UDC: 004.4
DOI: https://doi.org/10.66941/3106-9037/2026/iss3pp17-31

Azibek B.[1], Buganova S.[2], Kurmash A.[1], Kulmanov A.[3], Zhakiyev N.[2,*]
[1]Astana IT University, Astana, Kazakhstan;
[2]Kazakh Leading Academy of Architecture and Civil Engineering;
[3]International Educational Corporation, Almaty, Kazakhstan.
E-mail: n.zhakiyev@kazgasa.kz

REINFORCEMENT LEARNING FOR STRATEGIC BIDDING IN DAY-AHEAD ELECTRICITY MARKET

Balzhan Azibek, Senior Lecturer, School of Intelligent Systems, Astana IT University, Astana, Kazakhstan;
E-mail: balzhan.azibek@astanait.edu.kz, orcid.org/0000-0001-7389-7777
Buganova Svetlana, Professor, Candidate of Technical Sciences, Kazakh Leading Academy of Architecture and Civil Engineering (KazGASA), Almaty, Kazakhstan;
E-mail: svetlanabuganova7@gmail.com, https://orcid.org/0000-0003-2005-3305
Alisher Kurmash, junior researcher, School of AI and Data Science, Astana IT University, Astana, Kazakhstan;
E-mail: 243008@astanait.edu.kz
Azamat Kulmanov, International Educational Corporation, Almaty, Kazakhstan, E-mail: a.kulmanov@kazgasa.kz
Nurkhat Zhakiyev, Professor, Kazakh Leading Academy of Architecture and Civil Engineering (KazGASA), Almaty, Kazakhstan.
E-mail: n.zhakiyev@kazgasa.kz, orcid.org/0000-0002-4904-2047

Abstract

Reinforcement learning may improve strategic bidding in electricity markets, yet many studies rely on stylized or synthetic environments. This study evaluates reinforcement learning using bid-level data from the Kazakhstan Electricity and Power Market Operator (KOREM) covering July 2023–November 2025. Objectives include reconstructing endogenous pay-as-cleared prices through merit-order clearing, modeling one strategic generator alongside passive historical participants, and evaluating out-of-sample performance. Marginal costs are approximated by the 10th percentile of each generator's historical bid prices. Double Q-learning replaces classical Q-learning to address observed instability associated with overestimation bias, while bounded markup decisions and realistic market constraints ensure economically plausible bidding. Evaluation using a chronological 80/20 train–test split indicates an approximately 0.37% profit improvement relative to cost-based bidding, without distorting clearing price distributions or intraday patterns. Stabilized reinforcement learning can modestly enhance strategic bidding in an empirically grounded electricity market while preserving fundamental price formation properties.

Keywords: day-ahead electricity market; agent-based modeling; reinforcement learning; Q-learning; double Q-learning; strategic bidding; market power.

Б. Азибек[1], С. Буганова[2], А. Курмаш[1], А. Кулманов[3], Н. Жакиев[2]*
[1]Astana IT University, Астана, Казахстан;
[2]Казахская головная архитектурно-строительная академия (КазГАСА), Алматы, Казахстан;
[3]Международная образовательная корпорация, Алматы, Казахстан.
E-mail: n.zhakiyev@kazgasa.kz

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ ДЛЯ СТРАТЕГИЧЕСКИХ ТОРГОВ НА РЫНКЕ «НА СУТКИ ВПЕРЁД»

Балжан Азибек, старший преподаватель Школы интеллектуальных систем, Astana IT University, Астана, Казахстан;

E-mail: balzhan.azibek@astanait.edu.kz, https://orcid.org/0000-0001-7389-7777
Буганова Светлана, профессор, кандидат технических наук, Казахская головная архитектурно-строительная академия (КазГАСА), Алматы, Казахстан;
E-mail: svetlanabuganova7@gmail.com, ORCID: https://orcid.org/0000-0003-2005-3305
Алишер Курмаш, мнс Школы искусственного интеллекта и науки о данных, Astana IT University, Астана, Казахстан. E-mail: 243008@astanait.edu.kz
Азамат Кулманов, Международная образовательная корпорация, Алматы, Казахстан;
E-mail: a.kulmanov@kazgasa.kz
Нурхат Жакиев, профессор, Казахская головная архитектурно-строительная академия (КазГАСА), Алматы, Казахстан.
E-mail: n.zhakiyev@kazgasa.kz, ORCID: https://orcid.org/0000-0002-4904-2047.

Аннотация

Обучение с подкреплением может улучшить стратегическую подачу заявок на электроэнергетических рынках, однако многие исследования используют стилизованные или синтетические среды. Исследование оценивает обучение с подкреплением на основе данных заявок Казахстанского оператора рынка электрической энергии и мощности (KOREM) за июль 2023 - ноябрь 2025 года. Задачи включают воспроизведение эндогенного формирования единой равновесной цены посредством ранжирования заявок, моделирование одного стратегического генератора при пассивном поведении остальных исторических участников и вневыборочную оценку эффективности. Предельные издержки оцениваются по 10-му процентилю исторических цен заявок каждого генератора. Двойное Q-обучение заменяет классическое Q-обучение для устранения нестабильности, связанной с завышением оценок; ограниченные ценовые надбавки и реалистичные рыночные ограничения обеспечивают экономическую правдоподобность поведения. Хронологическое разделение данных на обучающую и тестовую выборки в соотношении 80/20 показывает увеличение прибыли приблизительно на 0,37% относительно заявок на основе издержек без искажения распределений равновесных цен и внутрисуточной динамики. Стабилизированное обучение с подкреплением позволяет умеренно улучшить стратегическую подачу заявок на эмпирически обоснованном рынке, сохраняя основные свойства ценообразования.

Ключевые слова: рынок электроэнергии «на сутки вперёд»; агентное моделирование; обучение с подкреплением; Q-обучение; двойное Q-обучение; стратегическая подача заявок; рыночная власть.

Б. Азибек[1], С. Буганова[2], Ә. Құрмаш[1], А. Кулманов[3], Н. Жакиев[2,*]
[1]Astana IT University, Астана, Қазақстан;
[2]Қазақ бас сәулет-құрылыс академиясы (ҚазБСҚА), Алматы, Қазақстан;
[3]Халықаралық білім беру корпорациясы, Алматы, Қазақстан.
E-mail: n.zhakiyev@kazgasa.kz

ЭЛЕКТР ЭНЕРГИЯСЫ НАРЫҒЫНДА «ТӘУЛІК БҰРЫН» СТРАТЕГИЯЛЫҚ САУДАЛАСУҒА АРНАЛҒАН НЫҒАЙТУМЕН ОҚЫТУ

Балжан Азибек, Зияткерлік жүйелер мектебінің аға оқытушысы, Astana IT University, Астана, Қазақстан;
E-mail: balzhan.azibek@astanait.edu.kz, https://orcid.org/0000-0001-7389-7777
Буганова Светлана, профессор, техника ғылымдарының кандидаты, Қазақ бас сәулет-құрылыс академиясы (ҚазБСҚА), Алматы, Қазақстан;
E-mail: svetlanabuganova7@gmail.com, https://orcid.org/0000-0003-2005-3305.
Әлішер Құрмаш, Жасанды интеллект және деректер ғылымы мектебінің кіші ғылыми қызметкері, Astana IT University, Астана, Қазақстан, E-mail: 243008@astanait.edu.kz
Азамат Кулманов, Халықаралық білім беру корпорациясы, Алматы, Қазақстан;
E-mail: a.kulmanov@kazgasa.kz

Нұрхат Жакиев, профессор, Қазақ бас сәулет-құрылыс академиясы (ҚазБСҚА), Алматы, Қазақстан.
E-mail: n.zhakiyev@kazgasa.kz, https://orcid.org/0000-0002-4904-2047.

Аннотация

Нығайтумен оқыту электр энергиясы нарықтарындағы стратегиялық өтінім беруді жақсарта алады, алайда көптеген зерттеулер шартты немесе синтетикалық орталарға сүйенеді. Зерттеуде Электр энергиясы мен қуат нарығының қазақстандық нарық операторының АҚ (KOREM) ұсынған 2023 жылғы шілде мен 2025 жылғы қараша аралығындағы өтінім деректері негізінде нығайтумен оқыту бағаланады. Зерттеу міндеттер өтінімдерді бағасы бойынша ранжирлеу арқылы бірыңғай тепе-теңдік бағаның эндогенді қалыптасуын қайта құруды, тарихи қатысушылардың енжар мінез-құлқы жағдайында бір стратегиялық генераторды модельдеуді және оқыту іріктемесінен тыс нәтижелерді бағалауды қамтиды. Шекті шығындар әрбір генератордың тарихи өтінім бағаларының 10-процентилі арқылы жуықталады. Q-мәндерін асыра бағалауға байланысты тұрақсыздықты азайту үшін классикалық Q-оқытудың орнына қосарлы Q-оқыту қолданылады; шектелген баға үстемелері мен шынайы нарықтық шектеулер экономикалық тұрғыдан негізді мінез-құлықты қамтамасыз етеді. Деректерді оқыту және тестілеу іріктемелеріне хронологиялық бөлу (80/20) кезінде алынған нәтижелер агенттің тепе-теңдік бағалардың үлестірімдері мен тәулік ішіндегі баға динамикасын бұрмаламай, шығындарға негізделген өтінімдермен салыстырғанда пайданы шамамен 0,37%-ға арттыратынын көрсетті. Тұрақтандырылған нығайтумен оқыту эмпирикалық негізделген нарықта баға қалыптасуының негізгі қасиеттерін сақтай отырып, стратегиялық өтінім беруді біршама жақсарта алады.

Түйін сөздер: «тәулік бұрын» электр энергиясы нарығы; агенттік модельдеу; нығайтумен оқыту; Q-оқыту; қосарлы Q-оқыту; стратегиялық өтінім беру; нарықтық билік.

Introduction. Liberalized electricity markets rely on auction-based mechanisms to coordinate decentralized supply and demand while maintaining efficiency, reliability, and competitive neutrality. In day-ahead markets, generators submit price–quantity bids for hourly delivery, and a system operator clears the market using a merit-order dispatch rule that determines a uniform clearing price. Although this framework is designed to promote competitive efficiency, strategic bidding remains an inherent feature of such markets. Heterogeneous cost structures, residual demand elasticity, repeated market interaction, and information asymmetry create incentives for generators to deviate from pure marginal-cost bidding.

Traditional approaches to analyzing strategic behavior in electricity markets rely on equilibrium-based models, most notably Nash–Cournot formulations and bilevel optimization frameworks. While analytically powerful, these approaches assume a high degree of rationality and informational completeness. In practice, market participants operate under uncertainty, observe only partial system states, and adapt behavior dynamically over time. As a result, equilibrium models often provide theoretical benchmarks rather than operational descriptions of strategic learning.

Agent-based modeling (ABM) has emerged as an alternative framework capable of representing electricity markets as complex adaptive systems. In ABM environments, prices and dispatch outcomes emerge from interactions among heterogeneous agents following behavioral rules. This paradigm enables modeling of bounded rationality, adaptation, and institutional detail within auction-based market designs.

Reinforcement learning (RL) provides a natural mechanism for embedding adaptive strategic behavior within such agent-based environments. By iteratively updating decisions based on realized profit signals, RL agents can learn effective bidding strategies without explicit knowledge of competitor cost functions or closed-form equilibrium conditions. Recent studies have demonstrated the potential of RL and deep RL in stylized electricity market simulations. However, the empirical grounding of such models remains limited in many cases, as synthetic demand curves and artificial cost functions are often employed.

The core problem addressed in this study is therefore the following: How does reinforcement learning perform when embedded within a reconstructed real-world day-ahead electricity market derived directly from historical bid-level data, rather than within a stylized synthetic environment?»

More specifically, the study investigates whether adaptive bidding via reinforcement learning can improve generator profitability while preserving structural properties of price formation in an empirically reconstructed market environment.

This problem is non-trivial. Reinforcement learning algorithms, particularly value-based methods, are susceptible to instability, overestimation bias, and excessive exploitation when interacting with auction mechanisms. Without careful design, RL-based strategies may generate unrealistic price escalation or volatility artifacts that do not reflect plausible market behavior. Evaluating RL in a realistic empirical setting therefore requires careful integration of auction reconstruction, cost proxy estimation, and stability safeguards.

The relevance of this study is both methodological and practical.

From a methodological perspective, embedding reinforcement learning within a reconstructed real market environment advances the realism of strategic bidding simulations. By reconstructing clearing outcomes directly from historical bid-level data, the model avoids reliance on synthetic cost assumptions and externally imposed clearing prices. This approach strengthens the credibility of conclusions regarding algorithmic performance.

From a practical perspective, electricity markets are increasingly influenced by algorithmic decision systems. Generators employ automated optimization tools to formulate bids, and the penetration of digital decision-making continues to grow. Understanding how adaptive learning algorithms interact with real auction structures is therefore important for both market participants and regulators.

If reinforcement learning algorithms systematically distort price formation, regulators may need to consider oversight mechanisms. Conversely, if bounded and stabilized RL can enhance efficiency without destabilizing markets, it may represent a viable strategic optimization tool under competitive constraints. By evaluating reinforcement learning within a reconstructed empirical market derived from nearly two years of real day-ahead bidding data, this study provides insight into how adaptive algorithms behave under authentic competitive conditions. The primary contribution of this paper is the integration of reinforcement learning into a fully reconstructed empirical market environment using bid-level data from KOREM. Rather than using published clearing prices or synthetic cost curves, the model reconstructs hourly supply stacks directly from historical bids and applies a pay-as-cleared clearing mechanism consistent with institutional market rules. The strategic agent operates within this reconstructed environment while all other participants retain their historical bids. This design ensures realistic competitive depth while isolating the effect of adaptive bidding.

Because marginal cost data are not publicly available, the study introduces a transparent proxy based on the 10th percentile of historical bid prices for each generator. This allows markup-based strategic behavior to be modeled without introducing artificial cost assumptions. Another key contribution is methodological. Initial experiments using classical Q-learning produced unstable markup escalation due to overestimation bias. Transitioning to Double Q-learning stabilized convergence and eliminated artificial price volatility, highlighting the importance of algorithm selection in electricity market applications.

Finally, the model is evaluated using a chronological 80/20 train–test split, ensuring that performance improvements reflect genuine out-of-sample predictive adaptation rather than in-sample overfitting.

Literature review

Strategic bidding in day-ahead electricity markets arises from the structure of uniform-price auctions and merit-order dispatch. When demand approaches available capacity, residual demand becomes less elastic, allowing generators to submit bids above marginal cost without losing dispatch. This behavior is a structural feature of electricity markets rather than an anomaly (Biggar and Hesamzadeh, 2014: 281–306).

Equilibrium-based models formalize these incentives using Nash–Cournot frameworks, showing that prices may exceed marginal cost under capacity and transmission constraints (Contreras et al., 2004: 195–206). When operational constraints such as unit commitment are included, the strategic problem becomes dynamic and dependent on generation feasibility (Baíllo et al., 2001: 227–248). However, these models rely on assumptions of full rationality and complete information, which limit their applicability in real market environments (Ventosa et al., 2005: 897–913).

Electricity markets operate as repeated and adaptive systems, where participants adjust strategies based on observed outcomes. Agent-based computational frameworks capture this behavior by modeling markets as evolving systems driven by decentralized interactions (Tesfatsion, 2002: 55–82). In such models, strategies emerge over time rather than being derived analytically (Tesfatsion, 2003: 262–268).

Agent-based simulations allow explicit representation of auction rules and heterogeneous participants, enabling analysis of price dynamics and market power under adaptive conditions (Guerci et al., 2010: 241–286). Platforms such as AMES demonstrate how learning agents interact with realistic clearing mechanisms (Li and Tesfatsion, 2009: 1–8), while dynamic testing shows that market outcomes depend on learning trajectories rather than equilibrium states alone (Sun and Tesfatsion, 2007: 291–327). These approaches establish a foundation for modeling adaptive strategic behavior in electricity markets.

Reinforcement learning provides a natural extension of agent-based approaches by formalizing adaptive bidding as a sequential decision-making process. In this framework, agents update their strategies based on realized profits without requiring explicit knowledge of competitor behavior or cost structures. Early applications of Q-learning to electricity supplier bidding demonstrate that such agents can improve profitability through iterative interaction with the market (Xiong et al., 2002: 1516–1521).

More advanced approaches incorporate probabilistic reasoning and structured learning, allowing agents to operate under uncertainty while maintaining stability (Dehghanpour et al., 2016: 4744–4754). Recent developments in deep reinforcement learning enable continuous bidding strategies and more complex state representations, improving flexibility and scalability (Ye et al., 2020: 1343–1353). Multi-agent reinforcement learning further shows that decentralized learning can approximate equilibrium-like outcomes under competitive conditions (Du et al., 2021: 534–544).

A key direction in recent research is the use of empirically calibrated environments rather than synthetic models. Reinforcement learning applied to realistic market data improves the credibility of results and reduces reliance on artificial assumptions (Di Persio et al., 2025). At the same time, safety-constrained reinforcement learning has been proposed to ensure that learned strategies remain economically and operationally feasible (Stanojev et al., 2023: 1–7). Dedicated simulation environments confirm that reinforcement learning must be evaluated within structured and institutionally consistent market settings (Jain and Saxena, 2023; Schimeczek et al., 2023).

Despite these advances, reinforcement learning models remain sensitive to instability. In particular, value-based methods such as Q-learning are prone to overestimation bias, which can lead to excessive markup strategies in stochastic environments (van Hasselt, 2010: 2613–2621). Addressing this issue is critical for applying reinforcement learning in electricity markets.

The effectiveness of reinforcement learning depends on the interaction between learning algorithms and auction design. Market microstructure influences both efficiency and the extent of strategic behavior (Nicolaisen et al., 2001: 504–523; Bower and Bunn, 2000: 1–29). Therefore, reinforcement learning must be evaluated within the constraints of the clearing mechanism.

Instability in value-based reinforcement learning, particularly overestimation bias, can lead to unrealistic outcomes if not controlled. This has motivated the development of more stable algorithms and constraint-based approaches that limit excessive strategic behavior (van Hasselt, 2010: 2613–2621; Stanojev et al., 2023: 1–7). These findings highlight that reinforcement learning in electricity markets must balance profitability with structural stability.

To clarify the conceptual differences between traditional equilibrium-based approaches and reinforcement learning methods, Table 1 provides a structured comparison across key modeling dimensions.

Table 1 - Comparison between equilibrium-based models and reinforcement learning approaches in electricity market modeling across key methodological and behavioral dimensions

Feature	**Equilibrium Models**	**Reinforcement Learning**	**Ref.**
Behavior assumption	Fully rational agents	Bounded rationality	(Weidlich and Veit, 2008: 1728–1759)
Information availability	Full information	Partial / local observations	(Tesfatsion, 2003: 262–268)

Adaptation	Static (one-shot solution)	Dynamic (learning over time)	(Guerci et al., 2010: 241–286)
Solution method	Analytical optimization (e.g., Nash)	Iterative learning (trial-and-error)	(Xiong et al., 2002: 1516–1521)
Strategic modeling	Explicit equilibrium formulation	Implicit strategy learning	
Scalability	Limited for complex systems	High (data-driven)	(Ye et al., 2020: 1343–1353; Du et al., 2021: 534–544)
Empirical realism	Often low (stylized assumptions)	High (when using real data)	
Uncertainty handling	Limited	Naturally incorporated	
Interpretability	High	Moderate	

As shown in Table 1, reinforcement learning provides a more flexible and adaptive framework for modeling strategic behavior under uncertainty. However, this flexibility comes at the cost of reduced interpretability and potential instability, particularly in value-based methods such as Q-learning.

The proposed framework evaluates adaptive strategic behavior under realistic competitive depth, demand variability, and institutional constraints. Therefore, this work contributes to bridging the gap between theoretical reinforcement learning models and empirically grounded electricity market analysis.

Model Framework

Market Environment. The simulated environment represents a uniform-price day-ahead electricity market reconstructed from historical KOREM bid-level data. For each delivery day d and hour h, the market consists of a set of generators $i \in \{1,\ldots,N\}$, each submitting price-quantity offers.

The use of a uniform-price, merit-order market structure follows the institutional design widely analyzed in electricity market economics (Biggar and Hesamzadeh, 2014: 119–138). Such market structures have been extensively modeled both in equilibrium formulations (Contreras et al., 2004: 195–206) and in agent-based simulation environments (Li and Tesfatsion, 2009: 1–8). The present framework adopts the institutional structure of these markets while embedding it within a reconstructed empirical dataset.

The model does not incorporate transmission constraints, nodal pricing, or unit commitment decisions. This abstraction isolates markup-based strategic behavior within a single clearing zone, consistent with many agent-based electricity market studies that focus on auction dynamics rather than network feasibility (Guerci et al., 2010: 241–286). The omission of intertemporal commitment constraints is deliberate, as strategic unit commitment introduces a substantially more complex bilevel optimization structure (Baíllo et al., 2001: 227–248).

Demand for each hour is derived from historical KOREM data. To avoid deterministic overfitting during training, stochastic multiplicative noise is introduced, scaled according to peak and off-peak regimes. This treatment aligns with the adaptive modeling perspective in which agents interact with a stochastic but institutionally consistent environment rather than a static deterministic schedule (Sun and Tesfatsion, 2007: 291–327).

Agent Structure and Roles. The reconstructed market consists of a set of agents corresponding to generators observed in the historical KOREM dataset. Each agent represents an individual market participant submitting price-quantity bids in the day-ahead market. In line with agent-based modeling principles, agents operate independently and make decisions based on local information rather than global system knowledge (Tesfatsion, 2002: 55–82; Tesfatsion, 2003: 262–268).

To isolate the effect of adaptive strategic behavior, agents in the model are divided into two categories: a single strategic learning agent and multiple passive agents.

The strategic agent represents one generator whose bidding behavior is modeled using reinforcement learning. Instead of submitting its historical bid price, this agent applies a markup over its estimated marginal cost and adapts this markup over time to maximize profit. The decision-making process of the strategic agent is based on observed market conditions, including clearing prices, demand

levels, and previous outcomes. This approach reflects adaptive learning behavior commonly studied in electricity market simulations (Ye et al., 2020: 1343–1353; Xiong et al., 2002: 1516–1521).

All remaining agents are modeled as passive participants. These agents submit their historical bids without modification and do not adapt their strategies over time. Their role is to provide a realistic and empirically grounded competitive environment for the strategic agent. By preserving the original bid structure of the market, the model maintains the actual supply stack and competitive depth observed in the real system.

This hybrid agent design, combining one adaptive agent with a population of fixed historical agents, is widely used in agent-based electricity market studies to ensure interpretability and stability of results (Pinto et al., 2013: 335–346; Di Persio et al., 2025). Unlike fully adaptive multi-agent systems, where all participants learn simultaneously, the current approach allows the impact of reinforcement learning on strategic bidding to be evaluated in isolation.

Although simplified, this structure captures the key dynamics of strategic interaction in electricity markets while avoiding the additional complexity and convergence issues associated with multi-agent reinforcement learning frameworks (Du et al., 2021: 534–544; Khamzina et al., 2025).

Reconstruction of Supply Stack and Clearing Mechanism. For each hour h, all submitted bids are aggregated and sorted in ascending order of offered price. Let $P_{i,h}$ denote the submitted price and Qi,h the offered quantity of generator i.

Cumulative supply is constructed as shown in Eq. (1):

$$S_{h(k)} = \sum_{j=1}^{k} Q_{(j),h} \tag{1}$$

The cumulative function $S_{h(k)}$ defines the total available supply up to the k-th bid, allowing construction of the aggregate supply curve. Bids are ordered such that:

$$P_{(1),h} \leq P_{(2),h} \leq \cdots \leq P_{(N),h} \tag{2}$$

This ordering represents the merit-order ranking of bids, where lower-priced offers are dispatched first, forming the basis of the supply curve.

The market-clearing price Ph* is determined as the price of the marginal accepted bid satisfying, as shown in (3):

$$S_h(k^*) \geq D_h \tag{3}$$

where D_h is hourly demand. The index k^* identifies the marginal unit required to meet demand D_h, determining the cutoff point for accepted bids.

This pay-as-cleared reconstruction reflects the clearing principle implemented in most pool-based electricity markets and studied extensively in computational market test beds such as AMES (Li and Tesfatsion, 2009: 1–8). Auction microstructure has been shown to influence strategic incentives significantly (Nicolaisen et al., 2001: 504–523; Bower and Bunn, 2000: 1–29), and therefore reconstructing clearing endogenously from bids ensures institutional fidelity. By avoiding direct use of published historical clearing prices, the model allows strategic bid modifications to propagate through the same merit-order logic that governs real market outcomes.

Marginal Cost Proxy. True generator marginal costs are not publicly available in the dataset. To enable markup modeling relative to cost, a proxy marginal cost MC_i is estimated for each generator i as the 10th percentile of its historical bid prices:

$$MC_i = \text{Percentile}_{10}\{P_{i,h,d}\} \tag{4}$$

This percentile-based proxy (4) assumes that the lower tail of historical bids approximates cost-reflective offers during competitive conditions. Similar empirical approximations have been employed in data-driven electricity market modeling when direct cost data are unavailable (Di Persio et al., 2025).

The approach avoids introducing synthetic cost curves while maintaining empirical grounding in observed bidding behavior.

While this proxy does not represent true fuel or operational cost data, it provides a consistent reference point for modeling markup-based strategic behavior within an empirically reconstructed environment.

Strategic Agent Structure. The reconstructed market includes all historical bidders observed in the dataset. However, only one generator is modeled as a strategic reinforcement learning agent. All other participants submit their historical bids unchanged.

This hybrid structure, one adaptive agent embedded within a population of passive participants, follows a common experimental design in agent-based market simulations where the objective is to isolate the behavioral impact of a single learning agent (Pinto et al., 2013: 335–346). The strategic agent replaces its historical price with a markup-based bid defined as in Eq. (5):

$$P_{s,h} = MC_s(1 + m_h) \tag{5}$$

where m_h is the markup selected by the reinforcement learning policy for hour h. The offered quantity remains equal to its historical quantity.

This formulation reflects the economic interpretation of strategic bidding as deviation from marginal-cost offers, consistent with equilibrium-based markup logic (Biggar and Hesamzadeh, 2014: 281–306; Contreras et al., 2004: 195–206), but without requiring explicit equilibrium solution.

Reinforcement Learning Formulation. The strategic bidding problem is modeled as a sequential decision process. At each hour, the agent observes a state vector incorporating recent clearing prices, demand conditions, and profit outcomes. Such state-based adaptation aligns with reinforcement learning formulations applied to electricity supplier bidding (Xiong et al., 2002: 1516–1521).

The action space consists of discretized markup levels bounded within economically plausible limits. Bounding the action space is essential in auction environments, as unbounded value-based learning may otherwise generate unrealistic escalation behavior, a concern emphasized in reinforcement learning applications to power systems (Stanojev et al., 2023: 1–7).

The reward function is defined as realized hourly profit:

$$R_h = (P_h^* - MC_s) \cdot Q_{s,h}^{accepted} \tag{6}$$

Equation (6) captures realized profit, where revenue depends on the clearing price and accepted quantity, while costs are approximated by the marginal cost proxy. The learning algorithm initially employed classical Q-learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[R + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \tag{7}$$

The update rule adjusts action values based on observed rewards and estimated future returns, with the maximization term introducing potential overestimation bias. Classical Q-learning has been previously applied to electricity bidding problems (Xiong et al., 2002: 1516–1521; van Hasselt, 2010: 2613–2621). To address this issue, the model adopts Double Q-learning, which separates action selection from value evaluation:

$$Q_1(s,a) \leftarrow Q_1(s,a) + \alpha \left[R + \gamma Q_2(s', \arg\max_{a'} Q_1(s',a')) - Q_1(s,a)\right] \tag{8}$$

By separating action selection and evaluation, this formulation (8) reduces overestimation bias and improves stability in stochastic environments. A symmetric update applies to Q_2. Bias reduction mechanisms are particularly relevant in competitive multi-agent contexts where value overestimation may lead to unstable bidding behavior (van Hasselt, 2010: 2613–2621; Du et al., 2021: 534–544).

Exploration follows an ε-greedy schedule with decay during training. After convergence, exploration is disabled during out-of-sample evaluation.

To improve clarity of the proposed framework, the overall workflow of the model is illustrated in Fig. 1. The diagram summarizes the interaction between data input, market reconstruction, and reinforcement learning-based strategic decision-making.

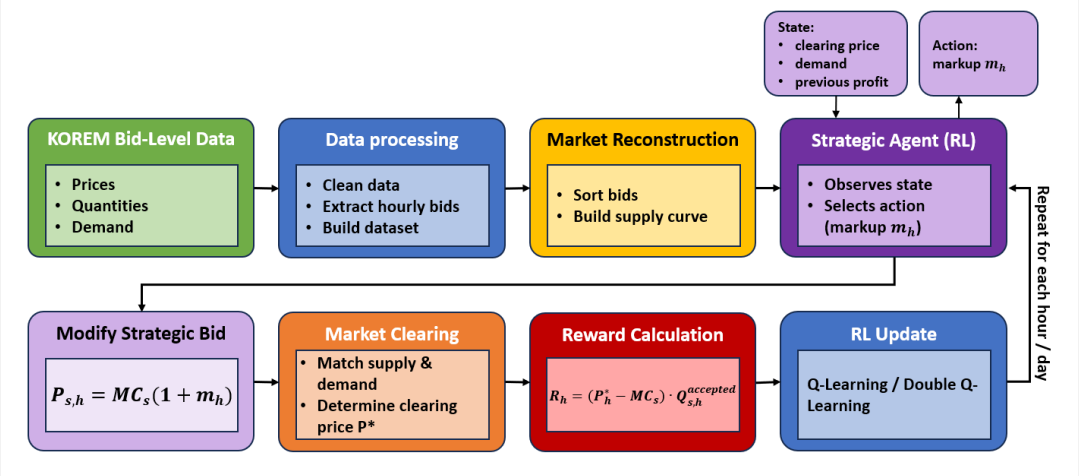


Figure 1: Workflow of the proposed agent-based reinforcement learning framework, illustrating the interaction between real bid-level data, market reconstruction, strategic agent decision-making, and learning through iterative updates

Training and Evaluation Protocol. The dataset is divided chronologically into training (80%) and testing (20%) segments. Chronological splitting avoids look-ahead bias and aligns with empirical validation practices in electricity market modeling (Di Persio et al., 2025). The reinforcement learning agent updates its policy sequentially across training days, interacting with reconstructed auction outcomes.

A central contribution of this study is its direct empirical grounding in real day-ahead electricity market data from the Kazakhstan Electricity and Power Market Operator (KOREM). Rather than relying on stylized synthetic datasets or abstracted cost curves, the proposed agent-based framework is constructed from historical bid-level data, allowing strategic adaptation to be evaluated within a realistic institutional and competitive environment.

Dataset and Market Coverage. The empirical dataset comprises bid-level submissions from the KOREM day-ahead electricity market covering days between July 2023 and November 2025. This period spans approximately two years of continuous day-ahead market operations, capturing seasonal variation, demand cycles, and heterogeneous bidding behavior across participants.

For each delivery day and hour, the dataset includes individual bidder identifiers, submitted bid prices, offered quantities, delivery hour, and delivery date. In addition, hourly system demand values are available. The dataset therefore enables full reconstruction of the supply curve for each trading hour.

Importantly, historical clearing prices were not directly used in the modeling process. Instead, clearing prices were reconstructed from bid-level data using the official pay-as-cleared mechanism. This approach ensures that any strategic modification of bids produces internally consistent clearing outcomes rather than relying on externally imposed historical prices.

Reconstruction of Market Clearing. For each delivery date and hour, the clearing procedure was reconstructed to replicate the pay-as-cleared uniform pricing mechanism used in day-ahead markets. All submitted bids were first aggregated and sorted in ascending order by price. Quantities were cumulatively summed until the total supply equaled or exceeded hourly demand. The price associated with the marginal accepted bid was identified as the market-clearing price for that hour.

This reconstruction ensures that the simulated market clearing process mirrors the actual auction logic of the KOREM day-ahead market. By reconstructing clearing prices rather than importing them directly, the framework allows strategic bid modifications to affect price formation in a structurally consistent manner.

Marginal Cost Proxy Construction. True marginal costs of generators are not publicly available within the dataset. To enable modeling of strategic markups over marginal cost, a proxy marginal cost was constructed for each bidder.

For each generator, marginal cost was approximated using the 10th percentile of that bidder's historical bid prices across the dataset. The rationale for this approach is that the lower tail of historical bids likely reflects cost-based offers submitted during periods of low demand or intense competition. Using a percentile rather than the minimum bid avoids distortions caused by occasional anomalous or extreme low-price entries.

This proxy introduces an approximation, as actual production costs are not observed directly. However, it allows strategic markups to be defined relative to empirically grounded bid behavior rather than synthetic cost curves. The implications of this assumption are discussed explicitly in the limitations section.

Strategic Agent Design within Real Market Environment. The reconstructed market includes all historical bidders observed in the dataset. However, only one bidder is modeled as a strategic reinforcement learning agent. All remaining participants retain their historical bids unchanged.

This modeling choice results in a hybrid market structure in which the competitive environment is fully empirical, while a single generator adapts its bidding behavior. The strategic agent replaces its historical bid price with a markup-based formulation defined as a function of its proxy marginal cost, while maintaining historically observed quantities. All other bidders act as passive participants, submitting the same prices and quantities recorded historically.

This design isolates the effect of reinforcement learning within a realistic competitive landscape. Because other bidders do not adapt strategically in the simulation, the learning dynamics of the focal agent can be interpreted without multi-agent interaction effects confounding the results.

Training and Evaluation Protocol. To prevent look-ahead bias and ensure credible empirical evaluation, the dataset was divided chronologically into training and testing segments. The first 80 percent of delivery days were used for training the reinforcement learning agent. The remaining 20 percent were reserved exclusively for out-of-sample evaluation.

The learning process was conducted sequentially across training days, allowing the agent to update its policy through repeated exposure to historical market conditions. After training convergence was achieved, the learned policy was fixed. Performance evaluation was then conducted on the test subset without further learning or parameter adjustment.

This chronological split ensures that performance improvements reflect genuine predictive adaptation rather than overfitting to future data.

Reinforcement Learning Implementation. The strategic agent employs value-based reinforcement learning. Initial implementation used classical Q-learning, a method previously applied to electricity supplier bidding problems (Xiong et al., 2002: 1516–1521; van Hasselt, 2010: 2613–2621). However, classical Q-learning is known to suffer from overestimation bias in stochastic environments, which may lead to aggressive action selection.

To mitigate this issue, the model adopts Double Q-learning, which separates action selection from evaluation and reduces upward bias in value estimation. Bias reduction is particularly relevant in auction-based markets where reward signals depend on stochastic clearing outcomes and competitor behavior.

Exploration is implemented using an ε-greedy policy with gradual decay. During the testing phase, exploration is disabled to evaluate deterministic policy performance.

The reinforcement learning configuration is designed to balance profitability improvement with structural stability, reflecting concerns raised in reinforcement learning applications to power systems regarding convergence and volatility (Stanojev et al., 2023: 1–7).

Results

Baseline Benchmark. The baseline configuration represents non-adaptive cost-based bidding by the strategic generator, while all other participants submit their historical bids unchanged. Market clearing is reconstructed through the pay-as-cleared merit-order mechanism described earlier, ensuring institutional consistency with wholesale pool markets (Biggar and Hesamzadeh, 2014: 119–138).

Out-of-sample evaluation on the 20% test dataset yields cumulative profit for the strategic agent of:

$$\text{Profit}_{\text{baseline}} = 21{,}491{,}955$$

The average clearing price during the test period equals:

$$P^{-}_{\text{baseline}} = 29.6008$$

These values establish the empirical reference against which reinforcement learning performance is evaluated and are presented in comparison in Table 2.

Table 2 - Out-of-Sample Performance Comparison Between Baseline and RL Configurations

Metric	Baseline	RL	Change
Total Profit	21,491,955	21,571,138	+0.37%
Mean Clearing Price	29.6008	29.6920	+0.31%

Reinforcement Learning Performance. Under Double Q-learning, the strategic agent adapts its markup decisions based on realized profit signals and observed market states. Out-of-sample cumulative profit increases to:

$$\text{Profit}_{\text{RL}} = 21{,}571{,}138$$

This corresponds to an absolute improvement of 79,183 monetary units and a relative increase of +0.37%.

Although moderate in percentage terms, such improvements are economically meaningful in wholesale electricity markets operating at large scale. Importantly, the improvement is achieved without extreme markup escalation, reflecting bounded strategic adaptation consistent with residual demand theory (Biggar and Hesamzadeh, 2014: 281–306).

The stabilization achieved through Double Q-learning is particularly relevant given known overestimation bias issues in classical value-based reinforcement learning (van Hasselt, 2010: 2613–2621). Earlier experiments with standard Q-learning produced higher short-run profit gains but introduced instability. The final configuration prioritizes structural plausibility over aggressive exploitation. The comparison between baseline and reinforcement learning performance is illustrated in Fig. 2.

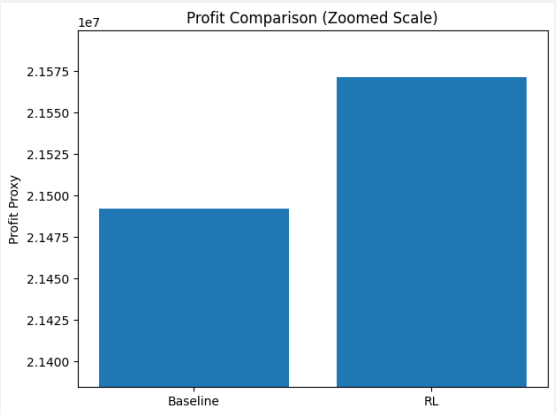


Figure 2: Out-of-sample cumulative profit comparison between baseline and Double Q-learning configurations

Clearing Price Distribution and Structural Stability. A critical concern in applying reinforcement learning to auction-based electricity markets is the potential distortion of price distributions. Auction microstructure studies demonstrate that market design strongly influences volatility and efficiency (Nicolaisen et al., 2001: 504–523; Bower and Bunn, 2000: 1–29). Therefore, reinforcement learning performance must be evaluated not only in terms of profit but also in terms of structural stability.

The empirical distribution of clearing prices under both configurations was compared using the reconstructed pay-as-cleared outcomes.

Under Double Q-learning, the distribution remains concentrated within historically observed intervals. No heavy-tailed amplification or abnormal price spikes are introduced. The modal structure associated with peak and off-peak regimes is preserved. Fig. 3 illustrates that reinforcement learning does not distort the overall price distribution.

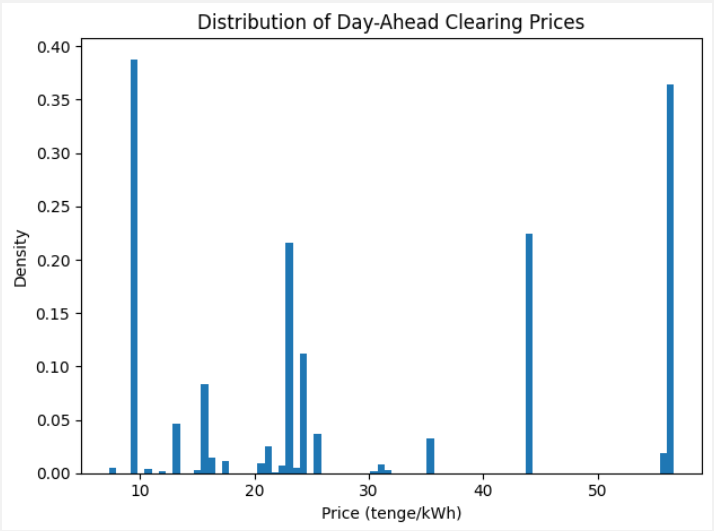


Figure 3: Distribution of day-ahead clearing prices under baseline and RL configurations

Intraday Price Structure. Electricity markets exhibit characteristic intraday price patterns driven by demand cycles and supply stack composition (Kabdygali et al., 2023: 16–29). Agent-based market studies emphasize that preserving such temporal structure is essential for institutional realism (Li and Tesfatsion, 2009: 1–8; Sun and Tesfatsion, 2007: 291–327).

The hourly average clearing price profile remains consistent under reinforcement learning. Prices increase during peak demand hours and decline during off-peak periods in both baseline and RL configurations. The RL scenario exhibits slightly elevated prices during peak periods, reflecting targeted markup adjustments.

The preservation of temporal price structure indicates that adaptive bidding operates within merit-order constraints rather than disrupting the institutional clearing logic. This pattern is illustrated in Figure 4.

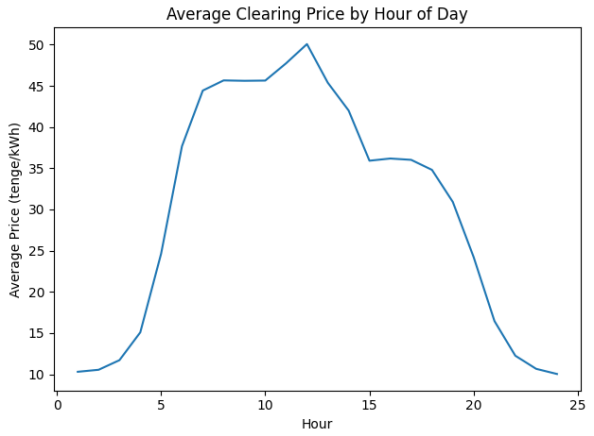


Figure 4: Average clearing price by hour of day

Strategic Markup Dynamics. The observed profit improvement is driven by context-dependent markup adaptation. Reinforcement learning increases markups primarily during hours of elevated demand, when residual demand elasticity permits price influence. During low-demand hours, markups remain minimal, reflecting competitive pressure.

This behavior is consistent with strategic bidding patterns predicted in equilibrium analyses (Contreras et al., 2004: 195–206) and observed in adaptive agent-based simulations (Pinto et al., 2013: 335–346). The bounded nature of markup adjustments demonstrates that Double Q-learning successfully mitigates overestimation bias and prevents unstable escalation.

Discussion

Interpretation of Results. The empirical results demonstrate that reinforcement learning can enhance strategic bidding performance within a reconstructed real-world day-ahead electricity market. However, the magnitude and stability of this improvement are strongly constrained by both the competitive structure of the market and the design of the learning algorithm.

The observed improvement in profitability, although moderate, reflects the realistic conditions of the reconstructed environment. Unlike stylized simulations where agents operate in simplified or synthetic settings, the strategic agent in this study interacts with a real bid stack derived from historical data. As a result, the opportunity for profit maximization is limited by the presence of competing bids and the merit-order dispatch mechanism. This indicates that reinforcement learning does not artificially inflate profits, but instead identifies economically feasible opportunities for strategic adjustment.

A key observation is the temporal distribution of markup behavior. The reinforcement learning agent increases markups primarily during peak-demand hours, while maintaining near-zero markups during off-peak periods. This behavior suggests that the agent implicitly learns the concept of residual demand elasticity, exploiting situations where demand approaches available supply. Such behavior is consistent with theoretical expectations from electricity market economics, where market power is most pronounced under capacity constraints (Biggar and Hesamzadeh, 2014: 281–306).

At the same time, the limited increase in average clearing prices indicates that the agent's influence on overall market outcomes remains constrained. This is an important result, as it demonstrates that reinforcement learning-based strategies do not lead to structural distortion of price formation when applied within a realistic institutional framework. The preservation of price distribution and intraday patterns further supports this conclusion.

As electricity trading becomes increasingly automated, stability and interpretability are essential criteria for evaluating learning-based strategies. The findings indicate that empirically grounded market reconstruction and appropriate algorithmic constraints can support adaptive participation without introducing excessive volatility (Azibek et al., 2024: 1749–1762).

Conclusion

This study evaluated reinforcement learning for strategic bidding in a day-ahead electricity market reconstructed from KOREM bid-level data. One adaptive generator interacted with passive historical participants within pay-as-cleared auctions. Reconstructing clearing outcomes from historical bid stacks preserved competitive depth and institutional structure, providing an empirical basis for evaluating strategic behavior.

The empirical grounding of the model plays a central role in these findings. By reconstructing clearing outcomes directly from bid-level data rather than relying on synthetic supply curves or predefined price series, the framework evaluates adaptive behavior under actual competitive conditions. This ensures that the observed results reflect realistic market dynamics rather than artifacts of simplified modeling assumptions.

Adaptive bidding improved profitability without fundamentally altering price formation in the reconstructed market. The agent increased markups during periods of high demand, while the competitive supply stack constrained excessive price manipulation. Bounded reinforcement learning may support strategic decisions within the existing auction rules.

Double Q-learning increased out-of-sample profit by approximately 0.37% relative to cost-based bidding. Although classical Q-learning produced larger preliminary short-term gains, its unstable markups and clearing prices highlighted the consequences of value overestimation. Separating action selection and evaluation improved convergence and supported bounded, economically plausible bidding.

References

Azibek, B., Zhakiyev, N., Kushekkaliyev, A., Zhalgas, A., & Mukatov, B. (2024). Optimal allocation of storage capacity in distribution network for renewable energy expansion. Electric Power Components and Systems, 52(10), 1749–1762. https://doi.org/10.1080/15325008.2023.2276835

Baíllo, A., Ventosa, M., Ramos, A., Rivier, M., & Canseco, A. (2001). Strategic unit commitment for generation in deregulated electricity markets. In B. F. Hobbs, M. H. Rothkopf, R. P. O'Neill, & H.-P. Chao (Eds.), The next generation of electric power unit commitment models (pp. 227–248). Springer. https://doi.org/10.1007/0-306-47663-0_13

Biggar, D. R., & Hesamzadeh, M. R. (2014). The economics of electricity markets. John Wiley & Sons. https://doi.org/10.1002/9781118775745

Bower, J., & Bunn, D. W. (2000). Model-based comparisons of pool and bilateral markets for electricity. The Energy Journal, 21(3), 1–29. https://doi.org/10.5547/ISSN0195-6574-EJ-Vol21-No3-1

Contreras, J., Klusch, M., & Krawczyk, J. B. (2004). Numerical solutions to Nash–Cournot equilibria in coupled constraint electricity markets. IEEE Transactions on Power Systems, 19(1), 195–206. https://doi.org/10.1109/TPWRS.2003.820692

Dehghanpour, K., Nehrir, M. H., Sheppard, J. W., & Kelly, N. C. (2016). Agent-based modeling in electrical energy markets using dynamic Bayesian networks. IEEE Transactions on Power Systems, 31(6), 4744–4754. https://doi.org/10.1109/TPWRS.2016.2524678

Di Persio, L., Garbelli, M., & Giordano, L. M. (2025). Reinforcement learning for bidding strategy optimization in day-ahead energy market. Energy Economics, 149, Article 108673. https://doi.org/10.1016/j.eneco.2025.108673

Du, Y., Li, F., Zandi, H., & Xue, Y. (2021). Approximating Nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning. Journal of Modern Power Systems and Clean Energy, 9(3), 534–544. https://doi.org/10.35833/MPCE.2020.000502

Guerci, E., Rastegar, M. A., & Cincotti, S. (2010). Agent-based modeling and simulation of competitive wholesale electricity markets. In S. Rebennack, P. M. Pardalos, M. V. F. Pereira, & N. A. Iliadis (Eds.), Handbook of power systems II (pp. 241–286). Springer. https://doi.org/10.1007/978-3-642-12686-4_9

Jain, P., & Saxena, A. (2023). A multi-agent based simulator for strategic bidding in day-ahead energy market. Sustainable Energy, Grids and Networks, 33, Article 100979. https://doi.org/10.1016/j.segan.2022.100979

Kabdygali, S., Omirgaliyev, R., Tursynbayev, T., Kayisli, K., & Zhakiyev, N. (2023). Deep recurrent neural networks in energy demand forecasting: A case study of Kazakhstan's electrical consumption. Scientific Journal of Astana IT University, 16, 16–29. https://doi.org/10.37943/16YIKA8050

Khamzina, A., Ten, A., Mukatov, B., Assylkhan, A., Omirgaliyev, R., Mankovska, N., & Zhakiyev, N. (2025). Critical analysis of tariff policy and legislative measures for renewable energy development: Medium-term challenges and prospects of Kazakhstan. ES Energy and Environment, 28, Article 1560. https://doi.org/10.30919/ee1560

Li, H., & Tesfatsion, L. (2009). The AMES wholesale power market test bed: A computational laboratory for research, teaching, and training. In 2009 IEEE Power & Energy Society General Meeting (pp. 1–8). IEEE. https://doi.org/10.1109/PES.2009.5275969

Nicolaisen, J., Petrov, V., & Tesfatsion, L. (2001). Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. IEEE Transactions on Evolutionary Computation, 5(5), 504–523. https://doi.org/10.1109/4235.956714

Pinto, T., Praça, I., Vale, Z., Morais, H., & Sousa, T. M. (2013). Strategic bidding in electricity markets: An agent-based simulator with game theory for scenario analysis. Integrated Computer-Aided Engineering, 20(4), 335–346. https://doi.org/10.3233/ICA-130438

Schimeczek, C., Nienhaus, K., Frey, U., Sperber, E., Sarfarazi, S., Nitsch, F., Kochems, J., & El Ghazi, A. A. (2023). AMIRIS: Agent-based Market model for the Investigation of Renewable and Integrated energy Systems. Journal of Open Source Software, 8(84), Article 5041. https://doi.org/10.21105/joss.05041

Stanojev, O., Mitridati, L., Prata, R. D. N. D., & Hug, G. (2023). Safe reinforcement learning for strategic bidding of virtual power plants in day-ahead markets. In 2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) (pp. 1–7). IEEE. https://doi.org/10.1109/SmartGridComm57358.2023.10333971

Sun, J., & Tesfatsion, L. (2007). Dynamic testing of wholesale power market designs: An open-source agent-based framework. Computational Economics, 30(3), 291–327. https://doi.org/10.1007/s10614-007-9095-1

Tesfatsion, L. (2002). Agent-based computational economics: Growing economies from the bottom up. Artificial Life, 8(1), 55–82. https://doi.org/10.1162/106454602753694765

Tesfatsion, L. (2003). Agent-based computational economics: Modeling economies as complex adaptive systems. Information Sciences, 149(4), 262–268. https://doi.org/10.1016/S0020-0255(02)00280-3

van Hasselt, H. (2010). Double Q-learning. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), Advances in neural information processing systems 23 (pp. 2613–2621). Curran Associates. https://papers.nips.cc/paper/3964-double-q-learning.pdf

Ventosa, M., Baíllo, A., Ramos, A., & Rivier, M. (2005). Electricity market modeling trends. Energy Policy, 33(7), 897–913. https://doi.org/10.1016/j.enpol.2003.10.013

Weidlich, A., & Veit, D. (2008). A critical survey of agent-based wholesale electricity market models. Energy Economics, 30(4), 1728–1759. https://doi.org/10.1016/j.eneco.2008.01.003

Xiong, G., Hashiyama, T., & Okuma, S. (2002). An electricity supplier bidding strategy through Q-learning. In 2002 IEEE Power Engineering Society Summer Meeting (Vol. 3, pp. 1516–1521). IEEE. https://doi.org/10.1109/PESS.2002.1043645

Ye, Y., Qiu, D., Sun, M., Papadaskalopoulos, D., & Strbac, G. (2020). Deep reinforcement learning for strategic bidding in electricity markets. IEEE Transactions on Smart Grid, 11(2), 1343–1353. https://doi.org/10.1109/TSG.2019.2936142