stochastic cloud-cover factor. The cloud-cover factor follows a mean-reverting AR(1) process:

$$\xi_t^{PV} = \mu + \phi(\xi_{t-1}^{PV} - \mu) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{23}$$

with autocorrelation $\phi = 0.85$, noise standard deviation $\sigma = 0.20$, and daily mean $\mu \sim \mathcal{U}[0.6, 0.95]$ drawn independently for each simulated day. To model transient cloud events, we superimpose 1–4 random Gaussian dips of width 0.4–1.2 h and depth 20%–70% during daylight hours. The resulting scaling factor is clipped to $[0.02, 1.0]$.

Demand. Demand d_t^i follows a diurnal baseline curve with characteristic morning (05:00–08:00) and evening (17:00–22:00) peaks and a midday dip (09:00–17:00). The midday dip coincides with peak PV generation, creating a systematic surplus that motivates intra-community trading. Intra-day variability is introduced via an AR(1) process (ϕ=0.6, σ=0.25) and 1–3 random Gaussian pulses of width 0.3–1.0 h and amplitude 0.1–0.5, representing stochastic appliance usage. Daily total demand is log-normally distributed (σ=0.2) to capture seasonal and occupancy variation. Final demand is normalised to match each prosumer's average daily consumption $\bar{d}^i$ and clipped to $[0.4, 1.8]$ times the baseline, see Table 1.

Dataset. We generate 1 000 days for training and 100 days for evaluation. All methods are evaluated on the same fixed test set. Training details (batch size, number of episodes) are given in Section 4.6

Real-world data. To test the findings against realistic profiles, we build real-data instances of the same six-archetype community. Household demand is taken from the CKW open smart-meter dataset for canton Lucerne, Switzerland [42], which records grid consumption only; for each archetype we form a band of metered households *without on-site PV* whose mean daily consumption lies within ±15% of the archetype level, and we sample K independent communities by drawing one real meter per archetype. PV generation is obtained from the European Commission's PVGIS service [43] (PVGIS-SARAH3) at Lucerne (47.05° N, 8.31° E) for the archetype-specified peak power (3/5/8 kWp, 14% system loss) over the same years as the demand.[4] Battery sizing and the controllable/passive split follow the archetype definition unchanged. Each community spans 730 days (2021–2022), split 80/20 into 584 training and 146 test days. This setup keeps the community *composition* fixed while replacing the synthetic demand and PV traces with real and irradiance-modelled ones; reporting results as mean ± s.d. across the K communities isolates whether the conclusions persist under realistic load and weather variability and are robust to the choice of metered households, rather than claiming a fully measured community.

4.3. Market parameters

All of the experiments use one of the pricing rules described in Section 2.2. The external tariff corridor is set to $\underline{p} = 0.05$ EUR/kWh (feed-in tariff) and $\bar{p} = 0.30$ EUR/kWh (retail grid import price), calibrated to typical central-European residential tariff structures. The resulting spread of $\bar{p} - \underline{p} = 0.25$ EUR/kWh provides strong arbitrage incentives for battery owners: even with round-trip efficiency losses of 10%, a full charge–discharge cycle yields a net margin exceeding 0.20 EUR/kWh in favourable market conditions. No platform fee is charged.

4.4. Benchmark methods

We compare the mechanism against a model-free MARL method, a heuristic baseline and a perfect-foresight LP as a lower bound.

[4] Because the smart-meter data records consumption only, PV is irradiance-modelled rather than measured – the standard way of adding generation to load data; we draw it for 2021–2022 to match the demand years, so there is no calendar offset.

Table 2
Method comparison by information structure. "Endogenous" means prices depend on agents' joint actions and gradients propagate through them; "exogenous" means prices are treated as given (stop-gradient)..

Method	Training	Price signal in training	Decentralised
MANNC-DTDE	Supervised	Endogenous (gradient)	Yes
MANNC-DTDE (SG)	Supervised	Exogenous (stop-grad)	Yes
MANNC-CTCE	Supervised	Endogenous (gradient)	No
MAPPO	RL (PPO)	Reward-only	Yes[a]
Perfect-foresight LP[b]	LP (foresight)	—	No
Rule-based	—	–	Yes

[a] Centralised training (shared critic), decentralised execution.
[b] Non-causal: optimises with full future information; a lower bound, not a deployable method.

- **MANNC-DTDE** (decentralised, price-aware): Per-agent feedforward networks trained end-to-end through the differentiable market layer. Gradients flow through the price formation process, enabling agents to learn how their actions affect clearing prices.
- **MANNC-DTDE (SG)** (decentralised, price-taking): Identical architecture and training procedure, but a stop-gradient operator detaches prices from the computational graph before back-propagation. Agents optimise against prices they treat as exogenous, analogous to the price-taking assumption in competitive equilibrium theory.
- **MANNC-CTCE** (centralised social planner): A single centralised network observes the full system state and outputs joint actions for all agents. Trained to minimise community cost, it serves as an efficiency upper bound under full information.
- **MAPPO** [39]: The model-free CTDE baseline introduced in Section 3.
- **Rule-based** (greedy self-consumption): Charges the battery from excess PV generation and discharges to cover demand deficits, with no lookahead or price awareness. This is the standard heuristic deployed in commercial home energy management systems.
- **Perfect-foresight LP** (perfect-foresight lower bound): A single planner with advance knowledge of the full day's demand and generation minimises the total community grid bill, subject to the *same* physical constraints as the learned policies – charge/discharge limits, SoC bounds, conversion efficiencies, and PV curtailment. Internal trades net out at the community level, so the total equals the external grid bill independent of the clearing rule; the problem is therefore a linear program, solved to optimality. Being non-causal and centralised it is not deployable, but its cost lower-bounds that of any causal policy.

Table 2 summarises the information structure exploited by each method.

4.5. Policy architectures and observations

Observation space. All decentralised agents (MANNC-DTDE, MANNC-DTDE (SG), MAPPO) receive the same observation vector at each time step t:

$$o_t^i = \left(\frac{S_t^i}{\bar{S}^i},\ \frac{t}{T},\ \tilde{d}_t^i,\ \tilde{g}_t^i,\ \frac{p_{t-1}^S}{\Delta p},\ \frac{p_{t-1}^D}{\Delta p},\ \frac{p_{t-1}^D - p_{t-1}^S}{\Delta p},\ \bar{d}_{t-1},\ \bar{g}_{t-1} \right), \tag{24}$$

where S_t^i is battery SoC, $\bar{S}^i$ battery capacity, $\tilde{d}_t^i$ and $\tilde{g}_t^i$ are the agent's normalised demand and PV generation, and $\Delta p = \bar{p} - \underline{p} = 0.25\,\text{EUR/kWh}$ is the tariff spread used for price normalisation. The lagged market signals p_{t-1}^S, p_{t-1}^D are the community clearing prices from the previous period, and $\bar{d}_{t-1}$, $\bar{g}_{t-1}$ are the community-average demand and generation. At t=0, all lagged features are initialised to zero.

The inclusion of lagged aggregate statistics reflects a realistic information structure in which the market platform publishes community-level summaries – but not individual prosumer data – after each trading period.

The centralised planner (MANNC-CTCE) observes the full system state $s_t = (t/T,\ \{S_t^i/\bar{S}^i,\ \tilde{d}_t^i,\ \tilde{g}_t^i\}_{i=1}^N)$. We give this benchmark system an artificial information advantage – full observability of all agents' states – to isolate the performance that can be reached in principle from the constraints imposed by partial observability. It does not observe market prices, as prices are a consequence of the joint actions it prescribes.

Table 3
Training hyperparameters. With γ=1 MAPPO's objective is undiscounted, matching the daily-cost evaluation metric and the full-episode MANNC objective; the MANNC methods carry no discount as they backpropagate the whole-episode cost directly.

Hyperparameter	MANNC-DTDE/MANNC-CTCE	MAPPO
Episodes	50	50
Batch size (days)	32	32
Optimiser	Adam	Adam
Learning rate	10^{-3}	5×10^{-4}
Gradient clipping	None	max norm 10.0
Discount factor γ	1.0	1.0
GAE λ	–	0.95
PPO clip ϵ	–	0.2
PPO epochs per episode	–	15
Minibatches per epoch	–	1 (full batch)
Entropy coefficient	–	0.05
Exploration	ϵ-greedy, $\rho_k = 0.7 \cdot 0.9^k$	Gaussian $\sigma_0 \approx 0.5$
Hidden layers	2×16 (2×16 N for CTCE)	2×64

MANNC networks. The policy architecture is described in Section 3.2. Decentralised agents use hidden dimension 16; the centralised planner scales this to $16N$ and produces $2N$ joint actions. Table 3 lists all architectural parameters.

MAPPO networks. MAPPO actors and critic use two hidden layers of 64 units with tanh activations, orthogonal weight initialisation [44] (gain $\sqrt{2}$ for hidden layers, 0.01 for the policy head), and layer normalisation on inputs. Actors produce tanh-squashed Gaussian policies with a learnable log-standard-deviation parameter initialised at -0.7 ($\sigma_0 \approx 0.5$). The centralised critic takes the global state vector $(t/T,\ \{S_t^i/\bar{S}^i,\ \tilde{d}_t^i,\ \tilde{g}_t^i\}_{i=1}^N,\ p_{t-1}^S/\Delta p,\ p_{t-1}^D/\Delta p)$ as input.

4.6. Training protocol

Table 3 summarises the training hyperparameters for all methods.

MANNC training. MANNC-DTDE and MANNC-DTDE (SG) policies are optimised with Adam (α=10^{-3}) without gradient clipping. Each episode simulates a full day (T=24 steps), after which each controllable agent's total daily cost is back-propagated through the environment dynamics and – in the case of MANNC-DTDE – through the market-clearing rule. Exploration follows an ϵ-greedy scheme: with probability $\rho_k = 0.7 \cdot 0.9^k$ (decaying from 70% to approximately 0.4% over 50 episodes), each action component is replaced by a uniform random sample from $[-1, 1]$. MANNC-CTCE follows the same protocol, except that a single centralised network outputs all agents' actions and the loss is the sum of all agents' costs (social welfare objective).

MAPPO training. MAPPO collects on-policy rollouts for all training days in the batch, yielding $T \times B = 24 \times 32 = 768$ transitions per agent per batch. Rollouts from all batches are aggregated before performing

15 PPO update epochs on the full dataset. Actor and critic are optimised separately with Adam (α=5 × 10^{-4}) and gradient clipping at max norm 10.0. A one-at-a-time sensitivity sweep over entropy coefficient, PPO epochs, mini-batch count, and initial policy standard deviation is reported in Appendix B.

Evaluation protocol. All trained policies are evaluated on the same fixed test set of 100 days. Battery SoC is initialised to zero at the start of each test day across all methods, ensuring comparable initial conditions. The rule-based heuristic requires no training and is evaluated directly on the test set.

4.7. Evaluation metrics

Community cost. The primary performance metric is the mean daily community cost on the test set:

$$C_{\text{total}} = \frac{1}{|D_{\text{test}}|} \sum_{d \in D_{\text{test}}} \sum_{i=1}^{N} C_d^i, \tag{25}$$

aggregating each agent's daily total cost (10) – including the terminal storage credit – over all agents and test days.

Per-agent cost. We report individual daily costs C^i to analyse distributional effects — specifically, whether learning based battery strategies shift costs onto agents without storage flexibility.

Unilateral deviation regret. We test whether the learned policies are close to a *Nash equilibrium* – whether any agent can lower its cost by deviating while the others hold fixed. A rollout gives each agent i a realised net-import trajectory $\mathbf{x}_i$; $\mathbf{x}^*_{-i}$ denotes the others' equilibrium trajectories (sampled from their policies – a Nash equilibrium is defined over policies, so these give a tractable proxy). With total cost $\text{Cost}_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \sum_{t=1}^{T} c_t^i - \underline{p}\, S_{T+1}^i$ (= C^i of eq. (10)) and the others frozen at $\mathbf{x}^*_{-i}$, the best-response cost and unilateral-deviation regret are

$$\text{Cost}^i_{\text{BR}} = \min_{\mathbf{x}_i} \text{Cost}_i(\mathbf{x}_i, \mathbf{x}^*_{-i}), \qquad \mathcal{R}^i = \text{Cost}_i(\mathbf{x}^*_i, \mathbf{x}^*_{-i}) - \text{Cost}^i_{\text{BR}} \;\geq\; 0. \tag{26}$$

Dynamic programming best-response. We solve agent i's best-response problem by dynamic programming (DP) over its battery schedule, with $\mathbf{x}^*_{-i}$ frozen and p_t recomputed at each (s, a) to include the deviator (a *strategic* best response); "exactly" is up to the discretisation of the state–action space (n_s=101 SoC and n_a=41 action points, PV fully used), whose effect is bounded in Appendix C.3. The recursion is run *after the entire trajectory has been observed* – it uses the realised future demand, generation, and others' actions – so it is a non-causal optimiser. With net grid trade $\text{net}_t^i(s,a) = d_t^i - g_t^i + b_t^{\text{grid}}(s,a)$ (the grid-side battery exchange of Eq. (3)) and terminal value $V_T(s) = -\underline{p}\, s$, the backward recursion

$$V_t(s) = \min_a \left[c_t^i\left(\text{net}_t^i(s,a),\, p_t(\text{net}_t^i(s,a), \mathbf{x}^*_{-i,t})\right) + V_{t+1}(s'(s,a)) \right] \tag{27}$$

gives $\text{Cost}^i_{\text{BR}} = V_0(0)$. Because this method optimises with full hindsight, $\text{Cost}^i_{\text{BR}} \leq \text{Cost}^i_{\text{BR,causal}}$ and hence $\mathcal{R}^i \geq \mathcal{R}^i_{\text{Nash}}$: the reported value upper-bounds the causal Nash regret, and a small $\mathcal{R}^i$ indicates proximity to equilibrium.

Training time. We report wall-clock training time measured on a single CPU (Apple M-series, no GPU) to quantify the computational trade-offs between differentiable and model-free training.

5. Experimental findings

All reported values are means ± one standard deviation over five random seeds. We first establish that the MANNC methods reduce cost and train efficiently (Section 5.1), verify convergence and numerical stability (Section 5.2), then analyse what the policies learn and why (Section 5.3), test generality across clearing mechanisms (Section 5.4), assess strategic stability via regret (Section 5.5), and validate on real Swiss smart-meter data (Section 5.6).

Table 4
Test-set performance (N=6, five seeds, SDR clearing). C_{total}: mean daily community cost; "Gap": excess over the perfect-foresight LP lower bound; training time on a single CPU (Apple M-series, no GPU). The Perfect-foresight LP is non-causal – it optimises with full knowledge of the day – and is a bound, not an achievable policy.

Method	C_{total} (EUR/day)	Gap (%)	Train (s)
MANNC-CTCE	9.34 ± 0.71	+3.9	≈75
MANNC-DTDE (SG)	9.73 ± 0.73	+8.2	≈61
MANNC-DTDE	10.06 ± 0.71	+11.9	≈121
MAPPO	11.22 ± 0.70	+24.8	≈184
Rule-based	11.51 ± 0.61	+28.0	0
Perfect-foresight LP	8.99 ± 0.72	–	–

5.1. Aggregate cost and training efficiency

Under the SDR clearing rule, the MANNC methods cut the mean daily community cost by 12.6–18.9% relative to the rule-based heuristic and outperform the model-free MAPPO baseline, training in one to three minutes on a single CPU (Table 4; N=6, five seeds). MANNC-CTCE, which has full state access, is cheapest (9.34 EUR/day); among the decentralised methods the price-taking MANNC-DTDE (SG) (9.73) edges out the price-aware MANNC-DTDE (10.06), and both beat MAPPO (11.22) and the heuristic (11.51).

A perfect-foresight LP that dispatches all batteries with full knowledge of the day (Section 4.4) reaches 8.99 EUR/day; MANNC-CTCE lies within 3.9% of this bound and the best decentralised policy within 8.2%, so little headroom remains. But the LP is a non-causal central planner: it does not operate through the endogenous price-formation mechanism and ignores incentive-compatibility.

The price-taking MANNC-DTDE (SG) attains a lower *community* cost than the price-aware MANNC-DTDE: by internalising their price impact, agents learn to extract surplus from others (Section 5.3), an effect specific to the clearing rule (Section 5.4).

MAPPO improves on the heuristic by only 2.5% and trains about three times slower than MANNC-DTDE (SG); the gap persists under systematic tuning (Appendix B), reflecting the learning signal rather than tuning. MANNC-DTDE is slower than MANNC-DTDE (SG) because its backward pass couples all agents through the shared price aggregates (Section 3.2).

Fig. 2 breaks cost down by archetype. The storage agent C (battery only) gains most, from 3.62 to 2.64–2.83 EUR/day via temporal arbitrage. Agent D (PV only) is cheapest under MANNC-CTCE (0.34), which times the other agents' charging to D's midday surplus. The pure consumer F pays more under price-aware MANNC-DTDE than price-taking MANNC-DTDE (SG) – a consequence of the capacity withholding analysed in Section 5.3.

5.2. Convergence and stability

Fig. 3 shows the training-batch community cost over episodes. The MANNC methods fall steeply in the first ten episodes and reach near-final cost within about 20 (MANNC-CTCE, MANNC-DTDE (SG)) to 30 (MANNC-DTDE) episodes. MAPPO starts lower – its Gaussian policy is initialised near zero action ("do nothing"), cheaper than the arbitrary dispatch of a randomly initialised network – but converges more slowly to a higher plateau, mirroring the test ranking of Section 5.1. The ±1 s.d. bands over the five seeds are narrow, indicating consistent convergence.

Numerical stability. The clearing rules contain non-smooth components – the $(\cdot)^+$ aggregation, the piecewise SDR price, and the $\min(\cdot, 1)$/clamp

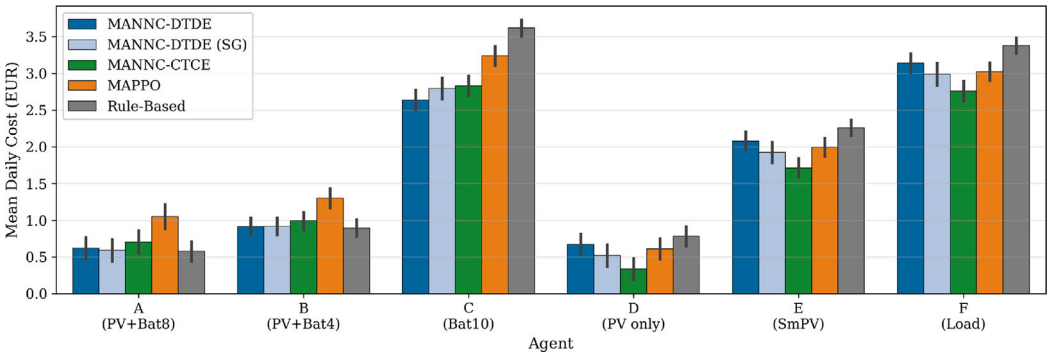


Fig. 2. Mean daily cost by prosumer archetype (N=6, SDR clearing). Error bars: ±1 standard deviation over five seeds.

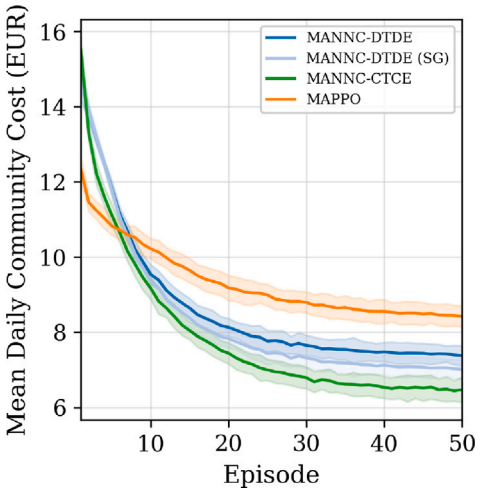


Fig. 3. Training convergence (N=6). Shaded regions: ±1 standard deviation over five seeds.

operators of MMR and LIN – yet gradient-based training is well-behaved. Across the SDR price's domain the automatic-differentiation gradient matches a finite-difference reference to within 4×10^{-3} (mean $8{\times}10^{-5}$), confirming the subgradients at the kinks are handled correctly. Over a full run (1,600 updates) the gradient norm stays bounded (mean 1.7, max 6.7) with zero NaN or infinite values, and the per-agent price sensitivity $|\partial p/\partial x|$ remains small (mean 0.016, max 0.60; Appendix C.1).

Exploration. The decaying ϵ-greedy noise gives a modest but consistent gain: MANNC-DTDE converges to 7.35 ± 0.22 EUR/day with exploration versus 7.46 ± 0.23 without (training-batch cost). The injected random actions raise the early-training curve but decay over training, helping agents escape suboptimal joint policies (Appendix C.2).

5.3. Learned strategies and the withholding mechanism

We now examine what the policies learn: their battery dispatch (Section 5.3.1), the resulting prices and trading flows (Section 5.3.2), the capacity-withholding effect of price-awareness (Section 5.3.3), and direct evidence that the price-gradient drives it (Section 5.3.4).

5.3.1. Battery dispatch

Fig. 4 shows the learned battery action as a function of hour-of-day and SoC for the three controllable agents. All learning methods recover a diurnal-arbitrage pattern. The PV+battery agents A and B charge around midday (06:00–16:00), when abundant PV depresses community prices, and discharge in the evening (17:00–22:00), with charge intensity rising in available PV and tapering as the battery fills. The pure-storage agent C, lacking PV, shows a fainter pattern, and MAPPO's C exhibits little structured arbitrage.

5.3.2. Market prices and trading flows

The learned policies produce a U-shaped price profile, see Fig. 5: internal prices approach the retail tariff $\bar{p}$ in the morning and evening and the feed-in tariff $\underline{p}$ at midday. The buyer–seller spread is narrowest under MANNC-CTCE and MANNC-DTDE (SG), indicating efficient internal matching, whereas the rule-based heuristic pins prices to the corridor extremes, having no price-smoothing behaviour. The per-agent net-import profiles show the controllable agents reshaping their flows – A and B charge in the morning, and C, despite owning no PV, becomes a midday buyer and evening seller – tracking the perfect-foresight optimiser's import/export shape closely, whereas MAPPO and the heuristic deviate more, see Fig. 6.

5.3.3. Capacity withholding

The per-agent breakdown of Section 5.1 localises the price-aware/price-taking gap. The storage agent C is *cheaper* under price-aware MANNC-DTDE than under price-taking MANNC-DTDE (SG) (2.64 vs. 2.79 EUR/day), while each passive buyer – D, E, F – pays about 0.15 EUR/
day more, for a net community increase of 0.34. Agent C achieves this by *withholding capacity*: it charges less and releases less into the evening demand peak, retaining more in storage at day's end and supplying ≈ 1 kWh less to the evening market (Appendix D), which sustains higher clearing prices on a smaller traded volume. This is textbook capacity withholding, learned purely from the price-impact gradient – which the next subsection isolates directly, and Section 5.4 shows to be specific to the clearing rule.

5.3.4. Direct evidence of gradient utilisation

To show directly that the stop-gradient changes the *learning signal* – not merely the reward – we compare the gradient with and without it on the same policy and data, see Table 5. On identical parameters and the identical batch we run two backward passes differing only in whether the cleared prices are detached: g_{full} (prices in the graph, MANNC-DTDE) and g_{direct} (prices detached, the MANNC-DTDE (SG) recipe); their difference $g_{\text{price}} = g_{\text{full}} - g_{\text{direct}}$ is exactly the signal the stop-gradient discards. At initialisation the two gradients coincide ($\cos = 1.00 \pm 0.00$, $\|g_{\text{price}}\| \approx 0.05$): while policies are naive, the stop-gradient is an exact approximation. Over training the discarded component grows about twentyfold – to $\|g_{\text{price}}\| \approx 1.2$–$1.3$, tight across seeds – becoming comparable in magnitude to the gradient the stop-gradient retains, while the two updates rotate apart (cosine falling from one towards zero). The price-impact gradient is thus real, sizeable, and increasingly distinct from the price-taking update: MANNC-DTDE demonstrably learns from information MANNC-DTDE (SG) discards.

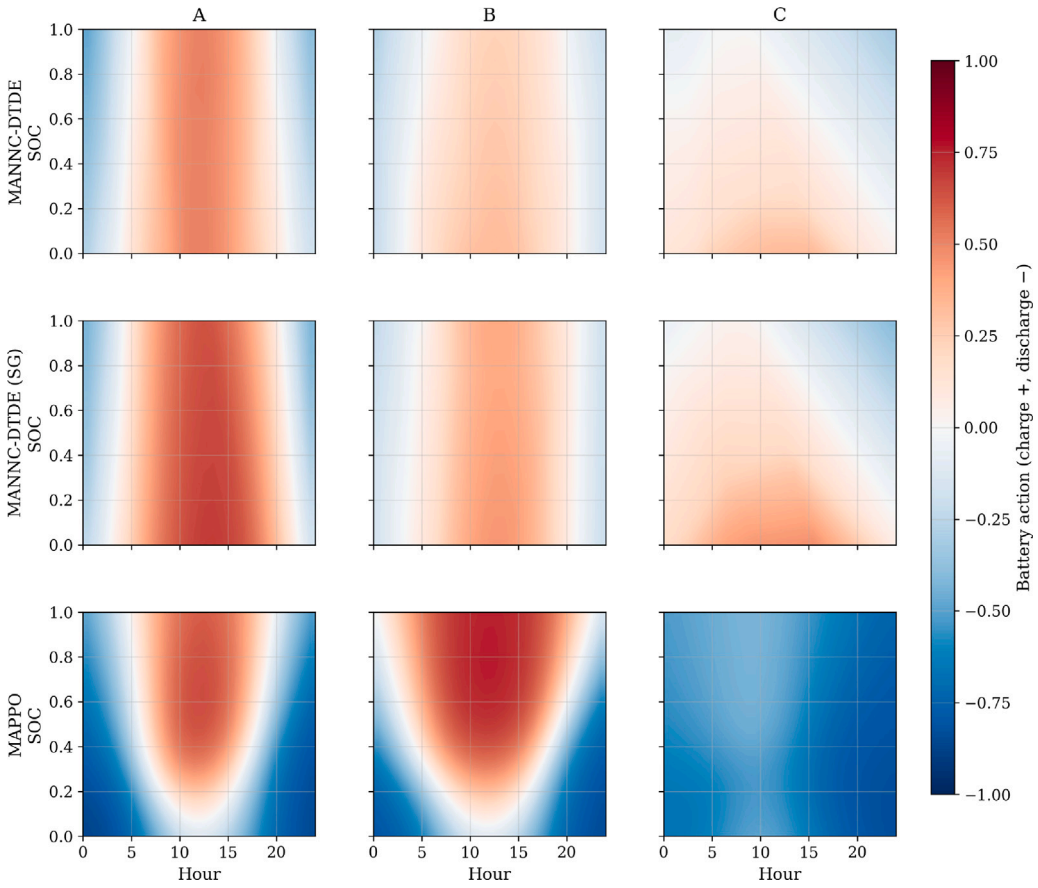


Fig. 4. Battery dispatch policies (hour × SoC) for the controllable agents A, B, C (columns) under MANNC-DTDE, MANNC-DTDE (SG), and MAPPO (rows). Red: charging; blue: discharging.

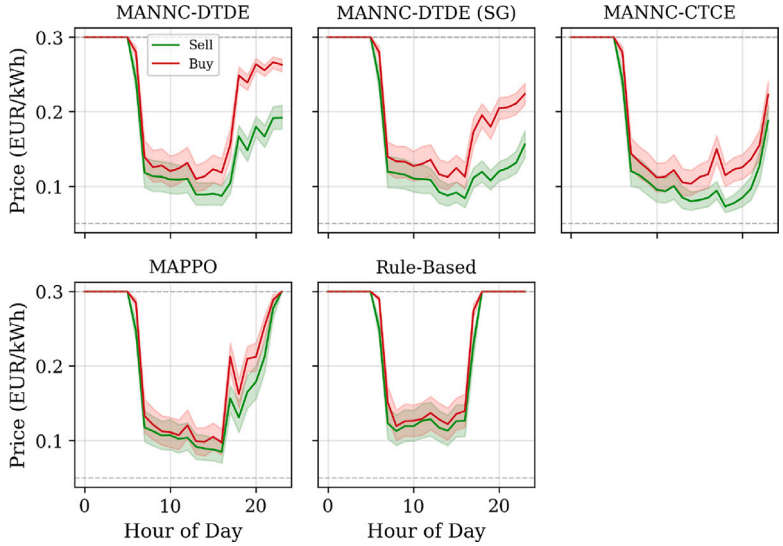


Fig. 5. Endogenous buyer/seller prices across the day. Dashed lines: tariff corridor $[\underline{p}, \bar{p}]$.

Section 5.4 confirms the causal link at the mechanism level by tuning a real clearing-rule sensitivity.

5.4. Generality across clearing mechanisms

Sections 5.1–5.3 used the SDR rule. We now repeat the comparison under all three clearing rules of Section 2.2. The three are nested: *MMR is exactly the β=0 member of the LIN family* (a constant mid-price), and LIN approaches an SDR-like full-corridor response as $\beta \to 1$, so the set spans price formation from a fixed internal price (MMR) to one that responds strongly to the supply–demand imbalance (SDR).

Table 6 reports community cost. MANNC-CTCE is lowest (≈ 9.35) and MAPPO and the heuristic highest under *every* rule, and all methods

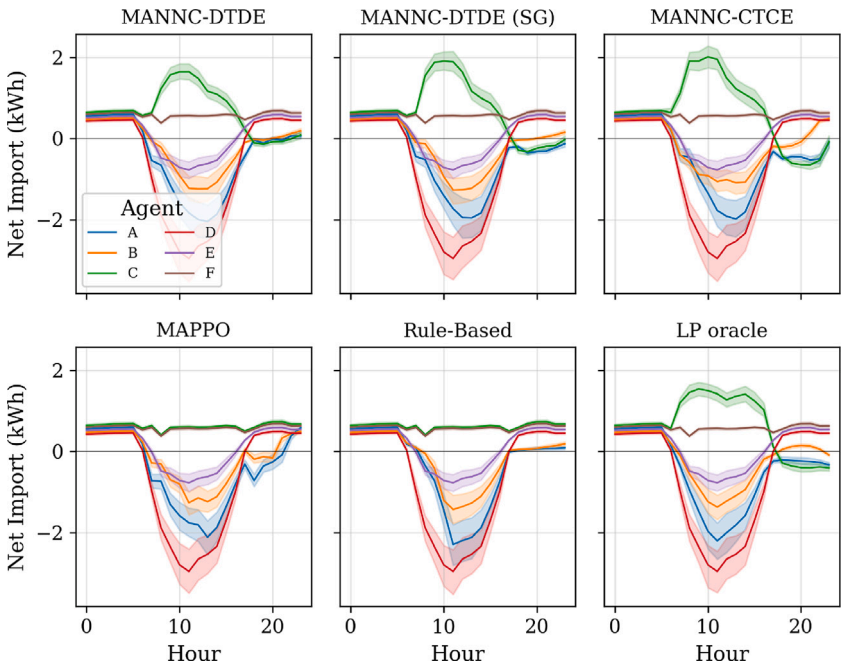


Fig. 6. Net-import profiles by agent and method, with the perfect-foresight LP optimiser as a reference panel. Positive: buying; negative: selling.

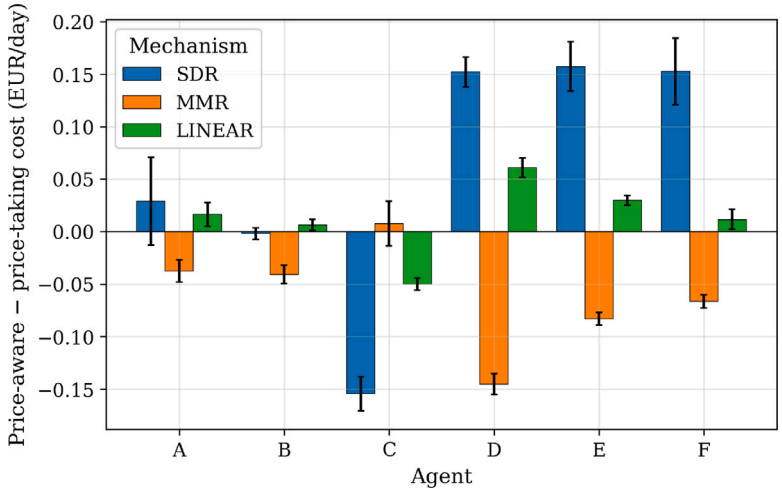


Fig. 7. Per-agent price-aware − price-taking cost gap by clearing rule. Negative = the agent gains from price-awareness. Under SDR the storage agent C gains at the passive agents' expense; under MMR C is roughly neutral and the passive agents gain.

Table 5
Norm of the price-impact gradient $g_{\text{price}} = g_{\text{full}} - g_{\text{direct}}$ and cosine between full and direct gradients at training checkpoints. Mean ± s.d. over five seeds. Late-training cosine variance is high because g_{full} is small near the optimum; $\|g_{\text{price}}\|$ is the robust quantity.

Episode	$\|g_{\text{price}}\|$ (dropped by SG)			$\cos(g_{\text{full}}, g_{\text{direct}})$		
	A	B	C	A	B	C
0	0.07 ± 0.01	0.04 ± 0.01	0.06 ± 0.13	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
9	0.75 ± 0.22	1.24 ± 0.24	0.63 ± 0.08	0.97 ± 0.03	0.98 ± 0.02	0.95 ± 0.04
24	1.13 ± 0.21	1.33 ± 0.20	1.14 ± 0.08	-0.14 ± 0.68	0.94 ± 0.08	-0.29 ± 0.32
49	1.17 ± 0.18	1.29 ± 0.18	1.30 ± 0.07	0.08 ± 0.53	0.30 ± 0.83	0.01 ± 0.70

stay above the mechanism-independent perfect-foresight bound. The same MANNC training runs unchanged under each rule – each admits a closed-form price gradient – so the method is not specific to SDR. What changes across rules is the strategic *behaviour*: the order of the two decentralised variants reverses, with MANNC-DTDE (SG) cheaper than MANNC-DTDE under SDR but dearer under MMR. That reversal is the withholding effect, which we now dissect.

The price-aware/price-taking gap $C_{\text{DTDE}} - C_{\text{SG}}$ (positive when price-awareness raises community cost) is $+0.34 \pm 0.05$ under SDR but -0.37 ± 0.02 under MMR and $+0.08 \pm 0.02$ under LIN(β=0.5): under MMR the sign flips. Fig. 7 shows why at the archetype level – under SDR the storage agent C profits (gap −0.15) while the passive buyers D, E, F each pay $\approx +0.15$; under MMR C's advantage vanishes (gap ≈ 0) and the passive agents gain.

The gradient itself is neutral: it tells an agent how its trades move the price it faces, and the mechanism sets whether acting on it helps or harms others. Under MMR the short side of the market receives the favourable mid-price while the long-side residual settles at the grid tariff, so the rule rewards being short – buying into community surplus,

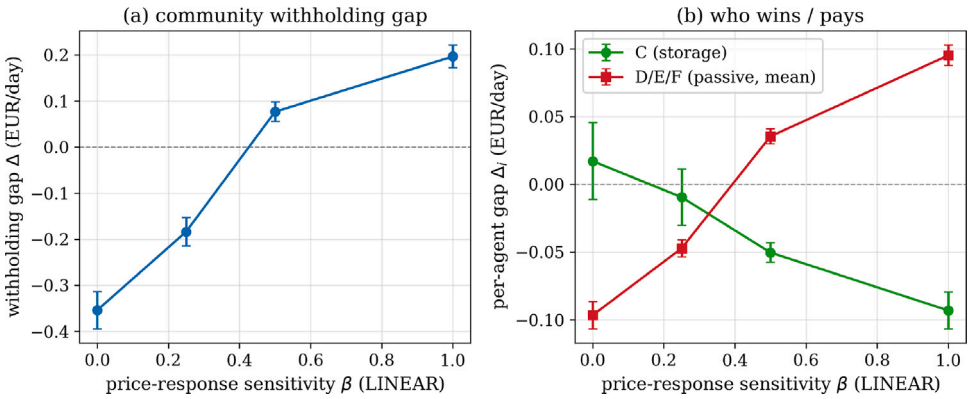


Fig. 8. LIN price-response sweep. (a) community withholding gap $\Delta = C_{\mathrm{DTDE}} - C_{\mathrm{DTDE(SG)}}$ versus the sensitivity β (β=0 is MMR); (b) per-agent gap Δ_i for the storage agent C versus the passive agents D/E/F. Mean ± s.d. over five seeds.

Table 6
Community cost (EUR/day, mean ± s.d. over five seeds) by method and clearing rule (LIN at β=0.5; MMR is LIN at β=0). The LP optimiser is mechanism-independent (internal trades net out at the community level).

Method	SDR	MMR	LIN
MANNC-CTCE	9.34 ± 0.71	9.35 ± 0.69	9.35 ± 0.70
MANNC-DTDE (SG)	9.73 ± 0.73	10.07 ± 0.72	9.59 ± 0.73
MANNC-DTDE	10.06 ± 0.71	9.70 ± 0.72	9.67 ± 0.73
MAPPO	11.22 ± 0.70	11.38 ± 0.65	11.26 ± 0.67
Rule-based	11.51 ± 0.61	11.51 ± 0.61	11.51 ± 0.61
Perfect-foresight LP	8.99 ± 0.72 (mechanism-independent)		

Table 7
DP unilateral-deviation regret $\mathcal{R}^i$ (EUR/day, mean ± s.d. over five seeds) for the controllable agents under SDR. The best-response oracle has full hindsight, so these upper-bound the causal Nash regret. Passive agents D–F have no flexibility and hence zero regret.

Method	A (PV+Bat8)	B (PV+Bat4)	C (Bat10)
MANNC-DTDE	0.19 ± 0.03	0.10 ± 0.01	0.24 ± 0.01
MANNC-DTDE (SG)	0.14 ± 0.02	0.09 ± 0.01	0.30 ± 0.01
MANNC-CTCE	0.22 ± 0.04	0.18 ± 0.03	0.35 ± 0.03
MAPPO	0.72 ± 0.04	0.55 ± 0.05	0.97 ± 0.02

selling into deficit. The price-aware C responds by cycling its battery markedly more (12.3 vs. 8.2 kWh/day), yet its own cost barely changes (per-agent gap ≈ 0): this is not cooperation. The extra balancing is a *positive externality* – the passive buyers D, E, F face lower prices and capture most of the −0.37 saving. Under SDR the same selfish gradient points the other way (restrict supply to lift the ratio-price), making the externality negative, so C profits and the others pay. The clearing rule fixes the *sign* of the externality that price-awareness creates, not whether agents act together.

The LIN family makes this a continuum through its price-response sensitivity β. Fig. 8 sweeps β from 0 to 1: the community gap rises monotonically from −0.35 at β=0 (where LIN is MMR, consistent with the standalone MMR value −0.37) through zero near $\beta \approx 0.4$ to +0.20 at β=1, and the per-agent split (panel b) crosses accordingly. The more the clearing price responds to aggregate imbalance, the stronger the incentive to withhold – the mechanism-level counterpart to the gradient evidence of Section 5.3.4.

These results clarify when withholding arises: it is a property of *quantity-sensitive* pricing (SDR, high-β LIN), not of price-aware learning per se. Rules that fix the internal price (MMR, β=0 LIN) turn the same gradient from a withholding incentive into a market-balancing one whose benefit spills over to the passive participants. Extending the analysis to auction-based mechanisms such as a double auction (which would require a differentiable relaxation of the discrete matching) is left to future work.

5.5. Strategic stability: regret

A low community cost matters only if the learned policies are stable – if no agent can improve by deviating unilaterally. We quantify this with the DP unilateral-deviation regret $\mathcal{R}^i$ of Section 4.7 (Eq. (26)): for each controllable agent, the gap between its equilibrium cost and the cost of its best response with the others held fixed. The best response is computed by dynamic programming over the realised trajectory and recomputes prices to include the deviator, so it is a non-causal, strategic oracle; $\mathcal{R}^i$ therefore upper-bounds the true causal Nash regret, and a small value indicates proximity to a Nash equilibrium. By construction $\mathcal{R}^i \geq 0$, and the three passive agents (no battery) have zero regret. All values are inflated by the oracle's hindsight; comparisons across methods and agents are the key quantity.

The decentralised MANNC policies are close to equilibrium: regret of 0.10–0.24 EUR/day for MANNC-DTDE (about 9% of the storage agent's daily cost) and 0.09–0.30 for MANNC-DTDE (SG). MAPPO is four-to-five times farther from Nash (0.55–0.97), likely because it learns from reward signals alone without access to the market structure, so its higher cost coincides with strategically unstable policies.

The price-aware and price-taking variants have the same average regret (0.18), but differ where it matters. For the pure-storage agent C – the only purely strategic actor – price-aware MANNC-DTDE has markedly lower regret than price-taking MANNC-DTDE (SG) (0.24 vs. 0.30). Because the best-response is strategic, this gap is the strategic margin that price-taking training leaves exposed: an agent that ignored its price impact during training is more exploitable by a strategic deviator. For the PV+battery agents A and B, which act less purely on price impact, the ordering is reversed and small.

The centralised planner MANNC-CTCE has positive regret for every battery owner (0.18–0.35), higher on average than the decentralised policies: the cost-minimising joint schedule is *not* a Nash equilibrium – each agent could cut its own cost by deviating. The same applies to the perfect-foresight LP of Section 5.1, likewise a cooperative, non-strategic solution. Low community cost and individual rationality are therefore distinct: the decentralised MANNC policies, each optimising only its own cost, settle closer to a self-enforcing equilibrium than the cooperative optimum does.

The best-response oracle is converged at the grid resolution used – refining it changes regret by under 0.003 EUR/day (Appendix C.3).

5.6. Real-world validation

The preceding sections used synthetic profiles. To test external validity we re-run the SDR comparison on real Swiss demand (Sec-

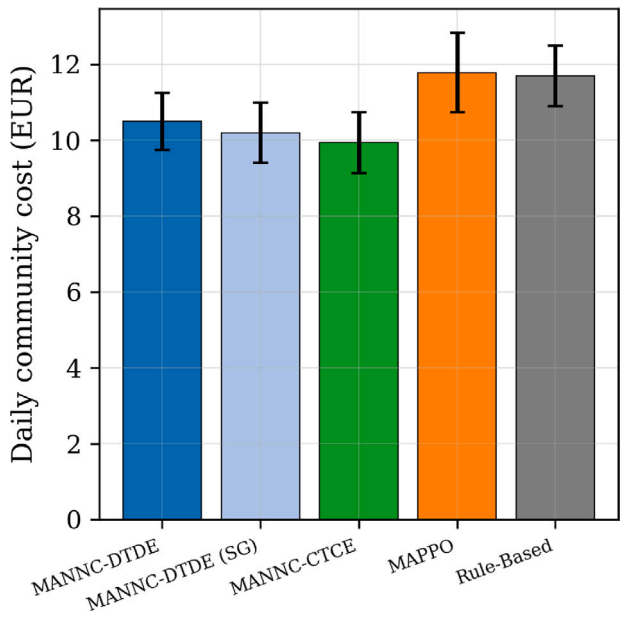


Fig. 9. Mean daily community cost across K=12 communities (CKW demand,PVGIS generation; SDR clearing). Error bars: ±1 s.d. across communities.

tion 4.2): K=12 communities, each formed by drawing one real CKW smart-meter household per archetype and pairing it with irradiance-modelled PVGIS generation, over 730 days (2021–2022). Costs are reported as mean ± s.d. across the twelve communities.

The method ranking is preserved on real data, see Fig. 9: MANNC-CTCE is cheapest (9.93 EUR/day), followed by the decentralised MANNC variants (MANNC-DTDE (SG) 10.19, MANNC-DTDE 10.50), with MAPPO (11.78) and the rule-based heuristic (11.69) clearly worse. The MANNC methods cut community cost by 10–15% over the heuristic on real demand, as on synthetic data; MAPPO, which narrowly beat the heuristic on synthetic data, no longer does.

The withholding signature also reproduces. Price-aware MANNC-DTDE costs more than price-taking MANNC-DTDE (SG) by +0.31 ± 0.19 EUR/day – positive in *all twelve* communities and close to the synthetic value (+0.34). The strategic effect of Section 5.3 is therefore not an artefact of the synthetic generator: it emerges under real load and weather variability too.

This is a real-*demand* validation rather than a deployment study. The demand is measured, but – because the CKW meters record consumption only – PV generation is irradiance-modelled (PVGIS), and battery sizing follows the archetype definitions. The results thus hold for real demand and realistic generation in a single region (canton Lucerne); pairing measured rooftop PV with measured demand, and varying region and storage, is left to future work (Section 6.5).

6. Discussion

6.1. Implications for platform design

Two findings from Section 5 carry direct implications for P2P platform design: the welfare effect of price-aware learning, and the incentive-compatibility of centrally planned dispatch.

First, the comparison between MANNC-DTDE (DTDE, endogenous prices) and MANNC-DTDE (SG) (DTDE, exogenous prices) isolates the welfare effect of agents internalising their price impact (the welfare *cost* arises under SDR; under MMR the same gradient produces a welfare *gain*, see Section 5.4).

Under SDR, internalising price impact triggers capacity withholding and raises community cost relative to the price-taking case (Table 4). Per-agent price impact diminishes with community size – each trade enters the clearing rule only through the aggregate, so individual influence is diluted as N grows – but the aggregate SDR cost gap stays at 2.5–3.8% across $N \in \{6, 12, 24, 48, 96\}$ as a fraction of MANNC-DTDE community cost (Table A.1). From a platform-design perspective, these two observations point to different forms of oversight. As community size increases, each individual agent's marginal price impact is diluted, reducing the relevance of interventions aimed at individual market power. At the same time, the persistence of the aggregate welfare gap suggests that mechanism-level monitoring remains important, since small individual incentives can still generate material community-level distortions. Under MMR the picture reverses (Section 5.4): the same gradient produces a positive externality and the gap turns into a welfare gain, so neither monitoring tool is warranted.

Second, the regret analysis, see Table 7, reveals that the socially optimal allocation produced by the centralised planner is not incentive-compatible: every battery owner could reduce its cost by 0.18–0.35 EUR/day through unilateral deviation. Platforms that rely on centralised dispatch instructions are therefore vulnerable to strategic non-compliance unless accompanied by binding commitment or compensation mechanisms. In contrast, the decentralised MANNC variants exhibit substantially lower regret ($\mathcal{R}^i \in [0.09, 0.30]$ EUR/day), indicating they approximate a Nash equilibrium and are self-enforcing without external coordination.

6.2. Centralised versus decentralised learning

In terms of incentive compatibility, both decentralised MANNC variants are close to a Nash equilibrium (regret 0.09–0.30 EUR/day, averaging ≈ 0.18 across the controllable agents), with MANNC-DTDE closer for the pure-storage agent C and MANNC-DTDE (SG) marginally closer for the PV+battery agents A and B. MANNC-CTCE has positive regret for every battery owner (0.18–0.35), confirming that the centrally planned schedule is not incentive-compatible. MAPPO has the highest regret (0.55–0.97 EUR/day), likely because it learns from reward signals alone without access to the market structure; its community cost is also ≈ 12% higher than MANNC-DTDE's and ≈ 15% higher than MANNC-DTDE (SG)'s. From a deployment perspective the two decentralised variants are a trade-off rather than a dominance ordering: MANNC-DTDE (SG) yields the lowest community cost among the decentralised methods, while MANNC-DTDE is closer to the strategic best-response of the pure-storage agent. Both are approximately self-enforcing.

In terms of computational cost, MANNC-DTDE (SG) and MANNC-CTCE scale linearly with N, see Appendix A, while MANNC-DTDE scales super-linearly due to the coupling of agents' gradients through the clearing rule. At N=96, MANNC-DTDE is approximately twenty times slower than MANNC-DTDE (SG). MAPPO scales linearly but with a larger constant. These observations motivate the stop-gradient approximation for communities exceeding approximately 24 agents, where the cost-performance difference between MANNC-DTDE and MANNC-DTDE (SG) is small relative to the computational savings.

In terms of information requirements, MANNC-CTCE requires centralised access to all agents' states, which raises privacy concerns and creates a single point of failure. MAPPO requires centralised training (shared value function) but permits decentralised execution. Both MANNC-DTDE variants require only local observations and the publicly broadcast clearing prices, making them compatible with privacy-preserving platform architectures.

6.3. Strategic behaviour and welfare redistribution

The capacity-withholding mechanism documented in Section 5.3.3 connects the MANNC framework to the broader literature on market power in electricity markets. In oligopolistic wholesale markets,

generators with inframarginal capacity can increase profits by restricting output, a phenomenon well documented in both theoretical and empirical studies [45–47]. An analogous mechanism arises endogenously in small-scale P2P markets when battery owners have access to price-impact information.

The impacts on individual costs are asymmetric. Battery owners (who can shift load intertemporally) benefit from the strategic premium, while passive consumers and prosumers without storage bear the cost through higher purchase prices, see Fig. 2, Agents D, E, and F. This distributional effect is relevant for regulators concerned with fairness in energy communities, particularly under EU directives that promote citizen participation in local energy markets. This effect is not inherent to price-aware learning but depends on the clearing rule: the mechanism sweep of Section 5.4 shows that withholding arises under quantity-sensitive pricing (SDR, high-β LIN) but reverses under fixed-price rules (MMR), where the same gradient produces a positive externality for passive participants.

Per-agent price impact diminishes with N – a single agent's trade enters only through the aggregate volume – but the community-level withholding gap stays at 2.5–3.8% (as a fraction of MANNC-DTDE community cost) across $N \in \{6, 12, 24, 48, 96\}$ under SDR (Table A.1). Under heterogeneous scaling, each agent's influence shrinks but the strategic actors and the agents bearing the resulting price distortion grow in proportion, so the aggregate distortion persists.

6.4. Comparison with model-free MARL

MAPPO serves as a model-free baseline. The MANNC methods reduce community cost by 12.6–18.9% over the heuristic versus 2.5% for MAPPO, see Table 4, and train faster. The gap reflects the learning signal: pathwise gradients through the clearing rule provide richer per-step information than scalar rewards, leading to faster convergence and more effective arbitrage. MAPPO's advantage is applicability to proprietary mechanisms where no model of the price-formation process is available.

6.5. Limitations

The framework requires a differentiable clearing rule; discrete matching mechanisms such as strict merit-order stacks would need smooth relaxations.

Absence of network constraints. The model treats the community as a single electrical node without grid constraints. In practice, voltage limits, line capacities, and network losses affect feasible trading outcomes. Ignoring these constraints may overstate the gains from P2P trading relative to grid-mediated exchange.

Training can converge to local optima, a known limitation of gradient-based policy search in the deep hedging architecture [38]. Extending the framework to stochastic policies may improve robustness.

The DP regret values are upper bounds due to the oracle's hindsight; discretisation error is below 0.003 EUR/day, see Appendix C.3.

Other practical constraints. The model also omits regulatory aspects (settlement rules, ex-post market-power mitigation), communication delays in the broadcast of cleared prices, and uncertainty in the published price signals themselves. Each would tighten the action space or perturb the observation channel; we treat them as future work rather than incorporate stylised versions here, to keep the strategic conclusions clean.

Regional scope of the real-data validation. The CKW case study uses smart-meter demand from a single Swiss region (canton Lucerne) paired with irradiance-modelled PVGIS generation and an assumed battery sizing (Section 5.6). The framework is region-agnostic – tariff corridor, demand profile, and asset mix are all parameters – but a multi-region evaluation with paired measured rooftop PV remains future work.

7. Conclusion

This paper introduced a learning framework for decentralised peer-to-peer electricity trading in which the platform's market clearing rule is embedded directly in the multi-agent training simulator. Each prosumer's policy is optimised end-to-end by backpropagation through a complete trajectory of physical dynamics and price formation, so that the gradient an agent receives reflects its own impact on the prices faced by all participants. A stop-gradient operator at the clearing prices isolates this price-impact signal, separating strategic learning – which internalises price impact – from competitive learning, which treats prices as exogenous. The same training graph admits three non-discriminatory clearing rules – a supply-to-demand ratio, a mid-market rate, and a linear pricing rule – in closed form, so the approach is not tied to a particular mechanism.

On a stylised six-prosumer community in a central-European setting, the pathwise methods reduce mean daily community cost by 12.6 to 18.9 per cent relative to a self-consumption heuristic, against 2.5 per cent for a representative multi-agent reinforcement-learning baseline, and train in one to three minutes on a single laptop CPU. The decentralised policies approximate a Nash equilibrium under a non-causal best-response oracle, with average unilateral-deviation regret of about 0.18 euro per day per controllable agent; the reinforcement-learning baseline is three to four times farther from equilibrium. Consistent with the general observation that learning through a known model is more sample-efficient than reward-only learning, embedding the price-formation mechanism in the training graph yields both lower cost and stronger strategic stability than treating it as a black box. Endogenous capacity withholding emerges from the price-impact gradient under quantity-sensitive clearing and reverses under fixed-price clearing – so the direction of the welfare effect of strategic learning is set by the clearing rule, not by the gradient itself. Method ordering and the withholding signature both reproduce on real Swiss smart-meter demand for twelve independently sampled communities.

CRediT authorship contribution statement

Nicolas Greber: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Nicolas Eschenbaum:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Oleg Szehr:** Writing – review & editing, Validation, Software, Methodology, Conceptualization.

Funding

This work was supported by the Swiss Federal Office of Energy (SFOE) under grant number SI/502525-01.

Declaration of competing interest

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

Appendix A. Scalability

We test scaling along two axes. *Community cost* is measured under *heterogeneous scaling*: for each $N \in \{6, 12, 24, 48, 96\}$ we draw three independent communities by sampling each agent's PV capacity, battery size, and average demand from the archetype ranges (yielding $\approx 50\%$ controllable agents), so per-agent cost is comparable across N (Table A.1). *Compute and memory* are measured under *replication scaling*, where the six-archetype base unit is replicated to the target N; this fixes the per-agent composition and isolates the algorithmic complexity from the cost trend (Table A.2).

Table A.1
Mean daily community cost (EUR/day, mean ± s.d. over three heterogeneous community draws) versus community size N under SDR clearing.

N	MANNC-DTDE	MANNC-DTDE (SG)	Rule-based	Perfect-foresight LP
6	10.06 ± 1.09	9.81 ± 1.32	11.17 ± 1.69	9.14 ± 1.28
12	21.67 ± 2.06	21.09 ± 2.15	23.50 ± 2.11	19.91 ± 2.37
24	44.08 ± 3.44	42.42 ± 3.61	48.83 ± 4.03	39.90 ± 3.47
48	86.90 ± 3.20	83.71 ± 3.79	97.25 ± 3.65	77.46 ± 3.61
96	175.16 ± 7.09	170.43 ± 7.29	195.60 ± 7.52	159.01 ± 6.30

Table A.2
Per-batch compute and memory under replication scaling (single CPU). The price-aware backward pass couples all agents through the shared price aggregates, so its cost grows quadratically with N; the stop-gradient, centralised, and model-free backward passes decouple across agents and grow linearly.

	N=6	N=12	N=24	N=48	N=96
MANNC-DTDE backward (s)	0.042	0.153	0.603	2.413	9.874
MANNC-DTDE (SG) backward (s)	0.011	0.031	0.053	0.095	0.190
MANNC-DTDE total (s)	0.064	0.194	0.682	2.570	10.19
MANNC-DTDE (SG) total (s)	0.032	0.073	0.144	0.243	0.489
MANNC-CTCE total (s)	0.040	0.076	0.149	0.314	0.697
MAPPO total (s)	0.085	0.153	0.288	0.563	1.130
MANNC-DTDE peak memory (MB)	369	375	390	419	490

Across all five sizes the MANNC methods cut community cost by 7.8 to 13.9% relative to the rule-based heuristic, and the withholding gap $C_{\mathrm{DTDE}} - C_{\mathrm{SG}}$ stays between 2.5 and 3.8% of community cost. The gap does not shrink with N: as the community grows, each agent's share of the price-forming volume falls, but the number of strategic actors and the number of agents bearing the resulting price distortion grow in proportion, leaving the aggregate distortion roughly constant.

The forward pass is linear in N for every method (prices depend only on the two aggregates M_t, Q_t). The price-aware backward pass, however, scales quadratically: over a 16× increase in N its cost rises ≈ 236× (0.042 → 9.87 s), consistent with the quadratic scaling derived in Section 3.2, whereas the stop-gradient backward pass rises only ≈ 17× (linear). At N=96 the price-aware step is therefore ≈ 21× slower than the price-taking one (10.2 vs. 0.49 s). Peak memory grows modestly – +120 MB for MANNC-DTDE across the range – so memory is not the binding constraint; compute is. Because each agent's trade enters the clearing rule only through the aggregate, individual price impact diminishes as the community grows, making the strategic effects of Section 5.3 most pronounced in small communities even though the aggregate withholding gap (Table A.1) persists at every tested size. For large communities the price-taking variant is the practical choice.

Appendix B. MAPPO hyperparameter tuning

MAPPO's underperformance (Section 5.1) is robust to hyperparameters. Starting from the reported configuration (γ=1, entropy 0.05, 15 PPO epochs, full-batch updates, initial policy std $\sigma_0 \approx 0.5$) we vary one setting at a time (Table B.1); each value is the mean daily community cost over five seeds.

No tuned setting moves MAPPO below 11.15 EUR/day. The initial policy std is near-optimal at the reference value: shrinking it to σ_0=0.25 leaves cost unchanged, while enlarging it to $\sigma_0 \in \{1.0, 1.5\}$ makes convergence within the 50-episode budget *worse* — so the reviewer-flagged Gaussian initialisation is not a confound. Across the entire sweep MAPPO remains $\gtrsim$ 1.4 EUR/day above MANNC-DTDE (SG); the deficit reflects the learning signal (reward-only versus pathwise gradients), not tuning.

Appendix C. Numerical and approximation diagnostics

Table B.1
MAPPO cost (EUR/day, mean ± s.d. over five seeds) under one-at-a-time hyperparameter changes from the reported configuration, including the initial policy std σ_0. Every variant stays ≈ 15–30% above the best decentralised method, MANNC-DTDE (SG) (9.73).

Configuration	C_{total} (EUR/day)
Reference (γ=1, ent. 0.05, 15 epochs, 1 minibatch, σ_0≈0.5)	11.21 ± 0.69
Entropy coefficient 0.0	11.31 ± 0.69
Entropy coefficient 0.01	11.25 ± 0.70
Entropy coefficient 0.1	11.31 ± 0.69
PPO epochs 5	11.26 ± 0.71
PPO epochs 30	11.15 ± 0.71
Minibatches 4	11.17 ± 0.69
Initial std $\sigma_0 = 0.25$	11.20 ± 0.71
Initial std $\sigma_0 = 1.0$	11.69 ± 0.59
Initial std $\sigma_0 = 1.5$	12.69 ± 0.56

Table C.1
DP best-response cost and regret (Agent C, EUR/day) versus grid resolution. The reported grid $(101, 41)$ is converged to within 0.003 EUR/day of a 4×-finer grid.

n_s	n_a			
	21	41	81	161
51	0.2207	0.2267	0.2284	0.2290
101	0.2216	**0.2276**	0.2292	0.2292
201	0.2220	0.2279	0.2295	0.2300
401	0.2220	0.2280	0.2296	0.2301

C.1. Gradient stability

Despite the non-smooth clearing components, the pathwise gradients are well-behaved. Across the SDR price's domain the automatic-differentiation gradient matches a finite-difference reference to within 4×10^{-3} (mean 8×10^{-5}), confirming the subgradients at the kinks are computed correctly. Over a full training run (1,600 updates) the gradient norm stays bounded (mean 1.7, max 6.7) with zero NaN or infinite values, and the per-agent price sensitivity $|\partial p/\partial x|$ remains small (mean 0.016, max 0.60).

C.2. Exploration-noise ablation

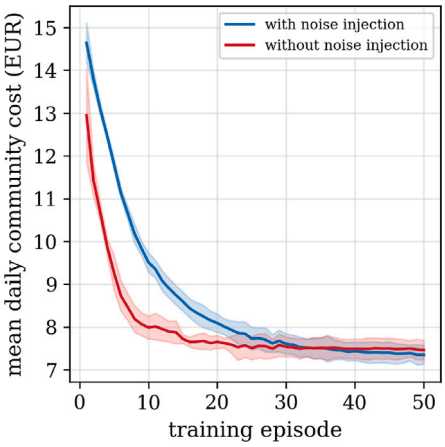
The decaying ϵ-greedy schedule yields a modest, consistent gain (Fig. C.1(a)): MANNC-DTDE converges to 7.35 ± 0.22 EUR/day with exploration versus 7.46±0.23 without (training-batch cost). The injected random actions inflate the early-training curve but decay over training, helping agents escape suboptimal joint policies.

C.3. DP best-response discretisation

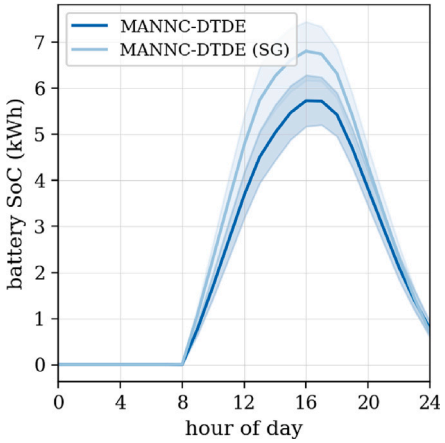
The regret in Section 5.5 uses a dynamic-programming best-response on a discretised state–action grid. Table C.1 shows it is converged: refining the grid 4× in each dimension beyond the reported (n_s, n_a)=(101, 41) changes the regret by under 0.003 EUR/day (≈ 1%), so the discretisation error is negligible relative to the reported values.

Appendix D. Supplementary strategy results

Fig. C.1(b) shows Agent C's mean SoC trajectory under price-aware and price-taking learning, the mechanism behind the capacity withholding of Section 5.3.3: the price-aware agent peaks lower (5.85 vs. 6.52 kWh) and retains more at day's end (1.00 vs. 0.72 kWh), supplying ≈ 1 kWh less to the evening market.



(a) Training convergence with and without ϵ-greedy exploration (MANNC-DTDE, N=6). Shaded: ±1 s.d. over five seeds.



(b) Mean SoC trajectory of Agent C under price-aware MANNC-DTDE and price-taking MANNC-DTDE (SG). Shaded: ±1 s.d.

Fig. C.1. Exploration ablation (a) and capacity-withholding mechanism (b).

Data availability

Data will be made available on request.

References

[1] Han J, E W. Deep learning approximation for stochastic control problems. 2016, http://dx.doi.org/10.48550/arXiv.1611.07422, arXiv preprint arXiv:1611.07422.

[2] Buehler H, Gonon L, Teichmann J, Wood B. Deep hedging. Quant Finance 2019;19:1271–91. http://dx.doi.org/10.1080/14697688.2019.1571683.

[3] Tushar W, Saha TK, Yuen C, Smith D, Poor HV. Peer-to-peer trading in electricity networks: An overview. IEEE Trans Smart Grid 2020;11:3185–200. http://dx.doi.org/10.1109/TSG.2020.2969657.

[4] Khorasany M, Mishra Y, Ledwich G. Market framework for local energy trading: A review of potential designs and market clearing approaches. IET Gener Transm Distrib 2018;12:5899–908. http://dx.doi.org/10.1049/iet-gtd.2018.5309.

[5] Long C, Wu J, Zhang C, Thomas L, Cheng M, Jenkins N. Peer-to-peer energy trading in a community microgrid. In: 2017 IEEE power & energy society general meeting. IEEE; 2017, p. 1–5. http://dx.doi.org/10.1109/PESGM.2017.8274546.

[6] Liu N, Yu X, Wang C, Li C, Ma L, Lei J. Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers. IEEE Trans Power Syst 2017;32:3569–83. http://dx.doi.org/10.1109/TPWRS.2017.2649558.

[7] Zhang S, May D, Gül M, Musilek P. Reinforcement learning-driven local transactive energy market for distributed energy resources. Energy AI 2022;8:100150. http://dx.doi.org/10.1016/j.egyai.2022.100150.

[8] Weidlich A, Veit D. A critical survey of agent-based wholesale electricity market models. Energy Econ 2008;30:1728–59. http://dx.doi.org/10.1016/j.eneco.2008.01.003.

[9] Harder N, Miskiw KK, Khanra M, Maurer F, Patil P, Qussous R, Weinhardt C, Klobasa M, Ragwitz M, Weidlich A. Assume: An agent-based simulation framework for exploring electricity market dynamics with reinforcement learning. SoftwareX 2025;30:102176. http://dx.doi.org/10.1016/j.softx.2025.102176.

[10] Harder N, Qussous R, Weidlich A. Fit for purpose: Modeling wholesale electricity markets realistically with multi-agent deep reinforcement learning. Energy AI 2023;14:100295. http://dx.doi.org/10.1016/j.egyai.2023.100295, URL https://www.sciencedirect.com/science/article/pii/S2666546823000678.

[11] Harder N, Weidlich A, Staudt P. Finding individual strategies for storage units in electricity market models using deep reinforcement learning. Energy Inform 2023;6:41. http://dx.doi.org/10.1186/s42162-023-00293-0.

[12] Harder N, Mitridati L, Pourahmadi F, Weidlich A, Kazempour J. How satisfactory can deep reinforcement learning methods simulate electricity market dynamics? benchmarking via bi-level optimization. ACM SIGENERGY Energy Inform Rev 2024;4:65–77. http://dx.doi.org/10.1145/3717413.3717419.

[13] Miskiw KK, Staudt P. Explainable deep reinforcement learning for multi-agent electricity market simulations. Proceedings of the 20th international conference on the European energy market, EEM. Istanbul, Turkey: IEEE; 2024, http://dx.doi.org/10.1109/EEM60825.2024.10608907.

[14] May DC, Taylor M, Musilek P. Decentralised coordination of distributed energy resources through local energy markets and deep reinforcement learning. Energy AI 2024;18:100446. http://dx.doi.org/10.1016/j.egyai.2024.100446, URL https://www.sciencedirect.com/science/article/pii/S2666546824001125.

[15] Ye Z, Qiu D, Li S, Fan Z, Strbac G. Federated reinforcement learning for decentralized peer-to-peer energy trading. Energy AI 2025;20:100500. http://dx.doi.org/10.1016/j.egyai.2025.100500.

[16] Ye Y, Papadaskalopoulos D, Yuan Q, Tang Y, Strbac G. Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets. IEEE Trans Smart Grid 2023;14:1541–54. http://dx.doi.org/10.1109/TSG.2022.3149266.
[17] Zhang K, Yang Z, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Springer; 2021, p. 321–84.
[18] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: A survey. Artif Intell Rev 2022;55:895–943. http://dx.doi.org/10.1007/s10462-021-09996-w.
[19] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. 2015, http://dx.doi.org/10.48550/arXiv.1509.02971, arXiv preprint arXiv:1509.02971.
[20] Szehr O. Hedging of financial derivative contracts via Monte Carlo tree search. J Comput Finance 2023. http://dx.doi.org/10.21314/jcf.2023.009.
[21] Cannelli L, Nuti G, Sala M, Szehr O. Hedging using reinforcement learning: Contextual k-armed bandit versus Q-learning. J Financ Data Sci 2023;9:100101. http://dx.doi.org/10.1016/j.jfds.2023.100101.
[22] Krabichler T, Teichmann J. A case study for unlocking the potential of deep learning in asset-liability-management. Front Artif Intell 2023;6:1177702. http://dx.doi.org/10.3389/frai.2023.1177702.
[23] Jaisson T. Deep differentiable reinforcement learning and optimal trading. Quant Finance 2022;22:1429–43. http://dx.doi.org/10.1080/14697688.2022.2062431.
[24] Tschora L, Guns T, Pierre E, Plantevit M, Robardet C. Electricity price forecasting based on order books: a differentiable optimization approach. In: 2023 IEEE 10th international conference on data science and advanced analytics. DSAA, Thessaloniki, Greece: IEEE; 2023, p. 1–10. http://dx.doi.org/10.1109/DSAA60987.2023.10302542.
[25] Štrupl M, Szehr O, Faccio F, Ashley DR, Srivastava RK, Schmidhuber J. On the convergence and stability of upside-down reinforcement learning, goal-conditioned supervised learning, and online decision transformers. 2025, http://dx.doi.org/10.48550/arXiv.2502.05672, arXiv preprint arXiv:2502.05672.
[26] Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the 31st international conference on machine learning, ICML. 2014, p. 387–95.
[27] Szehr O, Wolf MM. Connected components of irreducible maps and 1D quantum phases. J Math Phys 2016;57(8). http://dx.doi.org/10.1063/1.4960557.
[28] Hasan M, Zarin NA, Ahmed MR, Farrok O. Global pathways for hybrid renewable energy systems: Challenges, solutions, policy, and regulatory frameworks. Energy Convers Manag: X 2026;29:101534. http://dx.doi.org/10.1016/j.ecmx.2026.101534.
[29] Masroor MI, Sadaf F, Hossain M, Hasan M, Chowdhury N-U-R. A proposed optimization strategy for energy savings in residential sector. SSRN preprint, 2024, http://dx.doi.org/10.2139/ssrn.4992736.
[30] Janefar S, Chowdhury P, Yeassin R, Hasan M, Chowdhury N-U-R. Comparative performance evaluation of economic load dispatch using metaheuristic techniques: A practical case study for Bangladesh. Results Eng 2025;26:104720. http://dx.doi.org/10.1016/j.rineng.2025.104720.
[31] Talukder K, Zarin NA, Nobonita NR, Nawreen N, Sobhan S, Salsabil NA, Redwan R, Hasan M. Towards sustainable lithium-ion battery recycling in Bangladesh: Assessing environmental aspects and future prospects. Energy Nexus 2026;100698.
[32] Wang S, Ma C, Gao H, Deng D, Fernandez C, Blaabjerg F. Improved hyperparameter Bayesian optimization-bidirectional long short-term memory optimization for high-precision battery state of charge estimation. Energy 2025;328:136598. http://dx.doi.org/10.1016/j.energy.2025.136598.
[33] Wang S, Fan Y, Jin S, Takyi-Aninakwa P, Fernandez C. Improved anti-noise adaptive long short-term memory neural network modeling for the robust remaining useful life prediction of lithium-ion batteries. Reliab Eng Syst Saf 2023;230:108920. http://dx.doi.org/10.1016/j.ress.2022.108920.
[34] Wang S, Gao H, Takyi-Aninakwa P, Guerrero JM, Fernandez C, Huang Q. Improved multiple feature-electrochemical thermal coupling modeling of lithium-ion batteries at low-temperature with real-time coefficient correction. Prot Control Mod Power Syst 2024;9:157–73. http://dx.doi.org/10.23919/PCMP.2023.000257.
[35] Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the eleventh international conference on machine learning. 1994, p. 157–63. http://dx.doi.org/10.1016/B978-1-55860-335-6.50027-1.
[36] Oliehoek FA, Amato C. A concise introduction to decentralized POMDPs, vol. 1. Springer; 2016, http://dx.doi.org/10.1007/978-3-319-28929-8.
[37] Eschenbaum N, Mellgren F, Zahn P. Robust algorithmic collusion. 2022, http://dx.doi.org/10.48550/arXiv.2201.00345, arXiv preprint arXiv:2201.00345.
[38] Maggiolo M, Nuti G, Štrupl M, Szehr O. Deep hedging under non-convexity: Limitations and a case for AlphaZero. 2025, http://dx.doi.org/10.48550/arXiv.2510.01874, arXiv preprint arXiv:2510.01874.
[39] Yu C, Velu A, Vinitsky E, Gao J, Wang Y, Bayen A, Wu Y. The surprising effectiveness of ppo in cooperative multi-agent games. 2022, p. 24611–24.
[40] Ableitner L, Tiefenbeck V, Meeuw A, Wörner A, Fleisch E, Wortmann F. User behavior in a real-world peer-to-peer electricity market. Appl Energy 2020;270:115061. http://dx.doi.org/10.1016/j.apenergy.2020.115061.
[41] Statistisches Bundesamt (Destatis). Electricity consumption of private households by household size. 2023, URL https://www.destatis.de/EN/Themes/Society-Environment/Environment/Environmental-Economic-Accounting/private-households/Tables/electricity-consumption-private-households.html. [Accessed 2024].
[42] CKW AG. Swiss smart meter data – CKW 2021/2022 – anonymized individual metering points. Zenodo. 2023, http://dx.doi.org/10.5281/zenodo.7828796.
[43] European Commission, Joint Research Centre. PVGIS – Photovoltaic geographical information system. 2022, URL https://re.jrc.ec.europa.eu/pvg_tools/.
[44] Saxe A, McClelland J, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In: Proceedings of the international conference on learning representations 2014. International Conference on Learning Representations 2014; 2014.
[45] Wolfram CD. Measuring duopoly power in the British electricity spot market. Am Econ Rev 1999;89:805–26. http://dx.doi.org/10.1257/aer.89.4.805.
[46] Borenstein S, Bushnell JB, Wolak FA. Measuring market inefficiencies in California's restructured wholesale electricity market. Am Econ Rev 2002;92:1376–405. http://dx.doi.org/10.1257/000282802762024557.
[47] Hortaçsu A, Puller SL. Understanding strategic bidding in multi-unit auctions: A case study of the Texas electricity spot market. Rand J Econ 2008;39:86–114. http://dx.doi.org/10.1111/j.0741-6261.2008.00005.x.