DigIPlat

Deliverable 5.3

Approaches to procurement of balancing and redispatch and associated incentives of flexibility providers

Contributors:

AIT Austrian Institute of Technology

With the support from:

THU Technische Hochschule Ulm

KIT Karlsruher Institut für Technologie

University of St.Gallen
Institute for Operations Research and Computational Finance

TRANSNET BW

APG

FICHTNER IT CONSULTING

Funding from:

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

FFG Forschung wirkt.

Dieses Projekt wird aus Mitteln der FFG gefördert. www.ffg.at

powered by klima+ energie fonds

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Bundesamt für Energie BFE

Smart Energy Systems ERA-Net

This project has received funding in the framework of the joint programming initiative ERA-Net Smart Energy Systems' focus initiative Digital Transformation for the Energy Transition, with support from the European Union's Horizon 2020 research and innovation program under grant agreement No 883973.

Versioning and Authors

Version control Version

Version	Date	Comments
1	23.06.2025	Final Version

Authors

Viktor Zobernig,

Hemm Regina,

Stefan Strömer,

Sarah Fanta,

Tara Esterl (all AIT)

Executive Summary

The increasing share of renewable energy, progressive market integration, and slow grid expansion have significantly heightened the demand and cost for flexibility in European electricity markets. The growing mismatch between renewable generation and high-load demand areas has intensified congestion, often managed through redispatching. However, market-based remuneration of redispatch remains contentious due to the risk of inc-dec gaming, necessitating market design improvements that incentivize flexibility provision while ensuring system efficiency.

Within the DigIPlat project, we explore bid forwarding as a mechanism to enhance flexibility utilization. Our conceptual framework allows flexibility service providers (FSPs) to forward unaccepted bids across markets, increasing activation probabilities and available volumes. We develop a use case in which transmission system operators (TSOs) can activate balancing capacity market bids not only for balancing energy but also for redispatch. Assessing the strategic risks and impacts of such mechanisms is challenging due to scalability limitations and the unrealistic perfect information assumptions in conventional models. To address this, we develop an agent-based model leveraging deep reinforcement learning (DRL) to simulate multi-market interactions and strategic bidding behavior. We validate our approach by benchmarking it against state-of-the-art optimization methods and applying it to a four-market model, integrating balancing capacity, balancing energy, day-ahead, and redispatch markets, along with a grid model to capture congestion dynamics. Additionally, we analyze strategic behavior using a virtual power plant (VPP) representing a photovoltaic (PV) owner. Within this setup, we evaluate the agent's performance under two conditions: (1) a baseline scenario, where the four markets operate (i.e., without the use case), and (2) an extended use case, where bid forwarding is introduced as an additional market mechanism.

Our results confirm that DRL matches optimization benchmarks in controlled environments and outperforms them in complex, multi-market settings, where scalability constraints limit traditional approaches. In the four-market baseline scenario, the DRL agent successfully exploits flexibility markets while increasing total system costs, demonstrating its adaptability in this intricate market structure. However, in the use case scenario, its efficiency in redispatch gaming diminishes due to imperfect information, leading to more conservative bidding behavior despite its local market power. In contrast, the MILP-based VPP agent, operating under full market knowledge, exhibits less conservative behavior when applying the use case but ultimately follows a less profitable strategic approach compared to the DRL agent. Interestingly, these findings imply that higher system flexibility, despite the presence of local market power, does not inherently increase gaming risks. Instead, greater uncertainty in congestion patterns may discourage strategic exploitation, as failed attempts can carry legal and financial consequences. These insights highlight the need for robust market design that balances efficiency, flexibility, and safeguards against gaming risks in evolving electricity markets.

Kurzfassung

Der wachsende Anteil erneuerbarer Energien, die fortschreitende Marktintegration und der langsame Netzausbau haben den Bedarf an und die Kosten für Flexibilität auf den europäischen Strommärkten erheblich erhöht. Das zunehmende Ungleichmäßige Verteilung zwischen Erzeugungsgebieten erneuerbarer Energien und Regionen mit hoher Lastnachfrage verstärkt Netzengpässe und erhöht die Nachfrage nach Redispatch-Maßnahmen. Die marktbasierte Vergütung von Redispatch bleibt jedoch umstritten, da sie das Risiko für Inc-Dec-Gaming birgt. Um Anreize für die Bereitstellung von Flexibilität zu schaffen und gleichzeitig die Systemeffizienz zu gewährleisten sind daher Marktdesign Anpassungen erforderlich. Im Rahmen des DigIPlat-Projekts untersuchen wir daher Bid Forwarding als Mechanismus zur verbesserten Nutzung und Bereitstellung von Flexibilitäten. Unser konzeptioneller Rahmen ermöglicht es Flexibilitätsdienstleistern (FSPs), nicht angenommene Gebote auf andere Märkte zu übertragen, wodurch sowohl die Aktivierungswahrscheinlichkeit als auch das insgesamt verfügbare Angebotsvolumen erhöht werden. In diesem Zusammenhang analysieren wir einen Use Case, in dem Übertragungsnetzbetreiber (TSOs) Gebote aus dem Regelkapazitätsmarkt nicht nur für die Bereitstellung von Regelenergie, sondern auch für Redispatch aktivieren können.

Die Bewertung strategischer Risiken und Auswirkungen solcher Mechanismen stellt allerdings eine Herausforderung für konventionelle Modellierungsansätze dar, da diese nur schwer skalierbar sind und oft unrealistische Annahmen über perfekte Information treffen müssen. Um diesen Herausforderungen zu begegnen, entwickeln wir ein agentenbasiertes Modell auf Basis von Deep Reinforcement Learning (DRL), um strategisches Bietverhalten in einem Multi-Markt System zu simulieren. Wir validieren unseren Ansatz, indem wir ihn mit State-of-the-Art-Optimierungsmethoden vergleichen und auf ein Vier-Märkte-Modell anwenden, das Regelkapazität, Regelenergie, den Day-Ahead-Markt und einen Redispatch-Markt integriert, einschließlich eines Netzmodells zur Abbildung von Engpassdynamiken. Zusätzlich analysieren wir das strategische Verhalten eines virtuellen Kraftwerks (VPP), das einen Photovoltaik-(PV)-Betreiber repräsentiert. Innerhalb dieses Setups bewerten wir die Leistung der Agenten unter zwei Bedingungen: (1) einem Baseline-Szenario, in dem die vier Märkte ohne Bid Forwarding betrieben werden (d. h. ohne den Use Case), und (2) einem erweiterten Use Case, in dem Bid Forwarding als zusätzlicher Marktmechanismus eingeführt wird.

Unsere Ergebnisse zeigen, dass DRL die Ergebnisse etablierter Optimierungsmodelle replizieren kann und in komplexeren Multi-Markt-Szenarien überlegen ist, da die Skalierbarkeit dieser Modelle hier begrenzt ist. Im Baseline-Szenario mit vier Märkten nutzt der DRL-Agent seine Marktmacht, um die Flexibilitätsmärkte erfolgreich auszunutzen, was zu höheren Gesamtsystemkosten führt. Im Use-Case-Szenario nimmt jedoch seine Effizienz im Redispatch-Gaming aufgrund unvollständiger Informationen ab, was trotz weiterhin vorhandener lokaler Marktmacht zu vorsichtigerem Bietverhalten führt. Im Gegensatz dazu zeigt der MILP-basierte VPP-Agent, der mit vollständiger Marktinformation agiert, ein weniger konservatives Verhalten im Use Case, folgt jedoch letztlich einer weniger profitablen strategischen Herangehensweise im Vergleich zum DRL-Agenten.Diese Ergebnisse deuten darauf hin, dass eine höhere Systemflexibilität trotz lokaler Marktmacht nicht zwangsläufig zu einem Anstieg der Gaming-Risiken führt. Vielmehr kann eine erhöhte Unsicherheit über Netzengpässe strategische Ausbeutung erschweren, da gescheiterte Manipulationsversuche rechtliche und finanzielle Konsequenzen nach sich ziehen können. Unsere Erkenntnisse unterstreichen die Notwendigkeit eines robusten Marktdesigns, das Effizienz, Flexibilität und Schutzmechanismen gegen Gaming-Risiken in sich wandelnden Strommärkten vereint.

Table of contents

1. Introduction

The ongoing integration of renewable energy resources into the European grid system is altering market dynamics by changing the distribution of the actual dispatch and loads in time and space, leading to an increased need of congestion management. Consequently, this creates new possibilities for strategic behavior among power plant operators. For instance, electricity providers may engage in gaming tactics, such as intentionally contributing to congestion. As a result, topology-agnostics become more relevant for markets, creating situations where market players can suddenly achieve market power due to their location. A notable example is inc-dec gaming (increasing-decreasing gaming), where agents withhold or overbid their capacity on the day-ahead market to increase their margins on redispatch markets due to local market power [1]. These changes challenge the prevailing dynamics of existing short-term electricity markets while also offering new beneficial prospects.

To address these developments, enhancing the integration of flexibilities is crucial, not only to ensure grid security but also counteract gaming. However, this process is challenging due to the technical diversity of available flexibilities and the complex array of supply opportunities. Therefore, the aim of the project DigIPlat is (1) to develop a standardized framework for interoperable flexibility platforms, and (2) to standardize flexibility products. In this work package we focus on the latter by assessing the economic impact of coordinated capacity procurement from the existing balancing capacity market for potential use in redispatch scenarios. This corresponds to the definition of Use Case 2 “Coordinated Capacity Procurement” (see [2]). The coordination implies the idea of value stacking of flexibility products via bid forwarding[1], specifically, it is intended to increase the activation probability of capacities from flexibility suppliers.

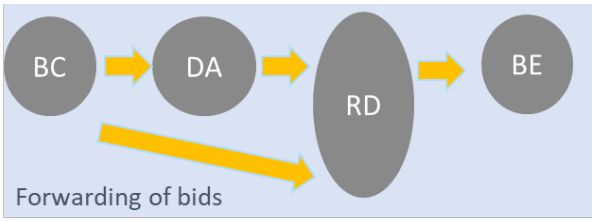


Figure 1: Illustration of the bid forwarding approach.

Beyond the notion of increasing incentives for firms, it is crucial to ensure that such changes do not adversely affect socio-economic costs. This deliverable emphasizes evaluating the added value of coordinated procurement while controlling for its impact on socio-economic costs. In other words, the main objective of this work package is to increase bidding flexibilities for firms without creating additional socio-economic burdens. To address these two perspectives, we employ a mixed integer linear program (MILP) to assess the enhanced flexibility of a virtual power plant comprising a solar and battery storage provider, alongside a deep reinforcement learning (DRL) approach to exploit bidding strategies across different markets. While MILP optimally determines volume allocation across markets, it is limited by the availability of historical data and faces challenges in identifying gaming opportunities between diverse markets. To overcome this, we apply DRL to simulate data from a non-existing market design, incorporating market-based redispatch and coordinated capacity procurement, allowing for exploration of this scenario. DRL generates the necessary data through simulation and is designed to explore optimal actions online without preexisting knowledge. In this context, we use DRL to act as an "attacker" in the analyzed market situation, aiming to identify

[1] For further details on bid forwarding, see Deliverable 3.2: "Standardized Flexibility Products and Attributes" (https://www.digiplat.eu/scientific-dissemination).

potential "worst-case" market abusive behavior. A similar notation for using DRL in electricity systems is also applied by [3].

The contributions of this deliverable can be summarized as follows:

1. Establishing how Deep Reinforcement Learning (DRL) can be utilized to assess strategic behavior across multiple interconnected markets, thereby overcoming existing model limitations.
2. Applying this method to gain new insights into the role of inc-dec gaming within the context of market-based redispatch procurement.
3. Leveraging this approach to explore new opportunities for coordinated flexibility procurement in balancing and redispatch markets.

In this deliverable, we first present an overview of the definition of Use Case 2, our key point indicators (KPIs) and the application of DRL in electricity auction modeling and. Next, we describe the model setup, including the electricity market environment, bidding-agent architecture, and data used for different scenarios. This is followed by the results section, which compares our use case to a benchmark without coordinated capacity procurement. Additionally, as the simulation approach is closely aligned with the methodology outlined in Deliverable 5.1 [4], we include the corresponding results for a flexibility provider, represented by the virtual power plant, in this deliverable. Finally, we discuss the outcomes with respect to pre-defined KPIs and conclude with our findings.

2. Description Use Case 2 and KPIs

The primary goal of coordinated capacity procurement for redispatch and balancing is to assist providers in offering their flexibility in various electricity markets through product standardization and bid forwarding. This method aims to increase the activation probability of offers by simplifying their availability across multiple markets, thereby enabling value stacking. The concept of bid forwarding is elaborated in more detail in Deliverable 3.2 – Standardized flexibility products and attributes [5]. To illustrate the benefits of product standardization and bid forwarding, we target enhanced capacity procurement for ancillary services in our use case. Specifically, we examine the impact of integrating redispatch - a currently non-market-regulated ancillary service that is gaining attention due to increasing needs and costs - with an established balancing market like aFRR. Our objective is to coordinate their procurement to potentially create positive synergies. Beyond the potential increase in market liquidity, the primary advantage aligns with our main goal: simplifying flexibility procurement. Therefore, our primary focus is not on reducing social welfare costs, but rather on increasing the capacity to provide and procure energy without disrupting existing market designs. This approach makes market participation more accessible, especially for flexibility providers not primarily involved in this business, such as demand response, households, or small power plant owners.

The use case can be summarized as follows:

"Assuming product standardization, enabling bid forwarding from accepted balancing capacity bids to be available for a potential redispatch market, and subsequently to the balancing energy market. For both balancing energy and redispatch, activated capacity is compensated with an additional energy price, based on bids that can be individually selected for each market."

To analyze this use case, the following key performance indicators (KPIs) have been identified:

I. **Procurement Deficit:** This KPI examines the impact of using balancing capacity for redispatch on balancing energy procurement. Since the required volumes for redispatch and balancing energy are highly divergent, activation of balancing capacity for redispatch might result in a deficit of capacity available for balancing energy, which will be highlighted in the analysis.

II. **Economic Impact:** This KPI assesses the effect of joint balancing capacity and redispatch procurement on overall socio-economic costs. The intention is to evaluate whether the use case leads to an increase, decrease or no significant change in gaming activities on redispatch markets due to the coordinated capacity procurement.

III. **Impact of Gaming:** This KPI investigates socio-economic costs? considering inc-dec gaming. Besides unknown potential gaming strategies arising from this use case, there are existing issues without this use case regarding the procurement of redispatch in the form of inc-dec gaming. To address this, we allow inc-dec gaming to occur with known parameters and observe its effects on the use case. This approach helps us understand whether the coordinated capacity procurement impacts inc-dec gaming without explicitly addressing the issue itself.

3. Strategic Bidding with Reinforcement Learning

Recent advancements in DRL have demonstrated its ability to navigate complex model environments and optimize strategic decisions [3], [6], [7], overcoming limitations of commonly used optimization algorithms. Traditional optimization solvers lack the capability to incorporate non-convex parameters, assume complete knowledge of market competitors and conditions, as well as model dynamics, and typically focus on either price or volume bidding [8], [9]. Based on the recent developments of reinforcement learning (RL) algorithms they gain growing attention in power systems research as an alternative to MPEC (Mathematical Program with Equilibrium Constraints) or EPEC (Equilibrium Program with Equilibrium Constraints) methods in electricity market modeling, e.g., in [6], [8]. A comprehensive review of RL in deregulated energy markets is given by [10]. In RL, an optimal solution is attained through the facilitation of a Markov Decision Process (MDP) [11]. This framework is described by having a state space S, an action space A, a transition probability distribution $P(s_{t+1}|s_t, a_t)$, and a reward function $R(s_t, a_t)$. The policy $\pi(s_t) = a_t$ determines the action executed at state s_t, iteratively updated to maximize rewards, effectively shaping the evolving strategy. The return $\mathcal{R} = \sum_{t=0}^{T} \gamma^t r_{t+1}$ describes the discounted cumulative reward, where $\gamma \in [0,1]$ serves as the discount factor accommodating immediate and future reward valuation, as expressed by the Bellman equation [11]. The MDP is defined as a sequential decision-making process, the future state and reward $(s_{t+1},\ r_{t+1})$ solely dependent on the current state and action $(s_t,\ a_t)$. An optimal policy entails maximizing the action value function, commonly recognized as the Q-function,

$$Q_\pi(s_t, a_t) \approx max_\pi E_\pi[R_t|s_t,\ a_t].$$

A detailed explanation about RL can be found in [12].

Unlike transforming the bi-level optimization into a single-level MPECs, RL solves it recursively: agents progressively enhance strategies through experiences from iterative interactions with the market clearing process. Here, market players rely solely on their operating parameters and observed market outcomes, without knowledge of competitors' operational details. Omitting the need for equivalent Karush-Kuhn-Tucker (KKT) optimality conditions enable the incorporation of non-convex agent attributes into the market clearing process (i.e., unit commitment), despite that solution optimality cannot be assured theoretically. Nevertheless, certain investigations center on confirming model reliability, demonstrating equivalent accuracy to game-theoretical driven approaches: [13] exhibits the effectiveness of utilizing RL in converging to Nash Equilibriums for learning bidding strategies in a duopoly day-ahead market. [14] shows the reliance of competition on the discount factor, ranging between complete competitiveness and tacit collusion. [8] compared a MPEC model with a deep reinforcement learning algorithm, demonstrating that while there is no significant difference in performance in convex environments, RL significantly outperforms MPEC in non-convex problem formulations. Conversely, EPEC and MPEC solvers encounter limitations as increasing model complexity challenges their foundational assumptions, leading to intractable solutions. In contrast, RL algorithms are versatile in diverse environments, albeit without guaranteed determining the optimal solution in finite-time; however, they frequently yield pragmatic strategies approximating real strategic behavior, giving valuable insights for adapting market-designs reasonably.

Hence, agent-based model frameworks employing RL for strategic decision-making processes can incorporate more complex models without imposing further assumptions. The framework necessitates solely a single assumption regarding market participant behavior: their pursuit of profit maximization

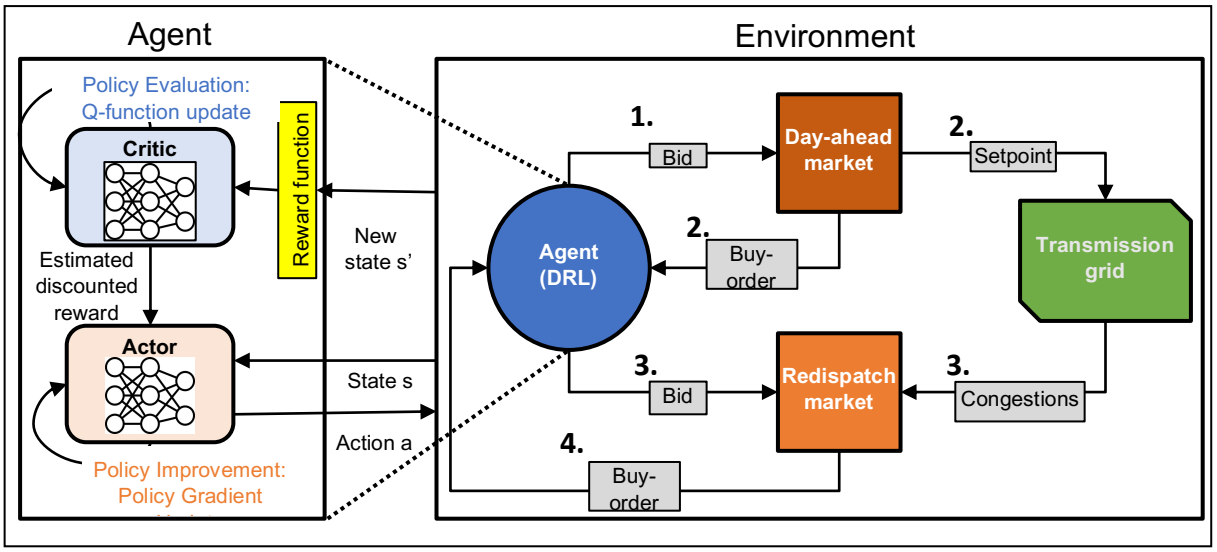


Figure 2: Deployment of deep reinforcement learning (DRL) integrated with an electricity market model within an agent-based model framework.

in the face of uncertainty. However, real-world agents often exhibit behaviors that deviate from rational profit optimization, despite their underlying objective to maximize returns. Considerations as risk-aversion and short-termism can be factored rational within the context of more intricate utility functions than the strict expected profits. Thus, the incorporation of uncertainties originating from the market environment can be readily accommodated as well. RL methodologies can adapt to such models merely by adjusting the reward function appropriately. These model designs enable the agent not only to learn optimal response strategies, akin to EPEC models, but also potentially to demonstrate genuine market manipulation tactics, such as spoofing[2] [15]. This ability for adapting behavior can further be exemplified by the recent findings of [16] who illustrated how RL algorithms can utilize strategic bluffing to outmaneuver opponents in the game Stratego.

It is important to acknowledge that the mentioned ability of utilizing RL for complex problems stems from the advancement of DRL. Algorithms like the deep deterministic policy gradient (DDPG) [17] are effectively applicable for continuous action spaces and improved learning efficiency through the integration of neural networks. In DRL, neural networks are utilized as function approximator using an actor-critic architecture. The critic, i.e., Q-network, $Q(s, a|\theta^Q)$ outputs a single value that rates the performance of the actor network, i.e., policy-network, $\pi(s|\theta^\pi)$, which outputs an action on a continuous action space. With θ representing the weights of the neural networks. The agent learns its behavior by storing its gathered experience in a memory buffer M and samples to train the neural networks. An illustrative overview of how DRL is deployed together with an electricity market model is given in Figure 2.

[2] Spoofing is a manipulative trading practice where a trader places orders with the intent to cancel them before they are executed, creating a false impression of market demand or supply. The trader benefits by executing trades at artificially favorable prices based on the misleading signals they generated.

4. Method

In this section, we detail the different types of scenarios simulated, along with an overview of the market environment and market players involved, to evaluate the KPIs for our Use Case (see Section 2.). We particularly emphasize the architecture of the reinforcement learning agents (RL-agents) used in the simulations. Alongside RL-agents, we employ mixed-integer linear programming-based agents (MILP-agents) and rule-based agents (RB-agents).

The MILP-agents and RL-agents are designed to analyze different types of behavior. RL-agents assess the impact of gaming behavior, while MILP-agents focus on optimal volume allocation of a virtual power plant (VPP) with installed photovoltaic (PV) and battery storage. Consequently, these two types of agents are never included in a single simulation simultaneously. The RB-agents are based on a well-known scenario from the literature [18], which we extended by incorporating historical data for Austria from November 2022 to October 2023. This approach results in two kinds of RB-agents: those that behave based on pre-determined best-response functions (based on [18]), representing conventional and renewable electricity providers, and additional conventional power providers that always bid their full capacity at actual market prices (based on historical data). This setup guarantees significant competition among agents with established Nash Equilibria, while also contending with actual historical clearing prices. The MILP-agents are outlined in more detail in Deliverable 5.1 [4] and therefore are not described further here.

4.1. Market Environment and Data

The market environment is based on the Chao and Peck 6-bus network [19], utilized for similar use cases in, e.g., [18] and [20] to assess the role of inc-dec gaming[3] in sequential electricity markets, including redispatch. Our focus is on the implementation from [18], expanding their scenario to encompass the balancing capacity and energy markets, alongside the day-ahead and redispatch markets. The paper's original example examines inc-dec gaming in market-based redispatch, finding that agents strategically leverage their local positions to maximize profits through elevated redispatch prices. Given that most European countries still compensate redispatch based solely on incurred costs, there is no comprehensive data available for our case study. Consequently, we rely on this studied example from the literature and incorporate real data where available. The limited data further highlights the benefits of using RL to learn “online” without any pre-existing knowledge (see more details about RL below in 4.2 Reinforcement Learning Agent Architecture. Additionally, this framework enables us to validate the learned policy of our RL-agents by comparing scenarios with known Nash Equilibria. Apart from redispatch, we use load and price data for Austria from November 2022 to October 2023, freely available from the Austrian transmission grid operator (TSO), Austrian Power Grid (APG). This approach is used because the literature example omits seasonal variations in market dynamics, instead of emphasizing optimization to achieve a Nash Equilibrium for a single time step. However, simulating balancing energy markets without any collected data is challenging because they do not follow any discernible pattern or distribution, evidenced by the difficulty in creating reliable forecast models for these markets [21]. In addition to market data, we incorporate weather forecasts and real-time availability for renewables from ENTSO-E's Transparency Platform [22].

[3] For more details on inc-dec gaming with potential redispatch markets see [1].

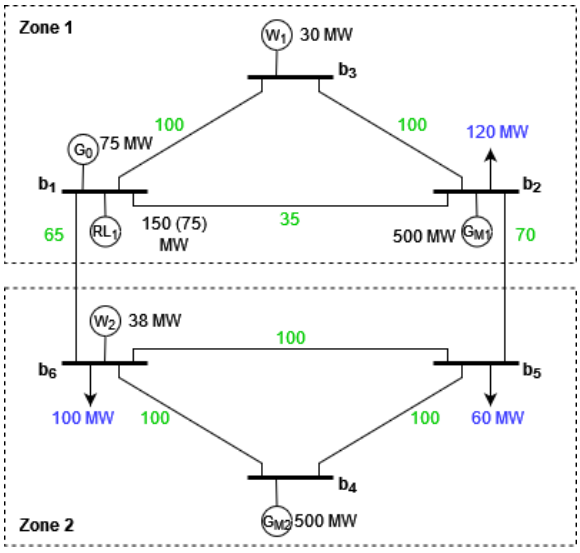


Figure 3: Overview of the 6-bus network, including generator and line capacities.

Table 1: Generator names, location, technology, marginal costs, and capacities.

Name	Bus	Carrier	Marginal Cost	Capacity
RL Conv 1	b1	Gas	90	130
Conv 2	b2	Gas	90	72
Solar 1	b3	PV	0	30
Solar 2	b4	PV	0	37.5
Marginal Conv 1	b5	Gas	100	500
Marginal Conv 2	b6	Gas	100	500

To align the applied data with the example from the literature, we scale it accordingly. For the day-ahead market, we employ a scaling factor α derived by dividing the fixed demand value μ_{Model} from the paper example by the mean of the historical data $\mu_{DA_{Data}}$.

$$\boldsymbol{Y} = \alpha \boldsymbol{X}, \qquad with\ \alpha = \frac{\mu_{Model}}{\mu_{DA_{Data}}}$$

This factor is subsequently applied to adjust the entire historical load dataset, including balancing capacity and energy. No scaling is applied to the price data from any markets or to the weather forecasts. Generally, in this setup, congestion occurs by separating the 6-bus network into two zones with limited transmission capacity between them on the day-ahead market. Congestion that needs to be solved via redispatch results from distributing the installed generation capacity and loads in the same ratio as in the paper example. By employing this scaling approach, we replicate congestion patterns similar to those analyzed in the literature, however, include fluctuation based on the actual historical load and weather data used. An overview of the 6-bus network and the capacities and MCs of Generators is given in Figure 3.

In reinforcement learning, we must decide the same number of discrete actions for each market, necessitating a uniform time resolution across all markets, which we set to an hourly basis. For the balancing capacity market, which operates in six 4-hour blocks, we replicated the data for each hour within each block. For the balancing energy market, we created hourly samples from 15-minute data entries. This sampling approach was extended to generate denser samples representative of each season by selecting data for each day of the week based on corresponding weekdays within the season. To further reduce data complexity for the neural networks, we created four representative hours for each day of the week for each season, effectively capturing daily, weekly, and seasonal variations. This

Table 2: Summary of data types, sources and applied scaling

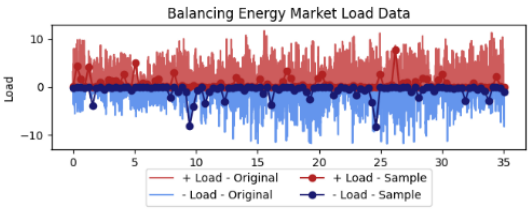


Figure 5: Distribution of balancing energy load data, including a sampled subset used for simulation.

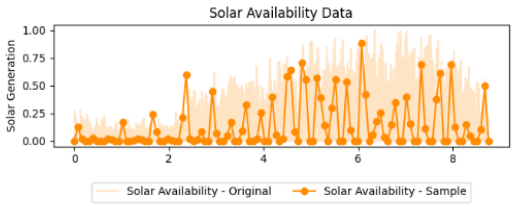


Figure 6: Distribution of solar data, including a sampled subset used for simulation.

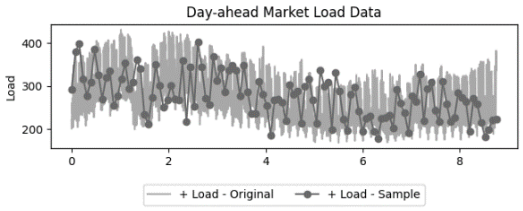


Figure 4: Distribution of day-ahead load data, including a sampled subset used for simulation.

method allows us to train on a variety of scenarios derived from the same dataset, incorporating stochasticity that reflects these temporal effects. A more detailed explanation of the data sampling process is provided in the Appendix. Figures 4-6 offer an overview of the data used, including an example of a sampled set employed in the simulation. Table 2 summarizes the data types, sources, and any applied scaling.

For the four distinct markets included in this study, we employ uniform pricing for the day-ahead and balancing energy markets and pay-as-bid pricing for the balancing capacity and redispatch market, reflecting real-world practices. All markets, except for redispatch, are modeled as zonal markets, consistent with common European practices.

The overall market clearing process is structured as followed:

- **Day-Ahead Market:** We solve a linear optimal power flow (LOPF) for market clearing, disregarding line capacities, except for the cross-border lines.
- **Post Day-Ahead Congestion Check:** We solve another LOPF, this time considering line capacities, to identify any congestion.
- **Redispatch Market:** If congestion is detected, a call for redispatch bids is issued, and the market is cleared using nodal pricing to resolve congestion.
- **Balancing Energy Market:** We focus solely on the secondary energy market (aFRR) due to its competitive pricing and operational flexibility compared to primary (FCR) and tertiary (mFRR) energy markets.

We assume a Colombian bidding format, with single divisible bids for each hour and agent in each market, involving volume and price decisions. Additionally, we apply the following criteria for balancing capacity and energy bids:

- **Accepted Capacity Bids:** *Negative* capacity bids from the balancing capacity market are mandated to be offered in the day-ahead market at the price floor. *Positive* capacity bids are withheld for the markets they are reserved for.
- **Balancing Energy Bids:** These are automatically generated based on accepted capacity bids, allowing only decisions on the energy price. No additional free bids are considered.

Table 3: Overview market implementations. All markets are solved in 4 x 1-hour blocks[4].

Market	GCT	Pricing Rule	Forwarding
Balancing Capacity	D-1, 10:00	Pay-as-Bid	+/- accepted bids get reserved
Day-ahead	D-1, 12:00	Pay-as-Cleared	- reserved capacities are forced to be offered at price floor
Redispatch	D-1, 18:00	Pay-as-Bid	**Only for UC2:** +/- reserved capacities are converted into energy bids. Free bidding is allowed; single price bids are added.
Balancing Energy	T-25min	Pay-as-Cleared	+/- reserved capacities are converted into energy bids. Free bidding is not allowed; only price bids are added.

4.2 Reinforcement Learning Agent Architecture

Our implemented RL-agents are based on the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [23], which is one of the state-of-the-art methods suitable for continuous action spaces. TD3 is model-free, meaning it does not have access to or information about the model environment itself, a common characteristic in Equilibrium Problems with Equality Constraints (EPEC) often used to determine the role of gaming, as in [18]. It trains off-policy and online, collecting data during training from the environment dynamics and storing it in a replay buffer. For policy updates, batches are sampled from this buffer. As described in Section 3., TD3 employs an actor-critic architecture for function approximation to estimate the maximum expected discounted future reward. The critic, i.e., Q-network, $Q(s, a|\theta^Q)$ outputs a single value that rates the performance of the actor network, i.e., policy-network, $\pi(s|\theta^\pi)$, outputting the action, i.e., bids. Where s, a and r represent the state, action, and reward, respectively, and θ the weights of the neural network. The Q-network is updated using the following loss function:

$$L = \frac{1}{N}\sum_j (y_j - Q(s_j, a_j|\theta^Q))^2$$

where N is the number of sampled transitions (s_j, a_j, r_j, s_{j+1}) from the memory buffer. The loss is calculated as the average loss over each of these sampled transitions. The target y_j is defined as:

$$y_j = r_j + \gamma Q'\left(s_{j+1}, \pi'\left(s_{j+1}\middle|\theta^{\pi'}\right)\middle|\theta^{Q'}\right).$$

Q' and π' are target networks, which represent copies of the actor and critic with parameters $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\pi'} \leftarrow \theta^\pi$ to avoid interdependencies during updating. The target networks are updated as follows:

[4] The implementation of all markets is also detailed in Deliverable 5.1 "Integration of standardized flexibility requirements and multi-market commercialization of flexibility in virtual power plant" (https://www.digiplat.eu/scientific-dissemination).

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'},$$
$$\theta^{\pi'} \leftarrow \tau\theta^{Q\pi} + (1-\tau)\theta^{\pi'}.$$

The actor network is then updated using:

$$\nabla_{\theta^{\pi}} J \approx \frac{1}{N}\sum_{j} \nabla_{a} Q\left(s, a \left|\theta^{Q}\right.\right)|_{s=s_j, a=\pi(s_j)} \nabla_{\theta^{\pi}} \pi(s|\theta^{\pi}) \mid s_j$$

Unlike the DDPG algorithm, TD3 addresses the overestimation of Q-values for certain states by utilizing a second Q-network [23]. The target value y (see Equation ??) is defined by taking the minimum of the two Q-values from the separate Q-networks, $min(Q1, Q2)$. The resulting loss from both Q-networks is then summed for updating the networks.

In reinforcement learning (RL), a critical aspect is the decision on state and action representation, essentially defining the input and output. The state refers to the data used as information to train the agent, providing it with the necessary context to recognize patterns, while the actions set the degrees of freedom for interacting with the environment. Additionally, defining the reward function is crucial, as it serves as the sole feedback mechanism for the agent's performance. Therefore, careful consideration of how the objective is defined significantly influences the resulting trained policy. The state, actions and rewards are defined as follows:

- **Actions**: As previously mentioned, we sample four hours for each representative time period to capture daily electricity consumption patterns and fluctuations. Consequently, an agent's action, comprising a price and volume bid, is represented as a 1x8 vector—consisting of 4 price bids and 4 volume bids for each market. We employ a hyperbolic tangent activation function to map all actions between -1 and 1, where 1 corresponds to the maximum available volume to regulate upwards and -1 to the maximum volume to regulate downwards for each agent, respectively. For prices, 1 equates to the price cap and -1 to the price floor. To ensure the agent explores the action space during training, we add Gaussian noise to each action.
- **States**: Based on a comprehensive investigation about various data usages, we have included the following variables in our final version:
  - The available capacity ("unused capacity") of the agent itself (for each hour),
  - the running energy ("used capacity") of the agent itself (for each hour),
  - the clearing prices of the last three days (for each hour),
  - available weather forecasts (sourced from ENTSO-E data) (for each hour), and
  - a proxy indicating the specific market (out of the four markets) for which the agent needs to output an action.
- **Reward**: For the reward R, we focus on immediate profits P from the current market session, represented by time step t. It is defined as the total profits from each hour h within this time step, excluding any fixed cost. This serves as the primary objective for the agent to maximize:

$$R_t = P_t = \sum_{h}^{H} SC_{h,t} * \left(CP_{h,t} - MC_{h,t}\right)$$

With $SC_{h,t}$ being the sold capacity, $CP_{h,t}$ the clearing price and $MC_{h,t}$ the marginal costs. Note that $MC_{h,t}$ can become negative if, for example, downward regulation leads to savings in fuel prices. However, since strategic interaction between markets might involve forgoing profits in one market to

increase profits in another, and thus overall profits, we use the return G instead of R. G is defined as the total accumulated reward from a particular time step t:

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$$

Here, T represents the finite time horizon, and γ is a discount factor between 0 and 1. Using the return enables the agent to make decisions considering the full bidding horizon of one day, thus we always take the return from a single day. Figure 7. provides a detailed overview of the TD3 architecture, highlighting the simulation iteration and update processes essential for optimizing the agent's performance.

4.3 Use Case Integration and Scenario Overview

To integrate our use case into the model, we extend the redispatch market framework to allow forwarded accepted bids from the balancing capacity (BC) market to be utilized for redispatch (RD), just as they are for balancing energy (BE). Specifically, in our use case, energy bids for redispatch are automatically formed based on the accepted capacity bids. The agent is also free to set the price at which this volume is offered, just as with balancing energy. However, note that free bids for redispatch are enabled, including those from agents with accepted capacity bids. Due to limitations in the RL-agent's action space, a single energy price bid applies to both free and capacity-based bids. Besides our use case, one of our primary objectives is to evaluate the effectiveness of using RL to address existing model limitations. As mentioned earlier, we draw on a well-known example from the literature that analyzes the occurrence of inc-dec gaming with market-based redispatch, both with and without market power [18].

In the results section, we first benchmark our approach and demonstrate the benefits of using RL to increase the degrees of freedom in deciding on price and volume bids simultaneously. Subsequently, we examine a scenario encompassing all four markets, incorporating data from Austria, with and without applying the use case. In this second batch of experiments, we also assess the role of local market power, a common real-world situation that enables market abuse, despite being generally prohibited. Finally, we include the results related to Deliverable 5.1, which involve the Mixed Integer Linear Program addressing perfect volume allocation for a virtual power plant comprising PV and battery storage.

Overview of the scenarios:

a. Benchmarking RL with established model from the literature
b. Evaluating the use case of coordinated capacity procurement, with and without market power
c. Assessing volume allocation for flexibility providers using a VPP featuring PV and battery storage

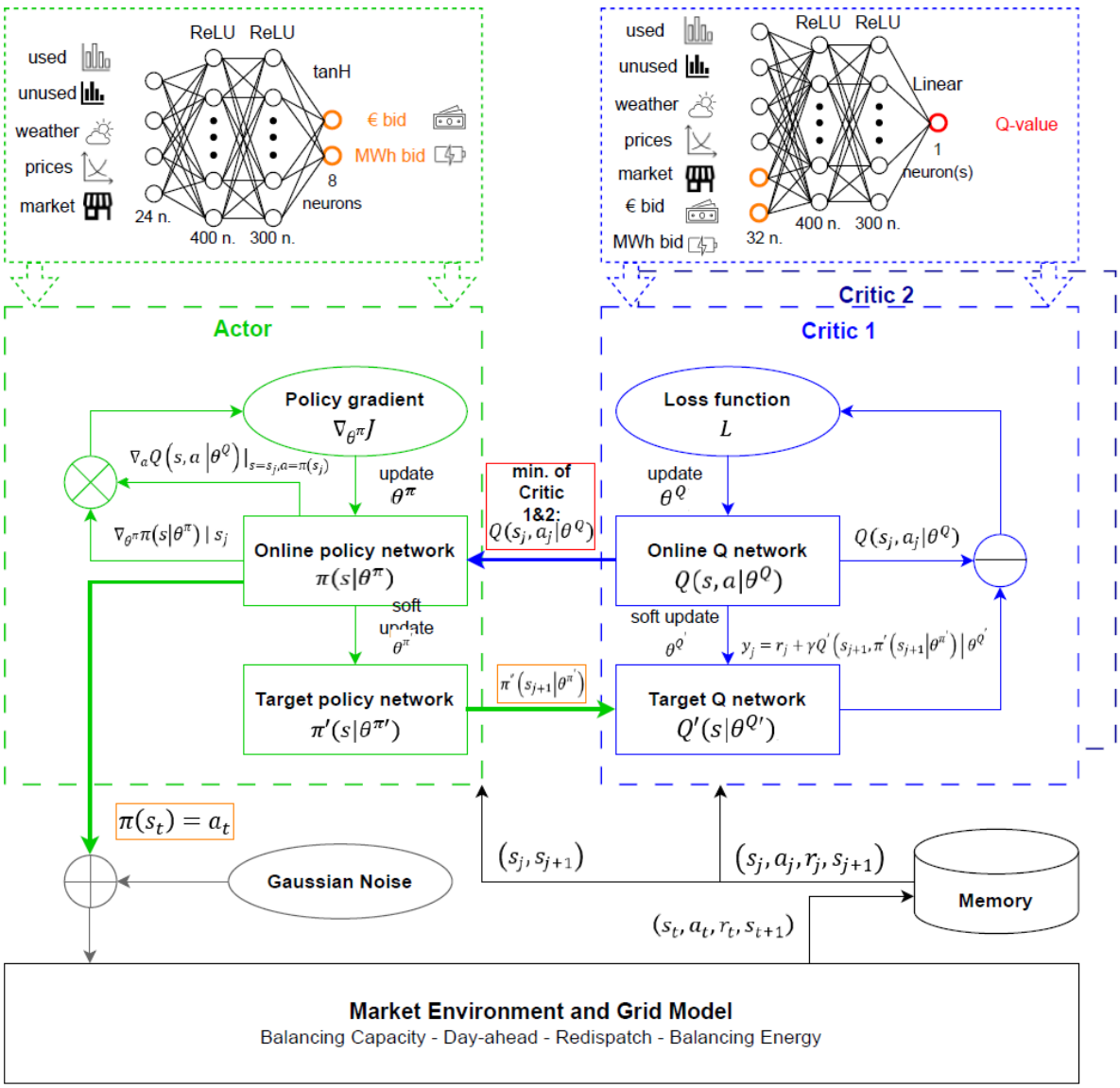


Figure 7: Comprehensive overview of the TD3 architecture, detailing the simulation iteration and update processes.

5. Results

This section presents the validation of the implemented TD3 algorithm against established two-market model from the literature as well the detailed analysis of our extended four market setup. The analysis covers two distinct setups: (1) Two-Markets Scenario: Day-Ahead and Redispatch. (2) Four-Markets Setup: Balancing Capacity, Day-Ahead, Redispatch, and Balancing Energy, including the evaluation of our defined use case on coordinated capacity procurement. Additionally, we present results from using the virtual power plant (VPP), described in detail in Deliverable 5.1, that models the behavior of a PV-generating entity to assess the impact of increased flexibility in bidding behavior. Both, the result from TD3 and VPP, are shown here to facilitate comparison and draw the final conclusions regarding our KPIs through an integrated assessment.

5.1 Two Markets: Model Validation and Gaming with Redispatch

To validate our TD3 agent application, we replicated the model from [18] and replaced one of the conventional generators with our RL-agent. In our use case, this corresponds to the “RL Conv 1” agent at bus-1, similar to agent “u1” in [18]. This agent was selected due to its local market power, enabling it to engage in inc-dec gaming by underbidding its marginal cost on the day-ahead market to ensure activation and then selling energy at higher prices on the redispatch market for downward regulation. In [18], an EPEC model is used to find a Nash Equilibrium for all five agents, where each bids a price according to their best response function within a discrete decision space. We included all optimal bidding decisions from the other agents to compete with our RL-agent, challenging it to find this known equilibrium.

The results in Figure 8 illustrate the distribution of total profits between the day-ahead and redispatch markets, showing that the majority originates from redispatch. This is driven by the agent exploiting its market power through inc-dec gaming. Our RL agent achieves nearly identical results, with only a 0.02% deviation, which stems from the residual noise added to its bidding actions to ensure sufficient exploration. To investigate the capabilities of RL further, we allowed the agent to decide on its volume bids as well, different to the EPEC model where only price decisions are considered. We tested an additional scenario where the agent could bid 20% more than its installed capacity to see if it would strategically bid "ghost" capacity to increase profits, displayed by the right bar labeled “RL-Agent

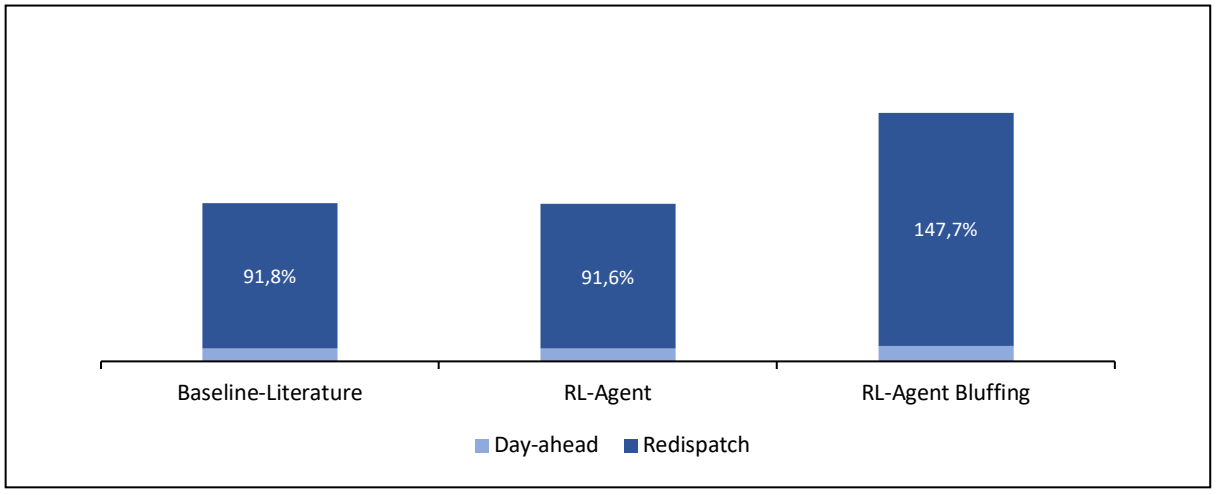


Figure 8: Profit distribution between day-ahead and redispatch markets across different setups. Both the Baseline-Literature and RL-Agent setups show similar redispatch profit shares, at 91.8% and 91.6%, respectively. The RL-Agent Bluffing setup exhibits an increased redispatch profit share of 147.7%, indicating more aggressive bidding strategies.

Bluffing". The agent learned to increase its volume bids, including ghost volume, up to a point where it could still predict being regulated downward via redispatch before needing to deliver the additional non-existent energy.

To further demonstrate the impact of increased degrees of freedom, we applied a bluffing scenario to a cost-based redispatch procurement model. In this model, redispatch is remunerated solely based on incurred costs, while allowing the agent to profit through savings. Specifically, savings occur if the agent secures profits in the day-ahead market when the clearing price exceeds its marginal cost. In this setup, the agent achieved 35% higher profits compared to the case without volume bid decisions. However, total profits remain approximately one-third of those obtained under extreme conditions when considering market-based remuneration. This increase is primarily driven by downward redispatch, where the agent avoids production costs, leading to negative costs (payments from the transmission system operator). Importantly, the agent does not have to return the difference between the marginal cost and the day-ahead clearing price, due to the unrestricted nature of market participation. We can conclude that even without market-based redispatch procurement, there is an incentive to engage in gaming; however, it is important to note that the ratio between the cost and the market-based remuneration is about five times lower due to the restricted price margins.

5.2 Four Markets: Analyzing Economic Impact with DRL

To assess the impact of our proposed use case, we extended the two-market model from Section 5.1 by incorporating balancing capacity and energy markets and integrating weather and load data. This enhancement improves the realism of market dynamics by accounting for non-Gaussian stochasticity, as detailed in Section 4.1. The final results are illustrated in Figure 9, where *Baseline* refers to the four-market scenario with a redispatch market but without applying the use case of combined procurement. In contrast, *Combined Procurement* represents the same scenario with the use case applied. Compared to the *Baseline* scenario, this market design increases awarded flexibility (including balancing energy and redispatch) by nearly 10%. However, the DRL agent's profits from redispatch provision decline by more than 40%, which also translates into reduced redispatch costs. Since the

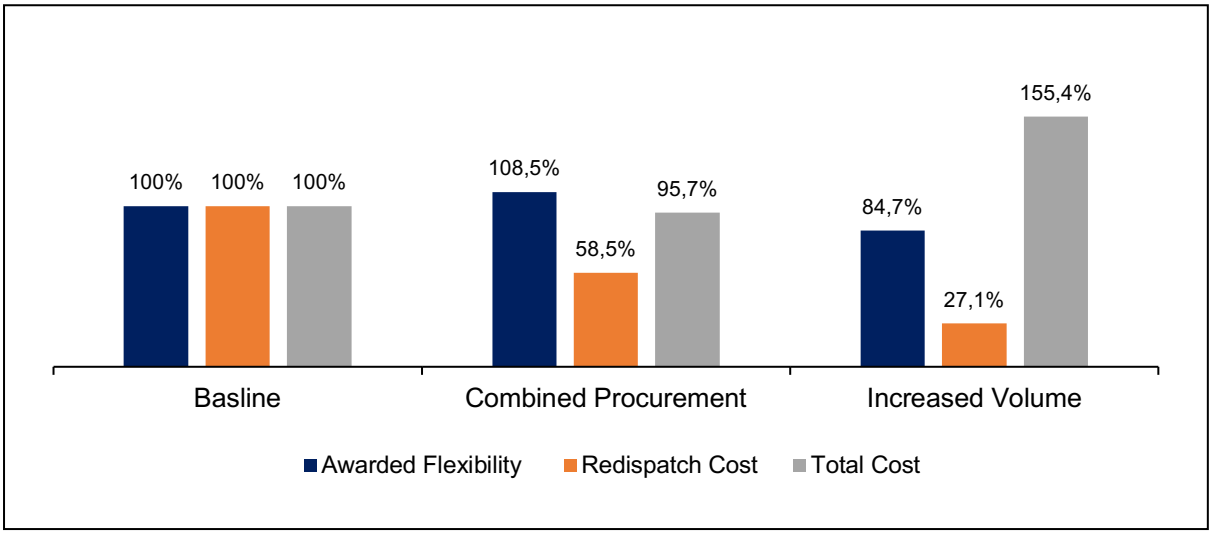


Figure 9: Final results comparison for the four-market scenario. The Baseline represents the setup without combined procurement but includes market-based redispatch remuneration. Combined Procurement builds on the baseline by incorporating the use case, while Increased Volume extends it further by increasing balancing capacity procurement to also cover redispatch needs. Awarded Flexibilities include balancing energy and redispatch, while Total Costs refer to overall system costs.

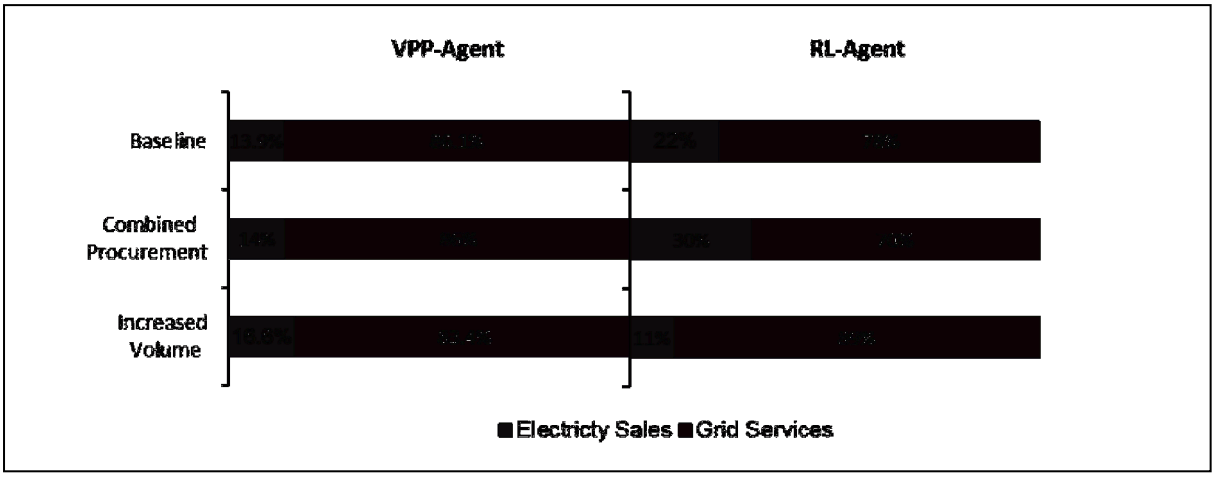
agent retains local market power, it remains the sole entity influencing redispatch market prices. Additionally, total system costs decrease by approximately 5%. These improvements result solely from the introduction of bid forwarding, with all other model parameters identical to the Redispatch Market scenario. A closer look at the results suggests that these systemic changes stem from a shift in RL-agents trading behavior. Increased system flexibility and a dilution of traceable action-reaction patterns make it more difficult to identify and exploit inefficiencies. Moreover, the remaining inefficiencies—despite the agent's initial market power—become harder to capitalize on due to heightened volatility, both in market conditions and the agent's profits. This increased uncertainty discourages opportunistic strategic behavior, leading to more risk-averse bidding decisions. Furthermore, these findings suggest that participants in a regulator-controlled system (e.g., without access to confidential market information) may struggle to identify and exploit existing inefficiencies.

The agent's difficulty in sustaining high profits provides further insight into the dynamics of highly interdependent markets. To deepen this understanding, we introduce an additional design modification: increasing procured balancing capacity to better accommodate potentially higher redispatch peaks, referred to as *Increased Volume* in Figure 9. This adjustment results in a nearly 75% reduction in redispatch profits compared to the Redispatch Market scenario. While awarded flexibility (energy) decreases by approximately 25% relative to the current use case, this does not yield a net benefit for the system. Instead, the increased demand for balancing capacity creates new opportunities for strategic behavior, driving up clearing prices and ultimately raising total system costs by roughly 60%.

5.3 Four Markets: Analyzing Firm Flexibility with VPP

The market simulation with the RL-Agent revealed a noticeable change in bidding behavior: a shift in profit allocation between conventional electricity sales and grid services. This shift suggests a potentially higher opportunity for exploiting structural issues related to grid services in the *Baseline* scenario where the use case is not applied. This observation assumes that introducing the use case increases market and system flexibility, which may also make strategic gaming more complex.

This change is almost negligible when analyzing the VPP results derived from the RL-Agent data, further supporting the previous assumption: The introduction of market-based redispatch procurement for



situations of dominant local market power positively impacts gaming potentials from the system's

Figure 10: Comparison of profit distribution between electricity sales (day-ahead market) and grid services (balancing capacity, balancing energy, and redispatch) for the VPP agent and RL agent. The VPP agent exhibits significantly less fluctuation in profit margins across scenarios, indicating lower sensitivity to strategic interactions compared to the RL agent.

perspective. In the case of the RL agent, this directly influences the market actor's behavior. However, due to the non-strategic nature of the VPP, this market design change does not demonstrate any significant effects. Refer to Figure 10 for a visual comparison. While the previous interpretation suggests that changes in market design are primarily necessitated when gaming actors make up a considerable portion of the overall system, the following observation highlights that such changes could also enhance fairness in the market. In contrast to the extreme results observed in the online simulation with the RL-Agent, the *Increased Volume* scenario under VPP participation provides a valuable initial insight. Restrictively low prices for (negative) balancing capacities largely hinder the VPP's participation in the common use-case. However, the extension of procured volume enables it to secure accepted bids in the market. As a result, the VPP refrains from “exploiting” additional grid services (balancing energy and redispatch), as the already procured - and therefore guaranteed capacity - appears sufficient to meet demand, outside of extreme situations. This observation suggests that a well-structured, regulated, and unified market for grid services, exemplified by bid forwarding through the combined procurement use case studied here, could offer an effective and efficient solution to the broader challenge of declining dispatchable capacities.

A careful analysis may suggest an apparent contradiction with the findings in Section 5.2. However, we argue that these results arise solely due to the extreme (and unrealistic) nature of high market-power: The RL agent operates freely in a currently non-existing market, competing against rule-based and highly conservative bidders aligned with historical market results, without assuming perfect information. In contrast, the VPP agent, utilizing MILP-based optimization, inherently benefits from complete knowledge of the exploited data, allowing it to make less conservative decisions in this high-flexibility market scenario. Mitigating such behavior—exemplified by the inherently non-gaming nature of the VPP—through sufficient competition could significantly reduce gaming opportunities. This aspect, we believe, is critical for further studies.

6. Conclusions

In this study, we explored the application of reinforcement learning (RL) to address limitations in state-of-the-art optimization methods for electricity markets. Our primary objective was to evaluate complex multi-market scenarios involving up to four distinct markets, where agents must align bidding strategies. RL serves as a tool to identify worst-case scenarios by leveraging market dynamics, contributing to the development of more resilient market designs resistant to gaming behavior. To evaluate the applicability of RL, we first validated the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm against a well-established benchmark before extending the analysis to a more complex four-market scenario. This scenario is based on a 6-bus grid model, capturing the occurrence of congestion and the influence of a potential redispatch market. Additionally, we integrated historical weather and load data from Austria, introducing non-Gaussian stochasticity to create a more realistic market environment. This setup allowed us to assess our proposed combined procurement use case, which enables the TSO to activate awarded flexibilities from the balancing capacity market for both balancing energy and redispatch, thereby enhancing overall system flexibility.

The RL agent successfully converged to the same equilibria as determined by the EPEC model used for validation. Under extreme conditions of local market power and high certainty in predicting congestion, the agent strategically exploited the redispatch market through inc-dec gaming. Allowing the agent to optimize both price and volume bids, a flexibility not feasible in the EPEC model, demonstrated that such strategies remain advantageous even under cost-based remuneration regimes. However, the resulting profit margins were still only one-third of those achieved under market-based remuneration. Extending the analysis to our four-market scenario revealed a 5% reduction in total system costs under the combined procurement use case. However, despite retaining local market power, the RL agent exploited the redispatch market less efficiently, highlighting the challenges posed by increased market flexibility. The absence of perfect information made it more challenging for the agent to detect action-reaction patterns in a highly flexible system, leading to more conservative bidding behavior. To deepen our understanding, we further increased the procured balancing capacity to fully cover redispatch needs. However, this raised total system costs by 60%, as it introduced new strategic opportunities for the RL-agent to leverage local market power. These findings highlight that while increased system flexibility can enhance efficiency; it must be carefully monitored by regulatory authorities to mitigate potential market manipulation risks.

Finally, we evaluated market outcomes by introducing a VPP agent based on MILP optimization to explore cross-market opportunities using the same dataset as the RL agent. Unlike the RL agent, the VPP agent exhibited no significant profit shifts between scenarios. We argue that this outcome stems primarily from the extreme market power granted to the RL agent: it competes exclusively against rule-based and conservative bidders without assuming perfect information, whereas the MILP-based VPP agent operates with complete knowledge of market conditions, enabling less conservative decision-making in a high-flexibility setting. Nevertheless, the resulting strategy still yields lower overall profits compared to the RL agent.

Based on these results, it is evident that while gaming remains an inherent risk, the presence of local market power amplifies its impact. However, our findings indicate that greater market flexibility does not necessarily increase gaming opportunities; rather, it can reduce them, as higher system flexibility adds uncertainty for strategic gaming while maintaining or even increasing activation probabilities for suppliers. Without confidence in congestion patterns, exploiting potential gaming opportunities becomes too risky, as failed attempts can have legal and financial consequences. Moreover, market-based remuneration can incentivize competition, reducing the risk of emerging local market power. This, in turn, enables a more efficient allocation of resources, ultimately lowering total system costs. Nevertheless, continuous market monitoring will be essential when introducing new mechanisms like the proposed use case. Further research should address how regulatory frameworks can effectively

balance market flexibility and competition while minimizing unintended strategic behavior. Ensuring this balance will be crucial as electricity markets evolve toward more dynamic and interconnected systems. Our results highlight the potential of reinforcement learning in enhancing market design and underscore the importance of carefully managing bidding flexibilities to maintain market efficiency and integrity. As regulatory frameworks adapt to these evolving conditions, thoughtful policy design will be essential for mitigating risks while fostering a competitive and efficient electricity sector.

7. References

[1] L. Hirth and I. Schlecht, "Market-Based Redispatch in Zonal Electricity Markets," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3286798.

[2] J. Dierenbach *et al.*, "Deliverable 3.3. Definition of multifunctional flexibility use cases," 2023. [Online]. Available: https://www.digiplat.eu/fileadmin//NES/DigIPlat_D3.3-UseCases_final..pdf

[3] T. Wolgast, E. M. Veith, and A. Nieße, "Towards reinforcement learning for vulnerability analysis in power-economic systems," *Energy Inform.*, vol. 4, no. S3, p. 21, Sep. 2021, doi: 10.1186/s42162-021-00181-5.

[4] V. Zobernig, R. Hemm, S. Fanta, S. Strömer, and T. Esterl, "D5.1_Integration_of_standardized_flexibility_requirements_and_multi-market_commercialization_of_flexibility_in_a_virtual_power_plant.pdf," Wien, Deliverable, 2024. Accessed: Jul. 18, 2024. [Online]. Available: https://www.digiplat.eu/fileadmin//NES/D5.1_Integration_of_standardized_flexibility_requirements_and_multi-market_commercialization_of_flexibility_in_a_virtual_power_plant.pdf

[5] K. Tolstrup *et al.*, "Deliverable 3.2. Standardized flexibility products and attributes," 2022. [Online]. Available: https://www.digiplat.eu/fileadmin//NES/DigIPlat_D3.2-Standardized_flexibility_attributes_final.pdf

[6] K. Poplavskaya, J. Lago, and L. de Vries, "Effect of market design on strategic bidding behavior: Model-based analysis of European electricity balancing markets," *Appl. Energy*, vol. 270, p. 115130, Jul. 2020, doi: 10.1016/j.apenergy.2020.115130.

[7] J. Tran, L. Gajewski, P. Pfeifer, S. Krahl, and A. Moser, "Simulation of strategic bidding for battery storage and e-mobility in local flexibility markets with multi-agent reinforcement learning," in *CIRED Porto Workshop 2022: E-mobility and power distribution systems*, Jun. 2022, pp. 716–720. doi: 10.1049/icp.2022.0804.

[8] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep Reinforcement Learning for Strategic Bidding in Electricity Markets," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1343–1355, 2020, doi: 10.1109/TSG.2019.2936142.

[9] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, “Deep Reinforcement Learning for Joint Bidding and Pricing of Load Serving Entity,” *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6366–6375, Nov. 2019, doi: 10.1109/TSG.2019.2903756.
[10] Z. Zhu, Z. Hu, K. W. Chan, S. Bu, B. Zhou, and S. Xia, “Reinforcement learning in deregulated energy market: A comprehensive review,” *Appl. Energy*, vol. 329, p. 120212, Jan. 2023, doi: 10.1016/j.apenergy.2022.120212.
[11] R. Bellman, “A Markovian Decision Process,” *J. Math. Mech.*, vol. 6, no. 5, pp. 679–684, 1957.
[12] R. S. Sutton and A. Barto, “Reinforcement Learning: An Introduction.” Accessed: Aug. 24, 2023. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html
[13] C. Graf, V. Zobernig, J. Schmidt, and C. Klöckl, “Computational Performance of Deep Reinforcement Learning to Find Nash Equilibria,” *Comput. Econ.*, Jan. 2023, doi: 10.1007/s10614-022-10351-6.
[14] Y. Liang, C. Guo, Z. Ding, and H. Hua, “Agent-Based Modeling in Electricity Market Using Deep Deterministic Policy Gradient Algorithm,” *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4180–4192, Nov. 2020, doi: 10.1109/TPWRS.2020.2999536.
[15] X. Wang, C. Hoang, and M. P. Wellman, “Learning-based trading strategies in the face of market manipulation,” in *Proceedings of the First ACM International Conference on AI in Finance*, New York New York: ACM, Oct. 2020, pp. 1–8. doi: 10.1145/3383455.3422568.
[16] J. Perolat *et al.*, “Mastering the Game of Stratego with Model-Free Multiagent Reinforcement Learning,” *Science*, vol. 378, no. 6623, pp. 990–996, Dec. 2022, doi: 10.1126/science.add4679.
[17] T. P. Lillicrap *et al.*, “Continuous control with deep reinforcement learning,” Jul. 05, 2019, *arXiv*: arXiv:1509.02971. doi: 10.48550/arXiv.1509.02971.
[18] M. Sarfati and P. Holmberg, “Simulation and Evaluation of Zonal Electricity Market Designs,” *Electr. Power Syst. Res.*, vol. 185, p. 106372, Aug. 2020, doi: 10.1016/j.epsr.2020.106372.
[19] H.-P. Chao and S. Peck, “A market mechanism for electric power transmission,” *J. Regul. Econ.*, vol. 10, no. 1, pp. 25–59, Jul. 1996, doi: 10.1007/BF00133357.
[20] A. Midttun Systad, J. Løken Eilertsen, F. Van de Sande Araujo, and R. Egging-Bratseth, “An analysis of mitigating measures for inc-dec gaming in market-based redispatch,” *Masteroppgave Ind. Økon. Og Teknol.*, Jun. 2022.
[21] S. Backe, S. Riemer-Sørensen, D. A. Bordvik, S. Tiwari, and C. A. Andresen, “Predictions of prices and volumes in the Nordic balancing markets for electricity,” in *2023 19th International Conference on the European Energy Market (EEM)*, Jun. 2023, pp. 1–6. doi: 10.1109/EEM58374.2023.10161961.
[22] “ENTSO-E Transparency Platform.” Accessed: Aug. 12, 2024. [Online]. Available: https://transparency.entsoe.eu/
[23] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing Function Approximation Error in Actor-Critic Methods,” Oct. 22, 2018, *arXiv*: arXiv:1802.09477. doi: 10.48550/arXiv.1802.09477.

8. Glossary

S	State space (set of all possible states in the environment)
A	Action space (set of all possible actions an agent can take)
R	Reward function (determines the reward for a given state-action pair)
P	Policy distribution (probability distribution over actions given a state)
Q	Q-value (estimates the expected return of taking an action under a policy)
π	Policy function (maps states ss to actions aa)
γ	Discount factor (weights future rewards in reinforcement learning)
τ	Learning rate for target networks (controls update speed of target networks)
θ	Parameters of neural networks (weights of function approximators)
L	Loss value for critic network (measures prediction error of the value function)
$\nabla\theta\pi J$	Gradient to update the actor network (adjusts policy parameters to maximize rewards)
y	Target value (expected return used for training value networks)
SC	Sold capacity (amount of energy sold in the market)
CP	Clearing price (final price at which trades are settled in the market)
MC	Marginal cost (incremental cost of producing one more unit of energy)
G	Return (cumulative discounted reward in reinforcement learning)
N	Total number of drawn samples (size of the sample set used for learning)
t	Time step index (discrete index for sequential decision-making)
h	Hour index (index representing different hours in a time series)
j	Index of drawn sample (index referring to a specific sample in a batch)
k	Window of visited markets at the current time step (number of past markets considered)

9. Appendix

Data Sampling Methodology

To analyze hourly data over the span of a year while accounting for seasonal and weekly variations, we implemented a comprehensive sampling strategy. This approach ensures that our dataset captures representative patterns and trends across different seasons and days of the week.

Dataset Overview

The original dataset D comprises hourly data points collected over an entire year, resulting in a total of 8760 data points:

$$D = \{x_t \mid t = 1,2, \dots ,8760\}$$

Seasonal Division

The year is divided into four seasons, each containing 13 weeks:

- S_1: Season 1 (Winter)
- S_2: Season 2 (Spring)
- S_3: Season 3 (Summer)
- S_4: Season 4 (Fall)

For each season S_j $(j \in \{1,2,3,4\})$, we extract the corresponding subset of the dataset:

$$D_j = \{x_t \mid t \in S_j\}$$

Weekly Sampling

From each season S_j, we sample one representative week W_j:

$$W_j \subset D_j$$

Daily and Hourly Sampling

Within each sampled week W_j, we further divide each day into four consecutive 6-hour intervals:

- Interval 1: $[0{:}\,00 - 05{:}\,59)$
- Interval 2: $[6{:}\,00 - 11{:}\,59)$
- Interval 3: $[12{:}\,00 - 17{:}\,59)$
- Interval 4: $[18{:}\,00 - 23{:}\,59)$

For each day $d \in \{1,2, \dots ,7\}$ of the sampled week, we select one data point from each of these intervals, resulting in four data points per day:

$$W_j = \{x_{t_{j,d,h}} \mid d \in \{1,2, \dots ,7\}, h \in \{1,2,3,4\}\}$$

Where:

- $t_{j,d,1}$: Time sampled from the interval $[0{:}\,00 - 5{:}\,59)$ on day d of season j.
- $t_{j,d,2}$: Time sampled from the interval $[6{:}\,00 - 11{:}\,59)$ on day d of season j.
- $t_{j,d,3}$: Time sampled from the interval $[12{:}\,00 - 17{:}\,59)$ on day d of season j.
- $t_{j,d,4}$: Time sampled from the interval $[18{:}\,00 - 23{:}\,59)$ on day d of season j.

Final Sampled Dataset

The final dataset $\boldsymbol{S}$ is the union of the sampled weeks from each season:

$$\boldsymbol{S} = \bigcup_{j=1}^{4} W_j$$

This results in a dataset containing four weeks of data, one from each season, with each week comprising seven days, and each day containing four data points sampled from each 6-hour interval.