# The Bidding Games: Reinforcement Learning for MEV Extraction on Polygon Blockchain


Andrei Seoev*
andrei.seoev@mev-x.com
MEV-X
Moscow, Russia

Leonid Gremyachikh*
Independent Researcher
Moscow, Russia
lgremyachikh@gmail.com

Anastasiia Smirnova†
Moscow Institute of Physics and Technology
Moscow, Russia
smirnova@mev-x.com

Yash Madhwal
Skolkovo Institute of Science and Technology
Moscow, Russia
yash.madhwal@skoltech.ru

Alisa Kalacheva
Moscow Institute of Physics and Technology
Moscow, Russia
kalacheva.ap@phystech.edu

Dmitry Belousov
Moscow Institute of Physics and Technology
Moscow, Russia
belousov.da@phystech.edu

Ilia Zubov
Moscow Institute of Physics and Technology
Moscow, Russia
ilia.zubov@phystech.edu

Aleksei Smirnov
MEV-X
Moscow, Russia
aleksei.smirnov@mev-x.com

Denis Fedyanin
HSE University
Moscow, Russia
dfedyanin@hse.ru

Vladimir Gorgadze
Moscow Institute of Physics and Technology
Moscow, Russia
vladimir.gorgadze@phystech.edu

Yury Yanovich
Skolkovo Institute of Science and Technology
Moscow, Russia
y.yanovich@skoltech.ru



## Abstract

In blockchain networks, the strategic ordering of transactions within blocks has emerged as a significant source of profit extraction, known as Maximal Extractable Value (MEV). The transition from spam-based Priority Gas Auctions to structured auction mechanisms like Polygon Atlas has transformed MEV extraction from public bidding wars into sealed-bid competitions under extreme time constraints. While this shift reduces network congestion, it introduces complex strategic challenges where searchers must make optimal bidding decisions within a sub-second window without knowledge of competitor behavior or presence. Traditional game-theoretic approaches struggle in this high-frequency, partially observable environment due to their reliance on complete information and static equilibrium assumptions. We present a reinforcement learning framework for MEV extraction on Polygon Atlas and make three contributions: (1) A novel simulation environment that accurately models the stochastic arrival of arbitrage opportunities and probabilistic competition in Atlas auctions; (2) A PPO-based bidding agent optimized for real-time constraints, capable of adaptive strategy formulation in continuous action spaces while maintaining production-ready inference speeds; (3) Empirical validation demonstrating our history-conditioned agent captures 49% of available profits when deployed alongside existing searchers and 81% when replacing the market leader, significantly outperforming static bidding strategies. Our work establishes that reinforcement learning provides a critical advantage in high-frequency MEV environments where traditional optimization methods fail, offering immediate value for industrial participants and protocol designers alike.



## CCS Concepts

• **Computer systems organization** → **Distributed architectures**; • **Computing methodologies** → *Reinforcement learning*; • **Applied computing** → Electronic commerce.

## Keywords

Maximal Extractable Value, Reinforcement Learning, Blockchain Auctions, Decentralized Finance, Auction Systems

**ACM Reference Format:**
Andrei Seoev, Leonid Gremyachikh, Anastasiia Smirnova, Yash Madhwal, Alisa Kalacheva, Dmitry Belousov, Ilia Zubov, Aleksei Smirnov, Denis Fedyanin, Vladimir Gorgadze, and Yury Yanovich. 2018. The Bidding Games: Reinforcement Learning for MEV Extraction on Polygon Blockchain. In


*Both authors contributed equally to this research.
†Anastasiia Smirnova is also with MEV-X, Moscow, Russia.



*Conference acronym 'XX, Woodstock, NY*



ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

*Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

The emergence of decentralized finance [20] has transformed blockchain ecosystems [18, 24] into dynamic financial markets where transaction ordering represents a fundamental source of value extraction [8]. Unlike traditional financial systems with established sequencing rules, blockchain consensus mechanisms grant block producers significant discretion over transaction inclusion and ordering, creating competitive markets for block space positioning. This competition is particularly pronounced on Ethereum [1] and its Layer-2 scaling solutions [14] like Polygon [1, 13], where the rapid growth of decentralized exchanges (DEX) [10] and lending protocols [4, 9, 16] has created fleeting arbitrage opportunities worth millions annually [21].

Maximal Extractable Value (MEV) represents the economic value achievable through strategic transaction ordering, with atomic arbitrage (AA) constituting a primary revenue source. These opportunities emerge when price discrepancies exist across DEX, allowing sophisticated actors (searchers) to execute risk-free profits through carefully sequenced transactions. Searchers aim to execute their transaction explicitly after the opportunity transaction. The ephemeral nature of these opportunities–often lasting less than a second–creates intense competition among searchers vying for optimal positioning within blocks.

The evolution from spam-based transaction inclusion mechanisms to structured auction systems represents a paradigm shift in MEV extraction dynamics [6, 19]. Traditional Priority Gas Auctions (PGAs) encouraged network-congesting bidding wars where searchers would spam the network with multiple identical transactions at incrementally higher gas prices–where gas represents the computational units required for blockchain execution–creating significant externalities through network congestion and wasted computational resources. Polygon's Atlas upgrade [22] introduces a fundamental redesign through sealed-bid (private bids without competitor visibility), per-opportunity auctions via its FastLane mechanism. This transformation repositions transaction prioritization from public gas wars to private valuation games, where bribes become strategic bids in a first-price auction format. The sealed-bid nature fundamentally alters the information environment–where PGAs allowed real-time observation of competitor behavior, Atlas auctions force participants to make one-shot decisions under extreme uncertainty about both competitor presence and their bidding strategies.

The core industrial challenge lies in solving this complex bidding game within sub-second decision windows. Polygon searchers must simultaneously: (1) detect and estimate the MEV of AA opportunities through real-time mempool monitoring (the temporary storage of pending transactions), (2) calculate optimal bid amounts considering uncertain competition, (3) construct valid transaction bundles, and (4) submit bids within ≈250ms auction windows–all while avoiding the winner's curse of overbidding (systematic overpayment due to incomplete competitor information). This high-frequency, partially observable environment demands sophisticated bidding strategies that traditional game-theoretic approaches struggle to address due to their assumptions of complete information and static equilibria.

This paper presents a comprehensive framework for MEV extraction on Polygon Atlas' FastLane mechanism (dedicated MEV auction channel) through reinforcement learning, addressing the critical gap between theoretical auction models and industrial practice. Our work bridges this divide through both methodological innovations and empirical analysis grounded in real-world blockchain data. The primary contributions of this research are:

(1) **A novel simulation framework** for Polygon Atlas auctions that accurately models the stochastic arrival of opportunities, probabilistic competition, providing a realistic training environment for bidding strategy development.
(2) **A Reinforcement Learning-based bidding agent** optimized for real-time constraints using Proximal Policy Optimization (PPO), capable of adaptive strategy formulation in continuous action spaces while maintaining the inference speed necessary for production deployment in high-frequency auction environments.
(3) **Empirical validation** showing our history-conditioned agent captures 49% of available profits when deployed as a competitor alongside existing searchers, and **81%** when replacing the market leader, demonstrating immediate industrial value and substantial market share capture potential.

## 2 Related Work

**The Polygon Atlas upgrade** [22] introduces a structured protocol, known as FastLane, for transparent MEV extraction via sealed-bid, per-opportunity auctions. This mechanism represents a significant shift from the spam-based PGAs of the past. For each publicly observed transaction that creates an MEV opportunity (**OppTx**), Atlas instantiates a short-lived auction where searchers submit private bids for the right to have their arbitrage transaction included immediately after the OppTx.

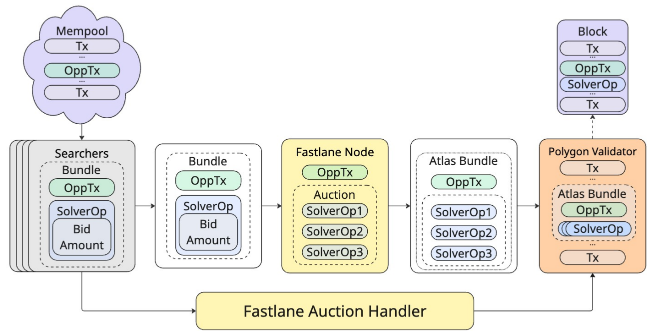


**Figure 1: Empirical transaction execution flow observed in the Polygon Atlas network.**

Our analysis of the live network reveals the following core execution flow (Figure 1), which emphasizes the practical constraints faced by searchers:

(1) A searcher detects an OppTx and must rapidly construct a bundle containing a signed **SolverOperation** (their bid and execution logic) and the OppTx hash.
(2) The bundle is submitted to a FastLane node, which, in practice, opens an auction window of approximately 250 ms. This effective window is a critical constraint dictated by network latency and validator processing.
(3) During this brief window, the node collects and validates competing bundles for the same OppTx. The number of competitors in any given auction is probabilistic and unobservable to participants.
(4) After the window closes, the node forwards the winning bundle to the validator. In the observed flow, the winning searcher's transaction is placed directly after the OppTx in the block.
(5) The **Auction Handler** smart contract executes the operation, collecting the bid amount. A critical operational component is **atlETH**, which searchers must bond (deposit as anti-spam collateral) to participate, creating a real capital requirement and cost for bidding.

From a strategic perspective, the empirical operation of Atlas creates a partially observable, sealed-bid auction with extreme time constraints. Searchers must decide on a bid without knowledge of competitors' actions, under a sub-second deadline dictated by their own latency and the observed 250 ms auction window. This structure eliminates the dynamic bidding wars of PGAs but introduces a complex game of incomplete information, where the value of the opportunity must be weighed against the unknown competitive landscape. This specific industrial context–a high-frequency, sealed-bid auction under probabilistic competition–is the precise problem our reinforcement learning framework is designed to solve.

**Blockchain Auctions:** The competition for MEV extraction has been extensively modeled through game-theoretic lenses. The paper [6] formalized Priority Gas Auctions (PGAs) as continuous-time competitions where participants iteratively replace transactions with higher gas fees–effectively creating all-pay auctions with imperfect information. While this framework captures Ethereum's historical dynamics of public bidding wars, its assumptions of observable bid patterns and reactive strategies become obsolete under Polygon Atlas's sealed-bid paradigm where participants submit single, private bids within sub-second windows.

Recent analyses reveal how auction mechanics shape strategic behavior. The paper [23] demonstrates how Ethereum's MEV-Boost auctions favor last-minute bidding strategies under 12-second time constraints, while [19] empirically show how open bidding structures incentivize vertical integration between builders and order-flow providers. However, both studies fundamentally rely on public bid observability–a critical assumption invalidated by Atlas's design that prevents participants from learning competitors' strategies through repeated interactions.

Theoretical extensions to distributed systems further highlight this limitation. Models of DAG-based ledgers and leaderless protocols employ Nash equilibrium concepts requiring either common knowledge of bid histories or repeated strategy updates [7, 17]. Atlas's combination of sealed bids, probabilistic opponent arrival, and sub-300ms decision windows collapses these dynamic games into partially observable Markov decisions. Where traditional game theory assumes players can gradually converge to equilibrium through observation and adaptation, Atlas's temporal and informational constraints demand instantaneous optimal decision-making under uncertainty–a regime where learning-based approaches become essential rather than optional.

This structural analysis reveals a fundamental mismatch: existing game-theoretic models of blockchain auctions presuppose either bid observability (PGAs), temporal flexibility (MEV-Boost), or repeated interactions (DAG strategies). None account for Atlas's combination of sealed bids, extreme time pressure, and probabilistic competition–a gap our reinforcement learning framework directly addresses.

**Reinforcement Learning for Mechanism Design:** Reinforcement learning (RL) has been widely applied to high-frequency online auctions such as display-ad real-time bidding (RTB), where bidders must optimise their actions under budget constraints and limited feedback. Early work modelled the problem as a Markov decision process in which an agent's state encodes remaining budget, remaining auctions and impression features, the action is the bid price, and the reward reflects click or conversion outcomes. Using value-function approximation and dynamic programming, these approaches derived bidding policies that significantly outperformed linear or static strategies on large benchmark datasets and in live deployment, establishing that sequential, learning-based policies can outperform fixed heuristics in high-volume auctions [2]. Building on this foundation, subsequent studies introduced constrained Markov decision process (CMDP) formulations to explicitly incorporate budget limits and alternative objectives, showing on the iPinYou benchmark that CMDP consistently achieves the highest number of clicks within restricted budgets while maintaining competitive cost metrics [11].

Beyond single-agent methods, researchers have turned to multi-agent reinforcement learning (MARL) to capture the strategic interaction among bidders. Distributed Coordinated Multi-Agent Bidding (DCMAB) groups hundreds of thousands of merchants and consumers into "super-agents" and "super-consumers" and applies a MADDPG-style actor-critic with a centralized critic to generate real-time bid adjustments on a Taobao scale. Offline experiments on millions of impressions show that DCMAB outperforms manual bidding, contextual bandits, A2C, and single-agent DDPG in terms of revenue, ROI, and CPA, and demonstrate that coordinated (social) rewards yield higher total revenue than purely self-interested bidding [12]. Related work on multi-channel ad auctions proposes MARL frameworks with reward shaping and simulation-based training to handle large state spaces and limited real-world data [5]. These studies demonstrate that MARL can model competitive bidding dynamics and achieve scalable performance in complex auction environments.

A complementary line of work addresses the "sim-to-real" gap faced by RL-based auto-bidding systems and explores model-design-oriented models. Li et al. formalize an iterative offline RL framework in which multiple auto-bidding agents collect real interaction data in parallel, then train a new policy offline and redeploy it for further data collection. They identify two key contributions: TEE (Trajectory-wise Exploration and Exploitation), which uses parameter-space noise for various high-return trajectories and a robust trajectory-weighting scheme to emphasize good trajectories during training; and SEAS (Safe Exploration by Adaptive Action Selection), an adaptive algorithm guaranteeing the safety of exploratory actions while preserving data set quality. Offline and real-world experiments on Alibaba's display-advertising platform show that TEE+SEAS achieves near-expert performance in a few iterations while satisfying safety constraints [15]. In parallel, Cai et al. generalize classic auction theory by explicitly modeling user response as a Markov Decision Process in repeated internet ad auctions. Instead of treating each impression as a one-shot auction, their model tracks the evolution of a user's click-through rate based on ad quality and designs mechanisms to maximize long-term discounted revenue, producing a Myerson auction with a modified virtual value and algorithmic sampling to learn approximately optimal mechanisms from data [3].

Together, these studies demonstrate how RL–ranging from single-agent MDP formulations through multi-agent actor-critic architectures to offline and mechanism-design-oriented approaches–can outperform fixed heuristics in high-volume, budget-constrained online auctions. Our work extends this literature to sealed-bid MEV extraction on Polygon Atlas, a setting with sub-second deadlines and sparse feedback that differs markedly from advertising auctions yet shares their sequential, competitive structure.

# 3 The MEV Searcher's Infrastructure: An Industrial Perspective

## 3.1 Architectural Overview

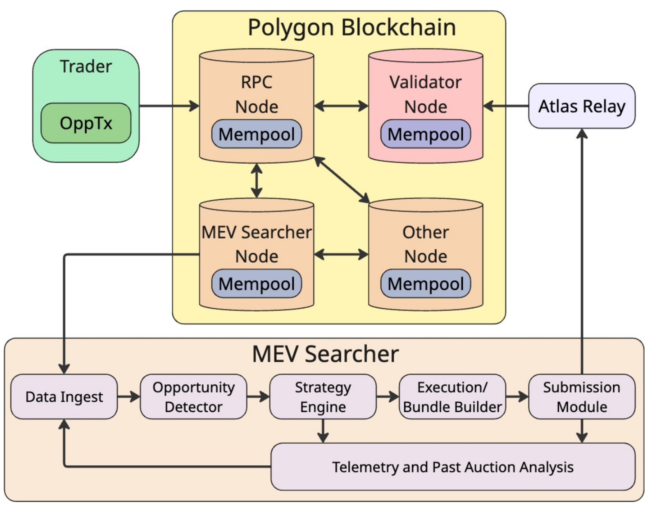


**Figure 2: End-to-end architecture of a professional Polygon MEV searcher system, illustrating data ingestion, opportunity detection, strategy formulation, execution pipeline, and sealed-bid submission mechanics.**

Professional Polygon MEV searchers implement a low-latency, end-to-end pipeline that transforms public mempool signals into executed, incentive-aligned back-runs. Aside from the node and the on-chain solver contract, all components are deployed as high-performance services (e.g., Rust/C++) to meet strict latency and throughput targets. Figure 2 illustrates the architecture.

**Polygon Node and Feeds:** Searchers maintain at least one full Polygon node to obtain an authoritative view of state and pending transactions and to enable deterministic simulation of candidate opportunities. Running an archival node further accelerates initialisation of historical pool states and supports rigorous post-mortem analysis of bot behaviour and performance metrics. The bot co-locates with the node and subscribes to IPC/WebSocket feeds of pending transactions and event logs, complemented by on-demand state reads from pools and oracles, to minimise end-to-end latency.

**Opportunity Detector:** This module filters incoming pending transactions in real time and simulates their effect on AMM state to infer prospective price imbalances and arbitrage opportunities.

**Strategy Engine:** Given an anticipated post-transaction state, the strategy engine computes an optimal arbitrage route, potentially spanning multiple AMMs, that maximizes expected profit net of gas and protocol fees. In our implementation, this is where the reinforcement-learning policy resides.

**Execution Engine and Solver Contract:** The execution engine instantiates the chosen route as an arbitrage transaction and combines it with the bid to form a submission intended to backrun the triggering transaction. Complex swaps are delegated to a dedicated solver smart contract, which executes the route on-chain and settles the winning bid to the protocol upon success.

**Submission Channels:** A dedicated networking module maintains a persistent, stable connection to the Atlas relay, ensuring continuous participation in sealed-bid auctions even under peak network load.

**Telemetry and Past-Auction Analysis:** An internal metrics system records state, actions, latency, and outcomes of each attempt. Offline analysis of past auctions using on-chain data and the Auction Explorer API identifies deficiencies in bot modules and competitor trends, providing training data and guiding the scaling of the searcher system.

This modular architecture enables searchers to ingest high-volume mempool signals, compute optimal strategies under sub-second deadlines, and submit sealed bids reliably to Atlas/FastLane while continuously improving through telemetry-driven analysis.

## 3.2 From Opportunity to Bid

We now detail how a detected opportunity transaction (OppTx) is transformed into a sealed bid under the sub-second deadlines imposed by Polygon Atlas.

**Block Tracking and Mempool Monitoring:** The searcher maintains per-block, synchronized states of all monitored pools while scanning the public mempool in real-time for opportunity transactions. Early detection, coupled with a state derived from the most recent finalised block, is essential: stale pool states lead to simulation errors and mispriced arbitrage, degrading bid quality and auction competitiveness. Conversely, earlier identification of opportunity transactions allows the searcher to trigger auctions and set deadlines that disadvantage slower competitors.

**Prefiltering:** Before expensive state simulation, incoming mempool transactions undergo lightweight heuristic screening to eliminate likely non-viable candidates (e.g., blacklisted addresses/tokens, declared gas limits outside a configured range, trivially short calldata). Given the volume of pending transactions, early rejection prevents unnecessary pipeline work, accelerates throughput, and improves reaction time relative to competitors.

**Simulation:** Transactions that pass prefiltering are executed against the latest known block state on a local node to produce a deterministic trace and a post-transaction pool configuration. Conservative early-exit checks suppress false positives; aggressive strategies may disable some of these guards. Although wall-time remains below the auction window, this stage benefits significantly from caching and adequate hardware.

**Candidate Route Filtering:** Using the simulated post-transaction state, the searcher enumerates cyclic routes traversing affected pools. The set is pruned by a necessary arbitrage condition–the directed product of spot prices along a route must exceed one; violating routes are discarded. Heuristic pruning reduces latency at the cost of possibly missing the global optimum, but typically yields competitive bids within deadline.

**Optimal Input Calculation and Route Selection:** For each remaining route the searcher computes the input size that maximises net profit, typically via numerical optimisation with closed-form solutions for certain pool types. Surviving routes are validated by deterministic forward execution and ranked by peak achievable profit. The top route is chosen, and non-overlapping secondary routes may be greedily bundled to form a composite arbitrage path. This stage trades accuracy for latency: high-precision methods improve per-route optimality but increase compute time, which can endanger the auction deadline.

**SolverOp Formation and Atlas Bundle Transmission:** The selected route is serialised as solver calldata and wrapped into an EIP-712-signed Solver Operation. The resulting Atlas bundle is dispatched over low-latency channels to the nearest ingress/operations relay. The bid is set as a fixed fraction of the expected net profit, configured as a constant in the bot. Professional searchers maintain multiple simultaneous relay connections for load balancing and instant failover to avoid packet loss at this decisive moment.

**Latency as a Binding Constraint:** In this sealed-bid auction the primary benefit of low transmission latency is not real-time competition but the expansion of the computational budget for earlier stages. Faster transmission directly provides more time for complex route search and precise optimal input calculation without risking missed deadlines, enabling more thorough profit maximisation and thus more competitive bids.

## 3.3 Numerical Insights from Polygon

Understanding the industrial landscape of MEV extraction on Polygon requires grounding the problem in empirical data. We analyze a dataset derived from the Polygon network, focusing on atomic arbitrage opportunities captured via the FastLane mechanism (Atlas upgrade) over the observation period from December 2024 to September 2025 containing 223,356 OppTx. This analysis provides crucial quantitative context for the bidding game searchers face.

**Searcher Multiplicity and Dynamics:** The identification of distinct searcher entities relies on analyzing "from" and "to" addresses associated with bids ("from" sends the bid, while "to" accumulates the profit), using graph-based clustering where addresses appearing in the same transaction are considered connected. Our analysis reveals a concentrated participant landscape: the active searcher base does not exceed 15 unique entities per week, with only 17 unique searchers observed over the entire observation period (Figure 3). Notably, participation frequently fluctuates between 5-8 active searchers during most periods, suggesting a core group of professional operators with occasional peripheral participants.

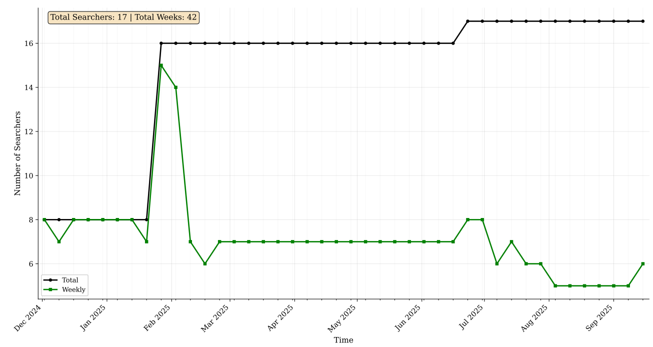


**Figure 3: Weekly statistics on unique MEV searchers, showing consistent participation from 5-8 core entities with occasional peripheral activity.**

This concentration validates the Atlas design assumption of limited concurrent bidders per opportunity. While blockchain anonymity permits address proliferation, the technical requirements (atlETH bonding) limit participation to professional entities.

**Searcher Operational Patterns:** Figure 4 reveals the operational evolution of a representative professional searcher over time. The visualization displays each address as a horizontal timeline spanning from its first to last observed participation week. "From" addresses (upper green bars) and "to" addresses (lower blue bars) are annotated with their average bribe percentage and total bid participation count.

The temporal analysis reveals several key operational insights: significant overlap among "from" addresses with minimal "to" address rotation indicates parallel deployment of multiple bot instances, suggesting sophisticated operational infrastructure. Furthermore, the searcher's activity cessation in August 2025 coincides with broader market consolidation that reduced active participants from the historical peak of 17 to just 5-6 persistent entities, potentially reflecting declining profitability due to intensified competition. Most crucially, the figure demonstrates the winner's curse through progressive bid inflation–addresses activated later in the observation period exhibit systematically higher mean bribe percentages, suggesting experienced searchers increasingly prioritized win probability over profitability, a behavioral pattern our RL framework specifically addresses through adaptive bidding strategies.

**Opportunity Complexity and Competition:** Arbitrage opportunities exhibit significant heterogeneity in complexity and competitive dynamics. Figure 5 categorizes opportunities by involved DEX protocols, revealing clear stratification in bidding behavior. Opportunities involving simpler, well-established DEXs attract more frequent participation with lower mean bids, indicating higher competition for standardized arbitrage routes. Conversely, complex

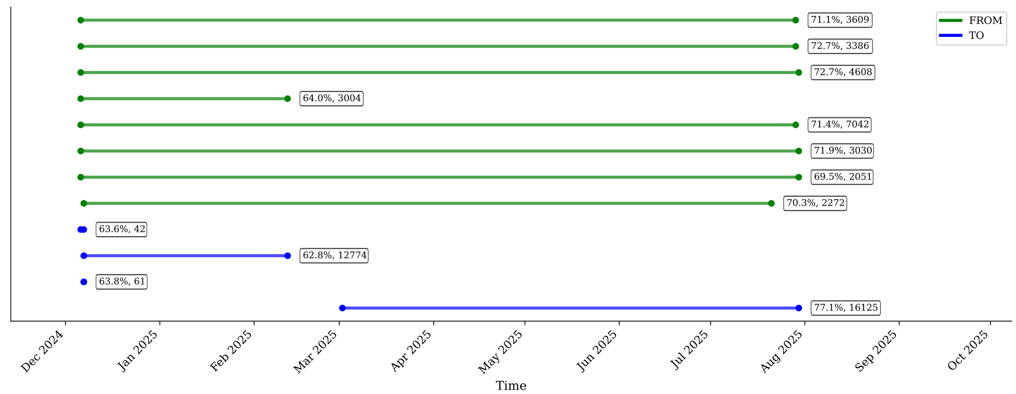


**Figure 4: Temporal activity patterns of a professional searcher's addresses, showing parallel bot deployment, operational timeline, and bid escalation patterns. Address annotations indicate mean bribe percentage and participation count.**

multi-protocol opportunities exhibit higher mean bids but less frequent participation, reflecting specialized expertise requirements and reduced competition.

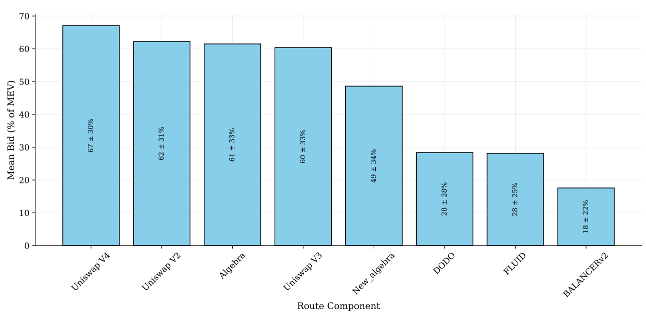


**Figure 5: Bribe percentage distribution by protocol type, showing inverse relationship between opportunity frequency and mean bid amount. Complex protocols attract fewer but more aggressive bidders.**

This stratification has direct implications for bidding strategy design. Complex opportunities demand sophisticated routing logic and risk assessment, creating natural barriers that reduce competition while enabling successful bidders to claim larger profit shares. The observed protocol-specific bid patterns directly inform our state representation in the RL framework, where opportunity complexity serves as a key feature for strategic adaptation across different market segments.

**Industrial Implications:** The empirical patterns observed–market consolidation, operational sophistication through parallel bot deployment, and opportunity stratification–collectively underscore the industrial maturity of MEV extraction on Polygon. The demonstrated winner's curse, evidenced by systematic bid escalation among experienced participants, emphasizes that naive bidding approaches systematically destroy value, creating the essential business case for our learning-based approach.

These numerical insights validate our problem formulation: successful MEV extraction requires adaptive strategies that account for dynamic competition, opportunity heterogeneity, and the fundamental tension between win probability and profitability. Our RL framework directly addresses these industrial realities by learning optimal bidding policies that avoid the observed pitfalls of manual strategy calibration.

## 4 Proposed Framework: Modeling the Auction as a PPO Problem

**Problem Formulation for Continuous-Action Auction Control:** We formulate the MEV bidding problem as a Partially Observable Markov Decision Process (POMDP) to capture the sequential decision-making under uncertainty inherent in Polygon Atlas auctions. The framework consists of three core components listed below.

*State Representation:* The state $x_t$ is a feature vector combining: (1) route-specific features including protocol types, route length, and historical frequency; (2) windowed system statistics such as recent win rates and average competition levels; (3) aggregated opponent actions from recent auctions. This representation captures both the immediate opportunity characteristics and the evolving market context.

*Action Space:* The action $b_t \in [0, 1]$ represents the continuous bribe fraction of the MEV value $v_t$. The absolute bid amount is $b_t v_t$, and upon winning, the net profit becomes $(1 - b_t)v_t$ (excluding minor blockchain network fees). This continuous formulation enables fine-grained strategic adaptation.

*Reward Function:* For auction $i$ with MEV $v_i \geq 0$, the agent bids fraction $b_i$ and competes against the highest competing bid $\tilde{b}_i$. The win indicator $w_i = \mathbb{I}\{b_i > \tilde{b}_i\}$ determines the net profit $\pi_i = w_i(1 - b_i)v_i$. We use a shaped reward to reduce variance: $r_i = \pi_i/(v_i+\varepsilon) - \lambda\mathbb{I}\{w_i = 0\} - \alpha(b_i - \tilde{b}_i)_+$, where $\varepsilon > 0$ provides numerical stability, $\lambda$ penalizes losing bids, and $\alpha$ discourages overbidding relative to the competition threshold.

**Evaluation Metrics:** We employ two complementary metrics for comprehensive performance assessment:

(1) *Win Ratio (WR):* $\mathrm{WR} = \frac{1}{N}\sum_{i=1}^{N} w_i$ measures the frequency of successful auction participation.
(2) *Maximum-Profit Capture (MPC):* $\mathrm{MPC} = \frac{\sum_{i=1}^{N}\pi_i}{\sum_{i=1}^{N}\pi_i^{\star}}$, where $\pi_i^{\star} = (1 - (\tilde{b}_i + \varepsilon))v_i$ represents the counterfactual profit achievable by bidding infinitesimally above the second-highest bid. This metric quantifies efficiency relative to perfect information.

**Algorithm Selection:** We select Proximal Policy Optimization (PPO) as our core learning algorithm based on several critical considerations for industrial deployment: the clipped surrogate objective ensures stable policy updates essential in competitive environments where training data is expensive to collect; PPO naturally accommodates continuous action spaces through Gaussian or Beta policy heads with entropy regularization; through minibatch SGD and Generalized Advantage Estimation (GAE), it achieves strong sample efficiency despite sparse reward signals; conservative on-policy updates provide resilience against non-stationary opponent behavior and noisy rewards; and fewer sensitive hyperparameters with straightforward action bounding facilitate reproducible deployment in production environments.

**Simulator Design for Polygon Atlas:** Our simulation environment captures three key industrial constraints:

(1) *Probabilistic Opponent Arrival:* Models heterogeneous detection rates across searchers using Bernoulli distributions parameterized by route complexity and historical participation patterns.
(2) *Latency Distributions:* Incorporates realistic reaction-time constraints through exponential distributions fitted from observed bid timing data.
(3) *Configurable Information Regimes:* Supports both real-time and delayed bid disclosure scenarios to evaluate strategy robustness under varying transparency conditions.

# 5 Experiments and Evaluation

## 5.1 Experimental Setup

**Data and Preprocessing:** We utilize a comprehensive Polygon Atlas snapshot spanning December 2024 to September 2025, containing 223,356 opportunity transactions. We filter approximately 10% of OppTx with smallest MEV (up to the MATIC equivalent of ≈0.1 United States dollar) as they typically have opportunity value less than blockchain network and Atlas protocol fees, resulting in negative net profit and low commercial interest. After filtering rows without recorded wins and ensuring non-negative MEV values, we sort by block height and employ a chronological 50/50 train-test split to prevent temporal leakage and maintain realistic evaluation conditions.

**Evaluation Scenarios:** We evaluate our agent under two complementary scenarios that capture different aspects of industrial deployment:

*Historical Participation* reflects realistic operational constraints where technical limitations–including message roundtrip delays (typically 50-150ms), network latency variations, and stochastic transaction propagation–prevent even sophisticated searchers from participating in every auction. This scenario measures performance when adding our algorithm as an additional market participant, providing a conservative baseline for real-world deployment impact.

*Market Leader Replacement* assesses direct competitive advantage by removing the dominant incumbent (MEV-X) and substituting our agent in their historical auctions. This demonstrates immediate deployment value: our algorithm can capture substantial market share from the current leader before competitors adapt to the new strategy, representing the most compelling industrial use case.

For both scenarios we compute Sum Profit (SP) and Upper Bound (UB) along with WR and MPC. For Market Leader Replacement we also report MEV-X's State-of-the-art (SOTA) performance as the primary industrial benchmark.

## 5.2 Results and Analysis

**Overall Performance Trends:** As shown in Table 1, our PPO agent demonstrates robust performance across both evaluation settings. We report all monetary values in MATIC, Polygon's native currency used for network fees and representing the standard profit denomination for industrial searchers. The consistent superiority of history-conditioned agents across key metrics validates our approach to incorporating temporal context.

**Historical Participation Analysis:** In the Historical Participation setting, our agents face the same operational constraints as professional searchers. The history-conditioned agent achieves significantly higher profit capture (48.79% vs. 44.90% MPC) despite lower win rates (35.23% vs. 39.94%), indicating more selective and economically efficient bidding behavior. This pattern suggests the temporal-aware agent learns to avoid overbidding in highly competitive auctions while aggressively pursuing undervalued opportunities–a crucial capability for sustainable profitability in production environments.

**Market Leader Replacement** scenario reveals our most compelling industrial result (Figure 6): when substituting for the dominant market participant, our history-conditioned agent captures **143%** of the incumbent's realized profit. This 43% outperformance demonstrates significant market inefficiencies in current bidding strategies that our learning-based approach exploits. The stateless agent's higher win rate (58.46% vs. 44.72%) but lower profit efficiency reveals an important trade-off: aggressive bidding wins more auctions but destroys value through overpayment, while the history-conditioned agent develops more nuanced strategies that optimize long-term profitability.

## 5.3 Environment-Specific Insights

**Stateless Environment:** In the stateless setting where each auction is treated as an independent one-step bandit, the observation space includes compact route representations with length, uniqueness counters, frequency metrics, and one-hot encoded first/last hop protocols. This approach provides a strong baseline but lacks the temporal reasoning necessary for optimal sequential decision-making in dynamic markets.

**History-Conditioned Environment:** To capture short-horizon dynamics without the computational overhead of full recurrent modeling, we develop a history-conditioned environment that provides limited temporal context. For each auction $t$, we maintain

**Table 1: Evaluation on Test Set Across Settings and Environments**

| Setting | Environment | WR | MPC | Sum Profit | UB |
|---|---|---|---|---|---|
| **Historical Participation** | Stateless | **39.94%** | 44.90% | 99,341 | 221,108 |
| | History-Conditioned | 35.23% ± 0.02% | **48.79%** | **107,881 ± 686** | |
| **Market Leader Replacement** | MEV-X (SOTA) | 53.93% | 56.54% | 33,998 | 60126 |
| | Stateless | **58.46%** | 74.95% | 45,070 | |
| | History-Conditioned | 44.72% ± 0.04% | **80.93%** | **48,660 ± 378** | |

a rolling window of size $H = 10$ and sample $K = 5$ historical observations without replacement, preserving temporal order. This approach simulates the partial observability inherent in live trading, where searchers see only subsets of recent market activity due to network delays and memory constraints.

The history-conditioned agent's superior performance across both evaluation settings demonstrates the critical importance of temporal context in competitive bidding environments. By incorporating recent market dynamics, the agent develops more sophisticated bidding strategies that optimize for long-term profitability rather than individual auction wins.

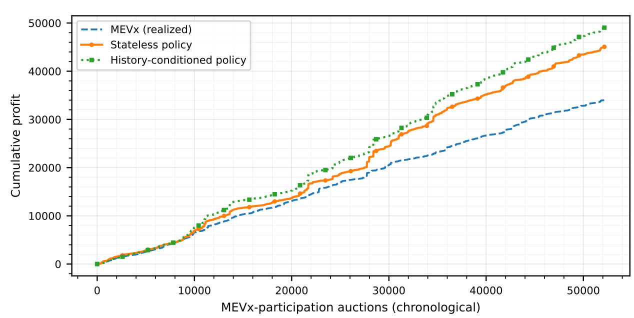


**Figure 6: Cumulative profit comparison on test set, demonstrating the consistent superiority of history-conditioned bidding strategies across the evaluation period.**

## 6 Conclusion

Our experimental results demonstrate the critical advantage of adaptive, learning-based strategies over static bidding approaches in sealed-bid MEV auctions. The history-conditioned PPO agent achieves substantial profit capture relative to the dominant market participant in counterfactual analysis, highlighting the value of sophisticated bidding strategies. For protocol designers, our findings indicate that information disclosure rules significantly impact strategic behavior, with limited historical context potentially improving market efficiency while maintaining auction integrity.

From an ethical perspective, it is important to distinguish atomic arbitrage (AA) from more harmful MEV forms. While front-running and sandwich attacks directly harm traders by manipulating transaction ordering, atomic arbitrage operates as back-running–executing after the trader's transaction without affecting its outcome. Although AA extracts value from market inefficiencies in DEX pools, it simultaneously contributes to market efficiency by eliminating price discrepancies. This aligns with emerging trends in decentralized exchange design, where protocols increasingly adopt batched transaction execution with solvers who distribute arbitrage revenue among traders, liquidity providers, and the protocol itself, creating more equitable value distribution.

Our simulation, while empirically grounded, necessarily simplifies certain market dynamics such as network latency variations and complex multi-block MEV opportunities. For industrial deployment, we address key considerations including sub-millisecond inference speeds that comfortably meet auction window constraints, continuous learning pipelines to mitigate model drift from market regime changes, and safety-constrained exploration to maintain risk management during operation.

Future research directions include extending the framework to other MEV strategies such as liquidations, modeling multi-dimensional strategy spaces with collusion detection, and incorporating cross-chain arbitrage opportunities. In summary, our reinforcement learning framework provides a robust solution to the complex bidding game in Polygon Atlas auctions, combining empirical market analysis with adaptive learning to significantly outperform traditional strategies while advancing automated mechanism design in blockchain environments.

## References


[1] Vitalik Buterin. 2014. Ethereum White Paper: A Next Generation Smart Contract & Decentralized Application Platform. *Ethereum* (2014), 1–36. Issue January. https://github.com/ethereum/wiki/wiki/White-Paper

[2] Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. 2017. Real-Time Bidding by Reinforcement Learning in Display Advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 2017)*. ACM, 661–670. doi:10.1145/3018661.3018702

[3] Yang Cai, Zhe Feng, Christopher Liaw, Aranyak Mehta, and Grigoris Velegkas. 2024. User Response in Ad Auctions: An MDP Formulation of Long-term Revenue Optimization. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*. ACM, 111–122. doi:10.1145/3589334.3645495

[4] Yatipa Chaleenutthawut, Vyacheslav Davydov, Michael Evdokimov, Sudarut Kasemsuk, Stanislav Kruglik, Grigorii Melnikov, and Yury Yanovich. 2024. Loan Portfolio Dataset From MakerDAO Blockchain Project. *IEEE Access* 12 (2024), 24843–24854. doi:10.1109/ACCESS.2024.3363225

[5] Chen Chen, Gao Wang, Baoyu Liu, Siyao Song, Keming Mao, Shiyu Yu, and Jingyu Liu. 2024. Real-time bidding with multi-agent reinforcement learning in multi-channel display advertising. *Neural Computing and Applications* 37, 1 (Nov. 2024), 499–511. doi:10.1007/s00521-024-10649-6

[6] Philip Daian, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels. 2020. Flash Boys 2.0: Frontrunning in Decentralized Exchanges, Miner Extractable Value, and Consensus Instability. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 910–927. doi:10.1109/SP40000.2020.00040

[7] Pranav Garimidi, Lioba Heimbach, and Tim Roughgarden. 2025. Transaction Fee Mechanism Design for Leaderless Blockchain Protocols. *Arxiv* (2025).

[8] Vincent Gramlich, Dennis Jelito, and Johannes Sedlmeir. 2024. Maximal extractable value: Current understanding, categorization, and open research questions. *Electronic Markets* 34 (2024), 49. Issue 1. doi:10.1007/s12525-024-00727-x
[9] Matthias Hafner, Romain de Luze, Nicolas Greber, Juan Beccuti, Benedetto Biondi, Gidon Katten, Michelangelo Riccobene, and Alberto Arrigoni. 2023. DeFi Lending Platform Liquidity Risk: The Example of Folks Finance. *The Journal of The British Blockchain Association* 6 (4 2023), 1–8. Issue 1. doi:10.31585/jbba-6-1-(5)2023
[10] Sascha Hägele. 2024. Centralized exchanges vs. decentralized exchanges in cryptocurrency markets: A systematic literature review. *Electronic Markets* 34 (12 2024), 33. Issue 1. doi:10.1007/s12525-024-00714-2
[11] Arti Jha, Harshit Jain, Parikshit Sharma, Yashvardhan Sharma, and Kamlesh Tiwari. 2024. Optimizing Real-Time Bidding Strategies: An Experimental Analysis of Reinforcement Learning and Machine Learning Techniques. *Procedia Computer Science* 235 (2024), 2017–2026. doi:10.1016/j.procs.2024.04.191
[12] Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. 2018. Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, 2193–2201. doi:10.1145/3269206.3272021
[13] Jaynti Kanani, Sandeep Nailwal, and Anurag Arjun. 2018. Matic Whitepaper Version 1.1. https://github.com/maticnetwork/whitepaper
[14] Stanislav Kruglik, Kamilla Nazirkhanova, and Yury Yanovich. 2019. Challenges beyond blockchain: scaling, oracles and privacy preserving. In *2019 XVI International Symposium "Problems of Redundancy in Information and Control Systems" (REDUNDANCY)*. IEEE, 155–158. doi:10.1109/REDUNDANCY48165.2019.9003331
[15] Haoming Li, Yusen Huo, Shuai Dou, Zhenzhe Zheng, Zhilin Zhang, Chuan Yu, Jian Xu, and Fan Wu. 2024. Trajectory-wise Iterative Reinforcement Learning Framework for Auto-bidding. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*. ACM, 4193–4203. doi:10.1145/3589334.3645534
[16] Ignat Melnikov, Irina Lebedeva, Artem Petrov, and Yury Yanovich. 2025. DeFi risk assessment: MakerDAO loan portfolio case. *Blockchain: Research and Applications* 6 (6 2025), 100259. Issue 2. doi:10.1016/j.bcra.2024.100259
[17] Sebastian Müller and Alexandre Reiffers-Masson. 2025. Game-Theoretic Analysis of Transaction Selection in DAG-Based Distributed Ledgers. In *2025 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 1–8. doi:10.1109/ICBC64466.2025.11114621
[18] Satoshi Nakamoto. 2008. Bitcoin: A Peer-to-Peer Electronic Cash System. *www.bitcoin.org* (2008), 1–9. https://bitcoin.org/bitcoin.pdf
[19] Burak Öz, Danning Sui, Thomas Thiery, and Florian Matthes. 2024. Who Wins Ethereum Block Building Auctions and Why?. In *6th Conference on Advances in Financial Technologies*. 1–22.
[20] Fabian Schär. 2020. Decentralized Finance: On Blockchain- and Smart Contract-based Financial Markets. *SSRN Electronic Journal* (2020). doi:10.2139/ssrn.3571335
[21] Daniil Vostrikov, Yash Madhwal, Andrey Seoev, Anastasiia Smirnova, Yury Yanovich, Alexey Smirnov, and Vladimir Gorgadze. 2025. Unpacking Maximum Extractable Value on Polygon: A Study on Atomic Arbitrage. *Arxiv* (2025). https://arxiv.org/abs/2508.21473
[22] Alex Watts, Jacob Greene, Jeremy Feral, Jordan Hagan, and Ben Sparks. 2024. Atlas. (2024), 9 pages. https://www.fastlane.xyz/Atlas_Whitepaper.pdf
[23] Fei Wu, Thomas Thiery, Stefanos Leonardos, and Carmine Ventre. 2024. Strategic Bidding Wars in On-chain Auctions. In *2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 503–511. doi:10.1109/ICBC59979.2024.10634354
[24] Yury Yanovich, Ivan Ivashchenko, Alex Ostrovsky, Aleksandr Shevchenko, and Aleksei Sidorov. 2018. Exonum: Byzantine fault tolerant protocol for blockchains. *bitfury.com* (2018), 1–36. https://bitfury.com/content/downloads/wp-consensus-181227.pdf

Received 7 October 2025; revised DD Month yyyy; accepted DD Month yyyy