COLLECTIVE INTELLIGENCE | ASSIGNMENT II

HALITE: a MARL Approach

Javier Martín Fuentes

ELTE | FACULTY OF INFORMATICS

IMPORTANT

See **README** for a much detailed explanation + game replay GIFs

Objective

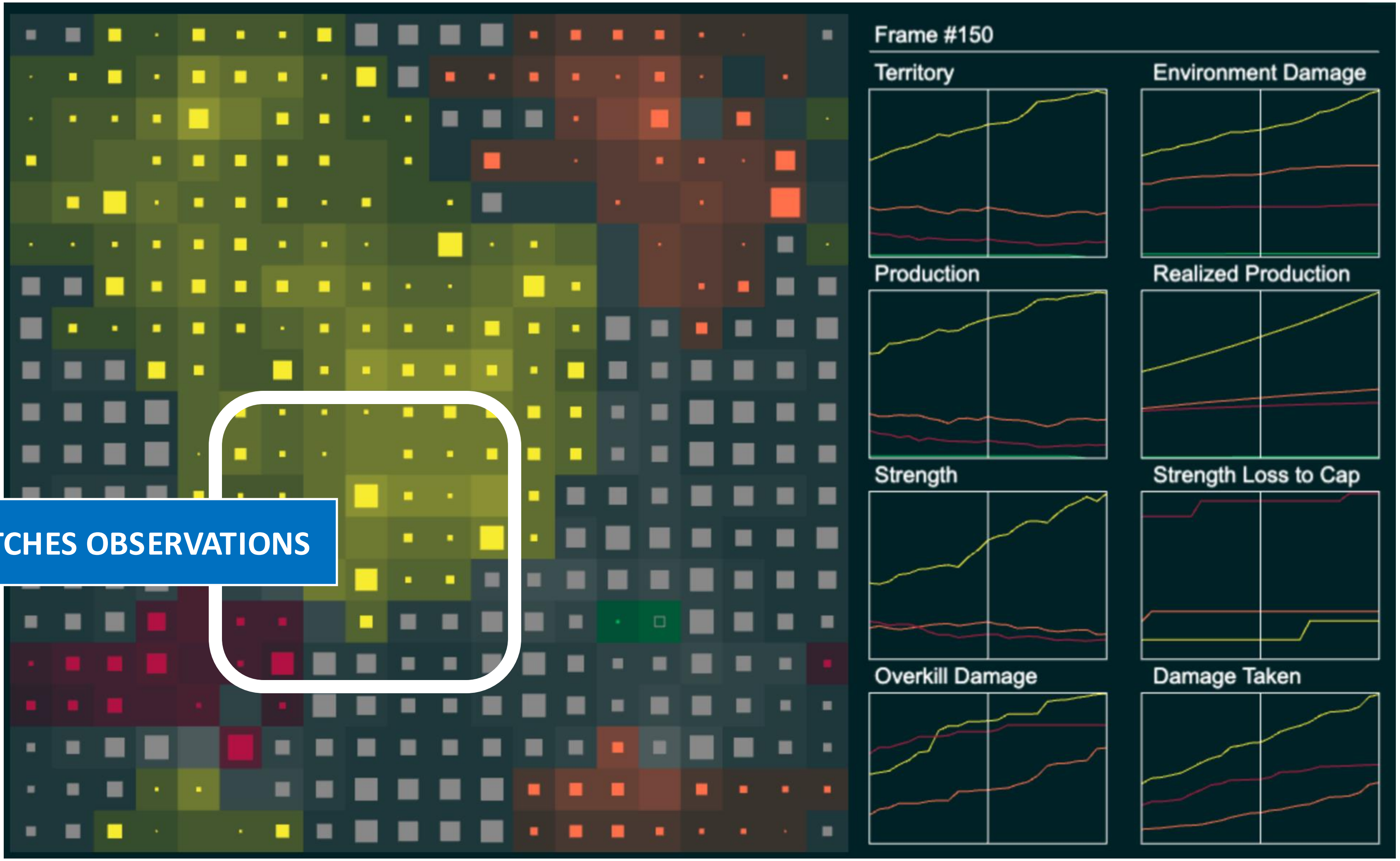
Train **MARL agents** to play **Halite** (territorial control strategy game)

Hypothesis

Reinforcement learning agents can **discover strategies** that are **competitive** with, or surpass, **rule-based bots** by exploiting emergent behaviors.

Different rewards will guide learning towards different strategic behaviors.

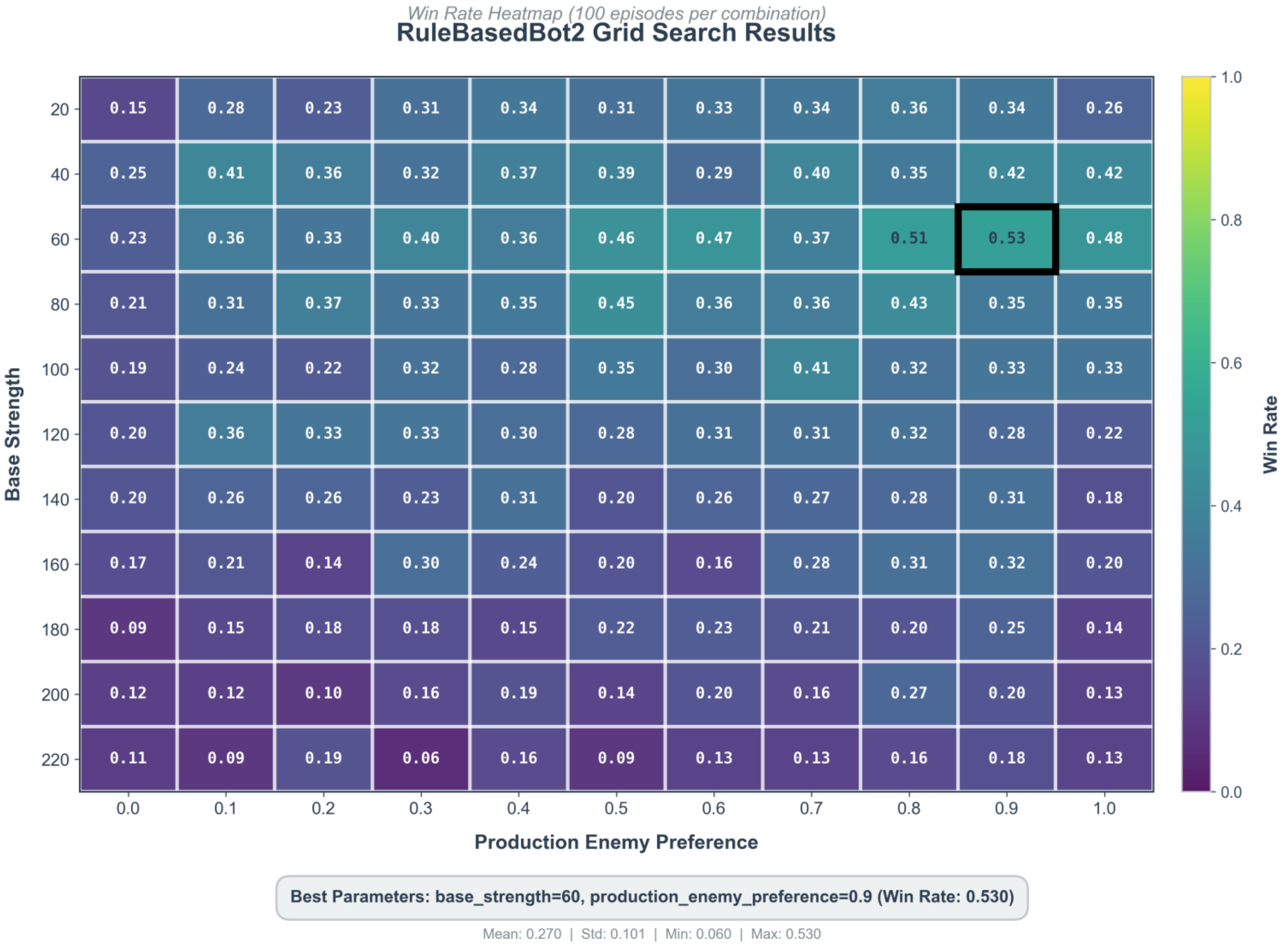
HALITE ENVIRONMENT



7X7 LOCAL PATCHES OBSERVATIONS

Rule-Based Bot

- **Expansion thresholding**
  - Dynamic strength requirements based on distance to frontier
- **Border pressure**
  - Prioritizes frontier cells, distance maps via BFS
- **Regroup/Hold logic**
  - Force combination for inner cells, STILL for accumulation
- **Anti-collision**
  - Conflict resolution (same destination, position swaps, moving targets)
- **Pathfinding**
  - Production-weighted enemy targeting



Win Rate Heatmap (100 episodes per combination)

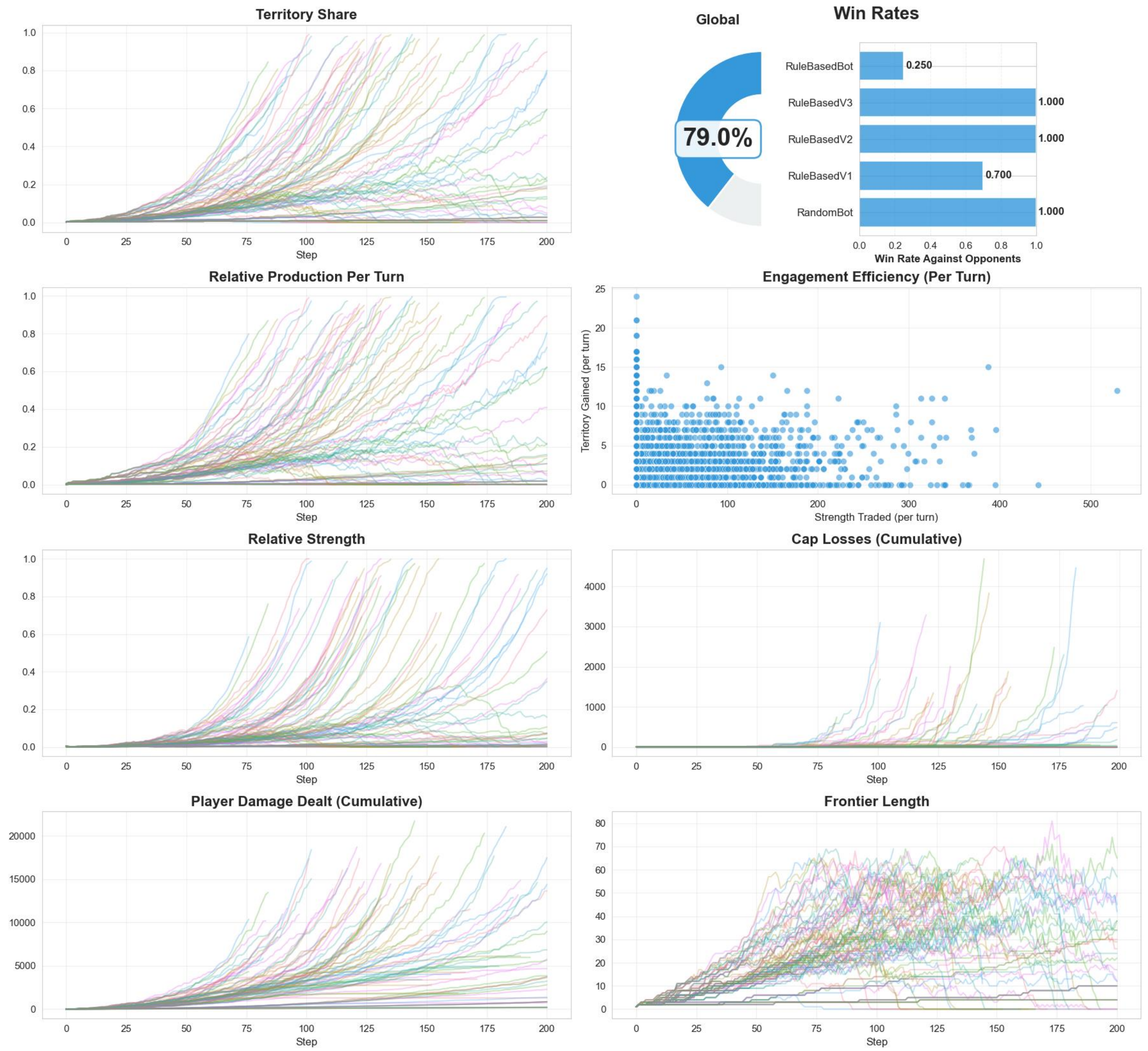
RuleBasedBot2 Grid Search Results

Base Strength	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
20	0.15	0.28	0.23	0.31	0.34	0.31	0.33	0.34	0.36	0.34	0.26
40	0.25	0.41	0.36	0.32	0.37	0.39	0.29	0.40	0.35	0.42	0.42
60	0.23	0.36	0.33	0.40	0.36	0.46	0.47	0.37	0.51	0.53	0.48
80	0.21	0.31	0.37	0.33	0.35	0.45	0.36	0.36	0.43	0.35	0.35
100	0.19	0.24	0.22	0.32	0.28	0.35	0.30	0.41	0.32	0.33	0.33
120	0.20	0.36	0.33	0.33	0.30	0.28	0.31	0.31	0.32	0.28	0.22
140	0.20	0.26	0.26	0.23	0.31	0.20	0.26	0.27	0.28	0.31	0.18
160	0.17	0.21	0.14	0.30	0.24	0.20	0.16	0.28	0.31	0.32	0.20
180	0.09	0.15	0.18	0.18	0.15	0.22	0.23	0.21	0.20	0.25	0.14
200	0.12	0.12	0.10	0.16	0.19	0.14	0.20	0.16	0.27	0.20	0.13
220	0.11	0.09	0.19	0.06	0.16	0.09	0.13	0.13	0.16	0.18	0.13

Best Parameters: base_strength=60, production_enemy_preference=0.9 (Win Rate: 0.530)

Mean: 0.270 | Std: 0.101 | Min: 0.060 | Max: 0.530

Baseline Evaluation for RuleBased



MARL Suite

Centralized Q-Learning (CQL)

- Joint-state centralized critic (all agents' observations)
- Decentralized action selection (epsilon-greedy)
- Off-policy with experience replay

Independent Q-Learning (IQL)

- Independent Q-networks per agent
- Parameter sharing for sample efficiency
- Off-policy with experience replay

IPPO (Independent PPO)

- Independent actor-critic pairs per agent
- Actor: local observations, Critic: global state
- On-policy with PPO clipping + GAE

MAPPO (Multi-Agent PPO)

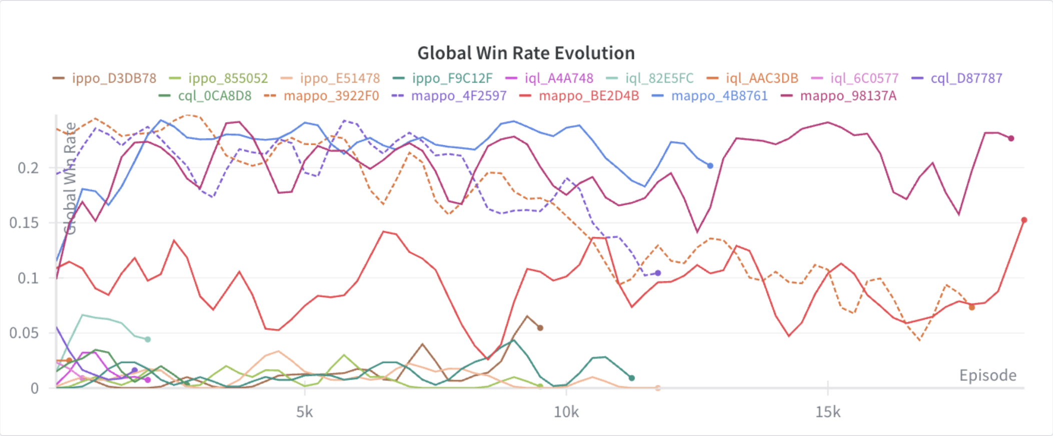
- Decentralized actors (local obs) + Centralized critic (global state)
- Centralized training, decentralized execution
- On-policy with shared episode buffer

Rewards

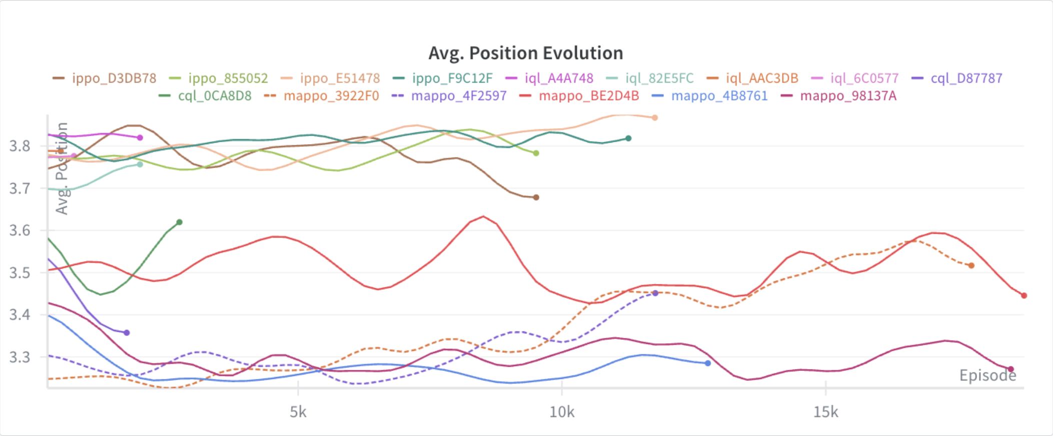
- **MinimalReward**: Sparse (+1 winner, 0 otherwise)
- **ProductionWeightedTerritoryRewardFn**: Territory weighted by production values
- **ShapedRewardFn**: Composite reward combining:
  - Territory (1.0) + Strength (0.05) + Production (0.3)
  - Expansion bonus (0.5) + Asymmetric loss penalty (1.5×)
  - Zero-strength move penalty (0.1)
- **CurriculumShapedRewardFn**: Blends shaped → minimal over time

REWARD ABLATION + VARIANCE ACROSS SEEDS IN MAPPO

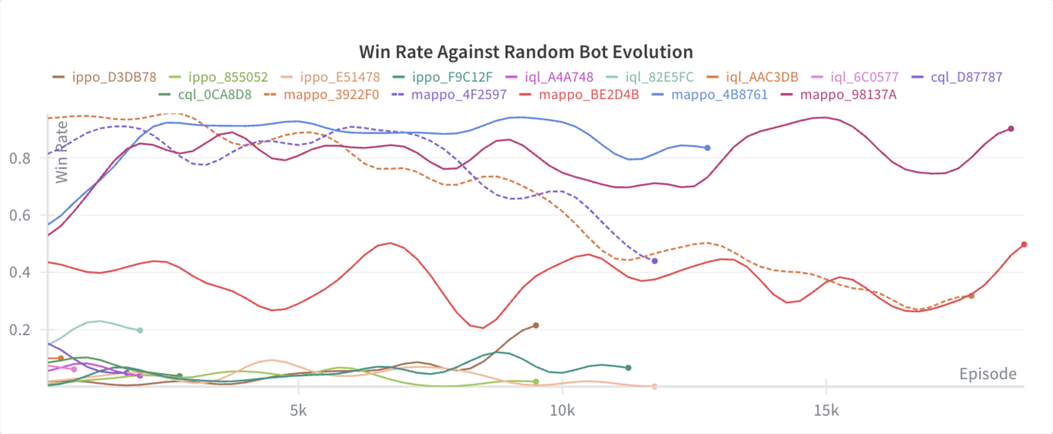
Results



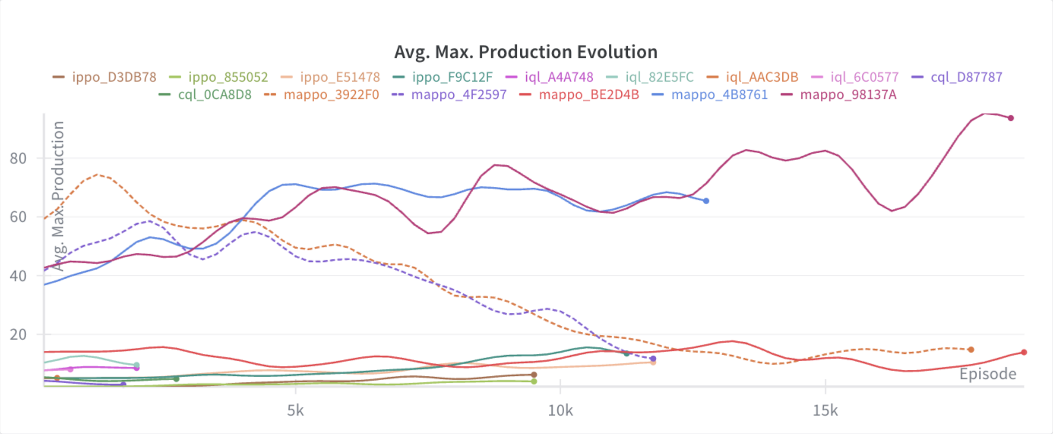
Global Win Rate Evolution



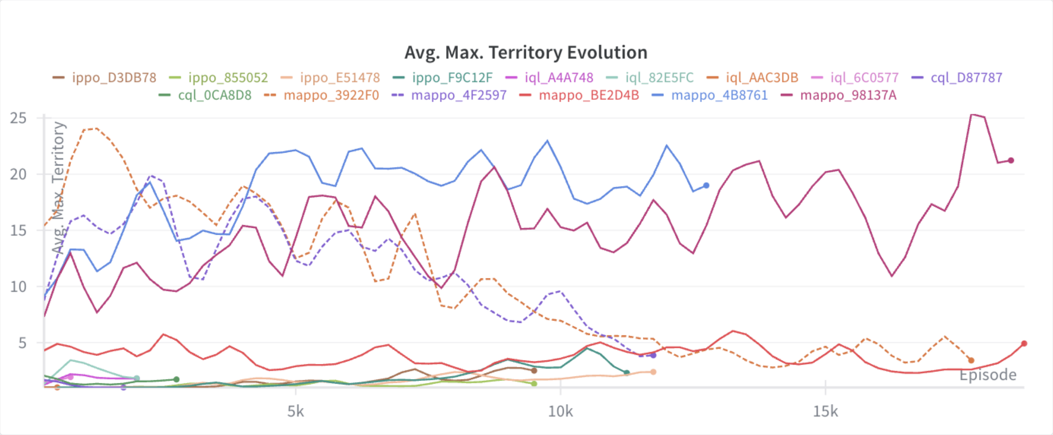
Avg. Position Evolution



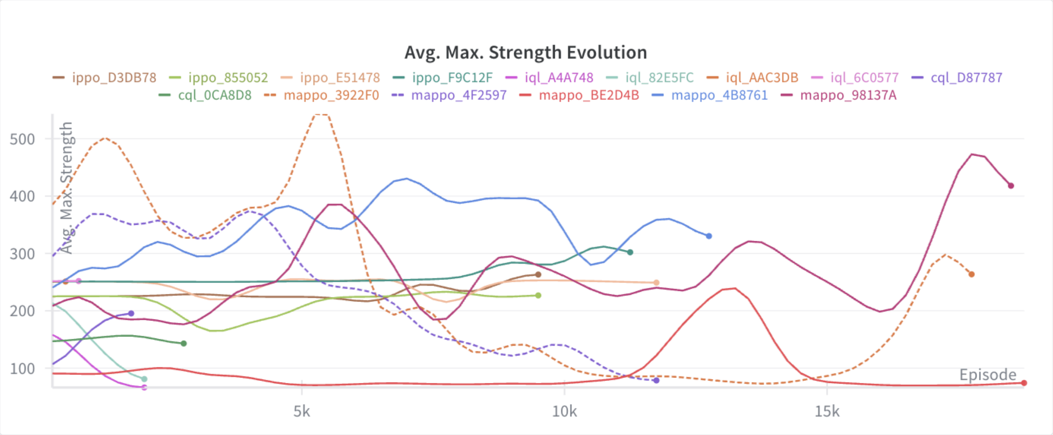
Win Rate Against Random Bot Evolution



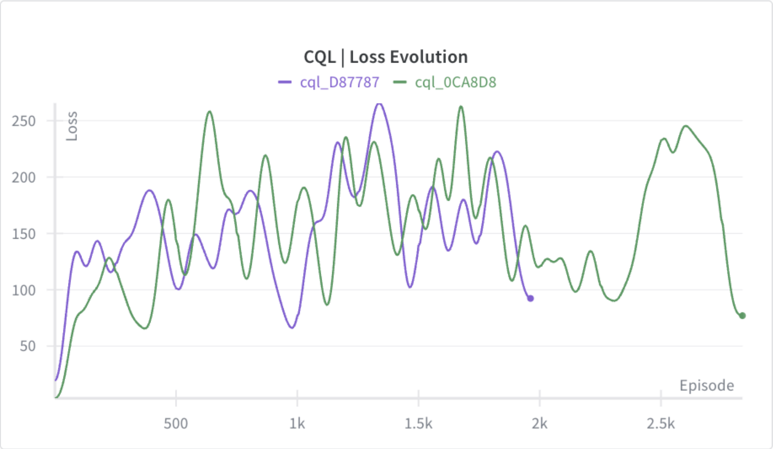
Avg. Max. Production Evolution



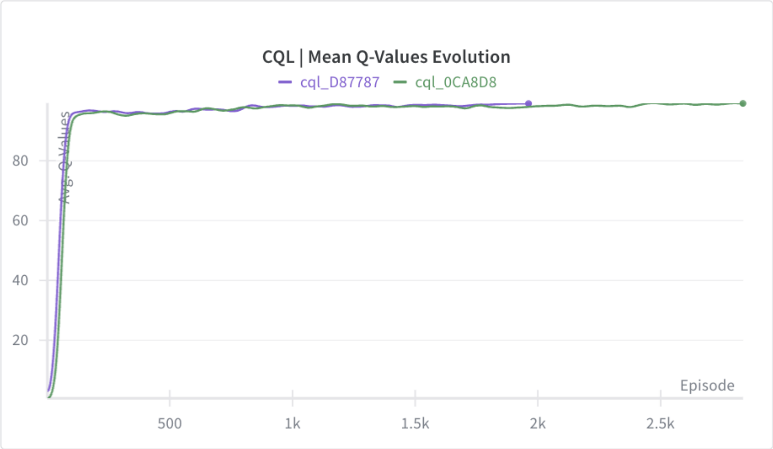
Avg. Max. Territory Evolution



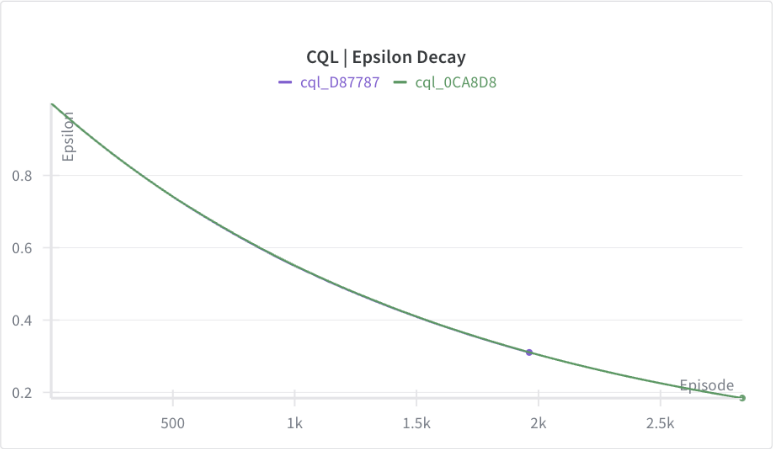
Avg. Max. Strength Evolution



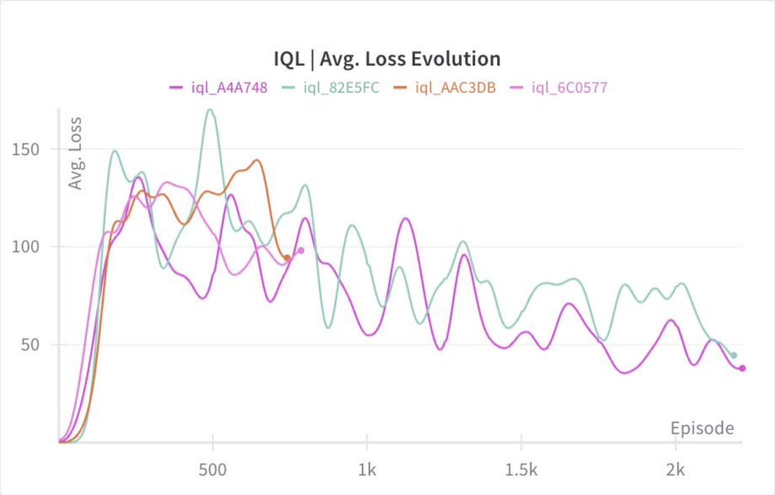
CQL | Loss Evolution



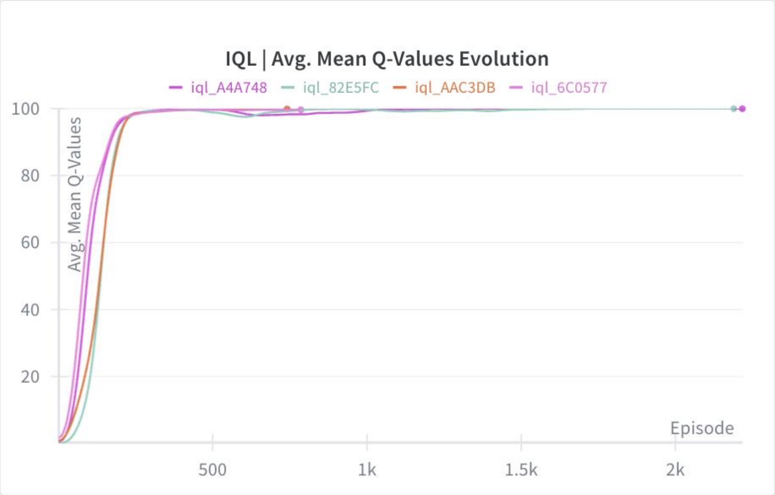
CQL | Mean Q-Values Evolution



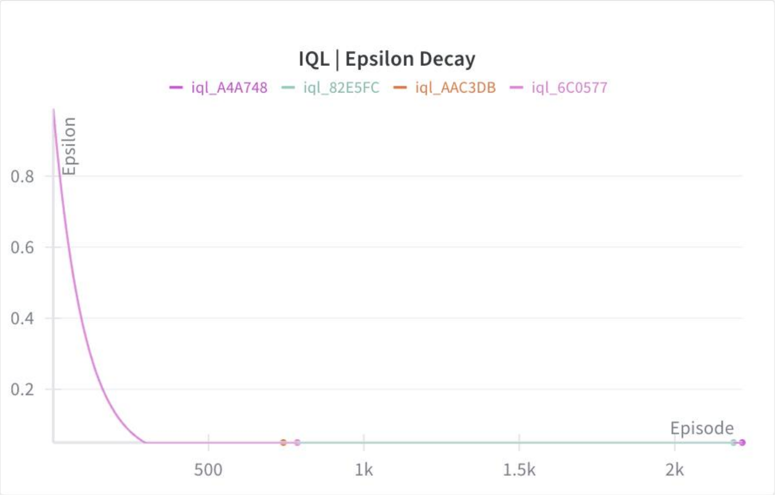
CQL | Epsilon Decay



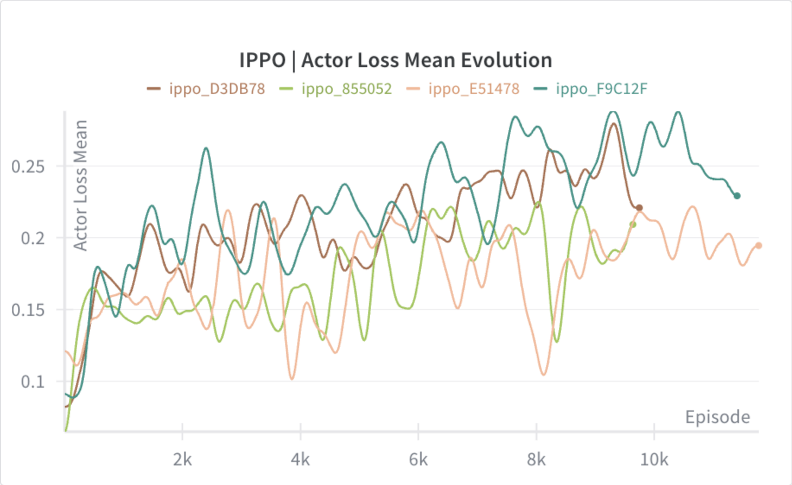
IQL | Avg. Loss Evolution



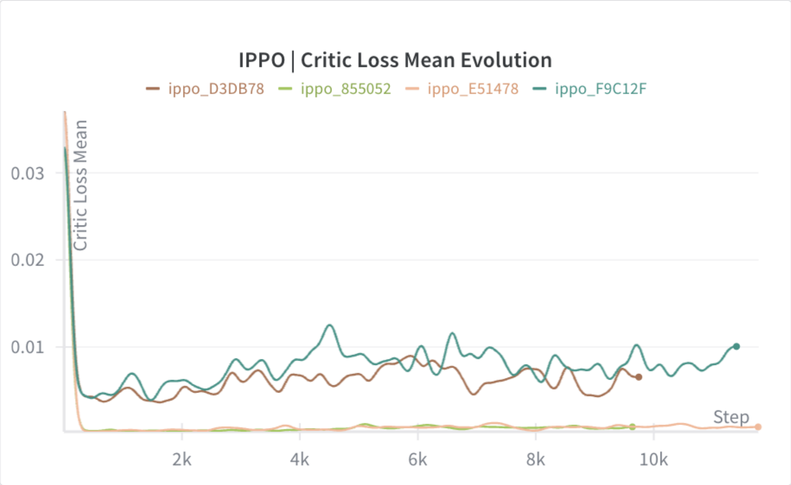
IQL | Avg. Mean Q-Values Evolution



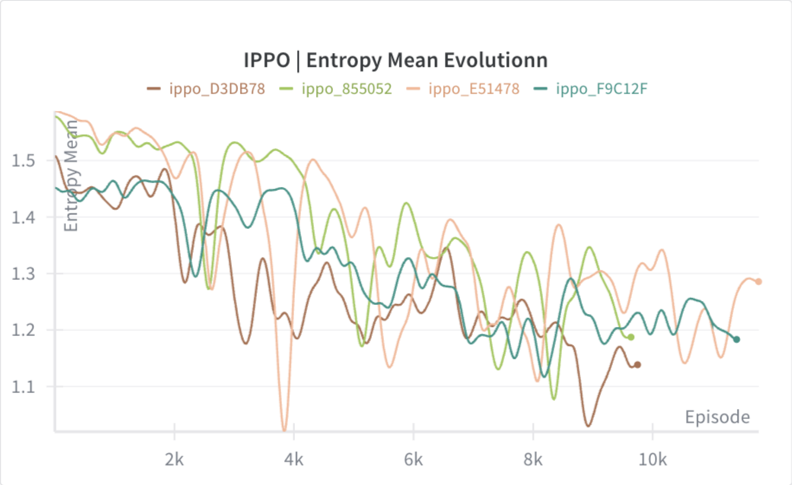
IQL | Epsilon Decay



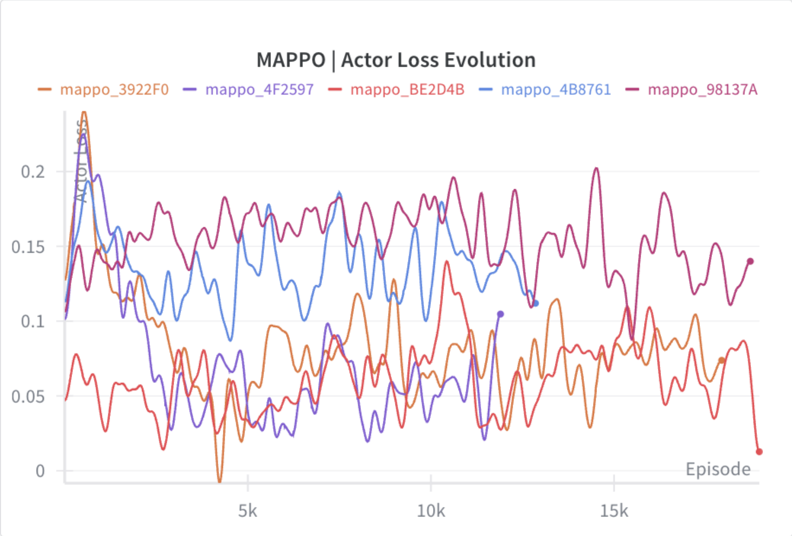
IPPO | Actor Loss Mean Evolution



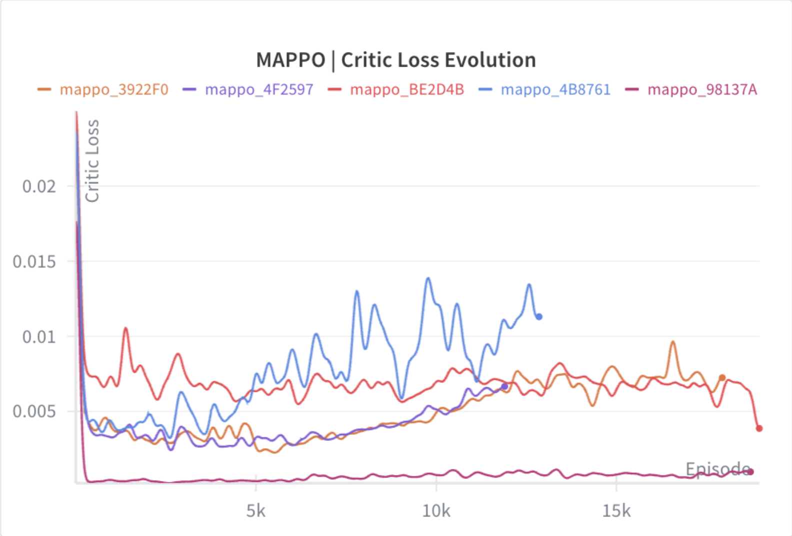
IPPO | Critic Loss Mean Evolution



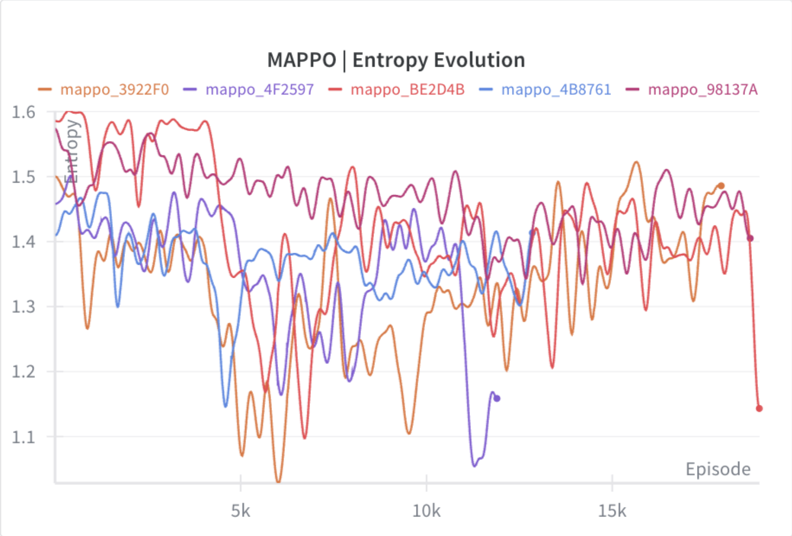
IPPO | Entropy Mean Evolutionn



MAPPO | Actor Loss Evolution



MAPPO | Critic Loss Evolution



MAPPO | Entropy Evolution

1. Algorithm Architecture Matters

- **MAPPO (centralized training + decentralized execution)** outperforms independent approaches
- Centralized critic enables coordinated strategies, maintains scalability

2. Reward Shaping is Critical

- **Dense reward shaping** essential for effective learning
- Minimal rewards fail; curriculum learning degrades performance

3. Policy Gradient > Value-Based Methods

- **MAPPO/IPPO** outperform CQL/IQL
- Value-based methods struggle with stability, sample efficiency, non-stationarity

4. Local Observations Enable Scalability

- **7×7 local patches** reduce complexity while maintaining effectiveness
- More scalable than global observations

5. Performance Gap with Strategic Opponents

- **MAPPO**: ~100% vs RandomBot, 0% vs rule-based bots
- Gap between learned policies and expert-designed strategies

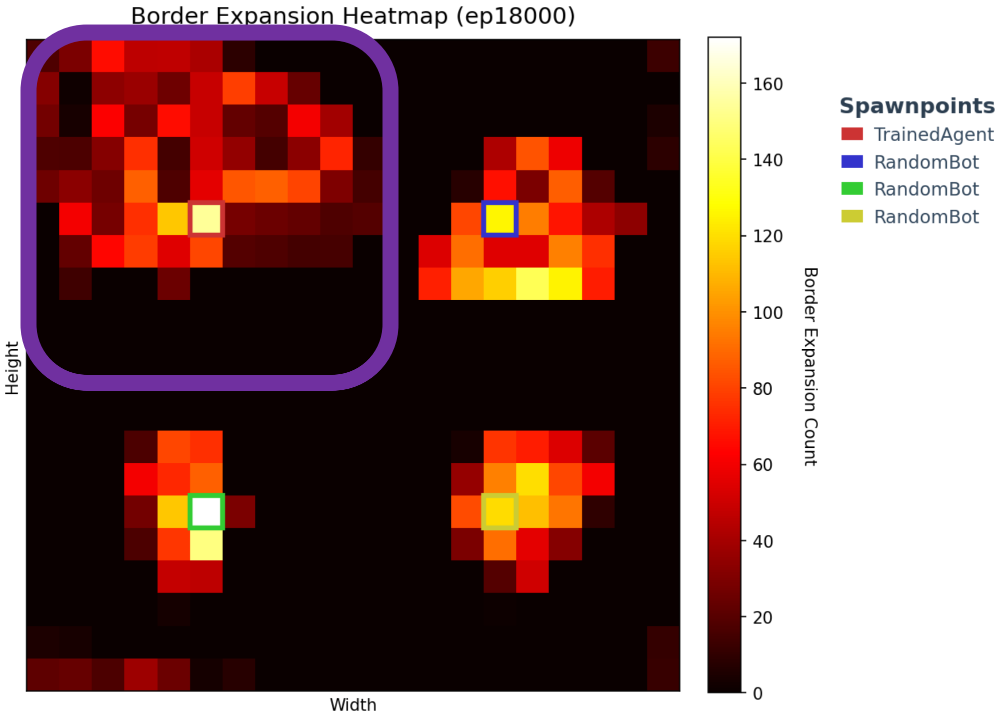
6. Behavioral Observations

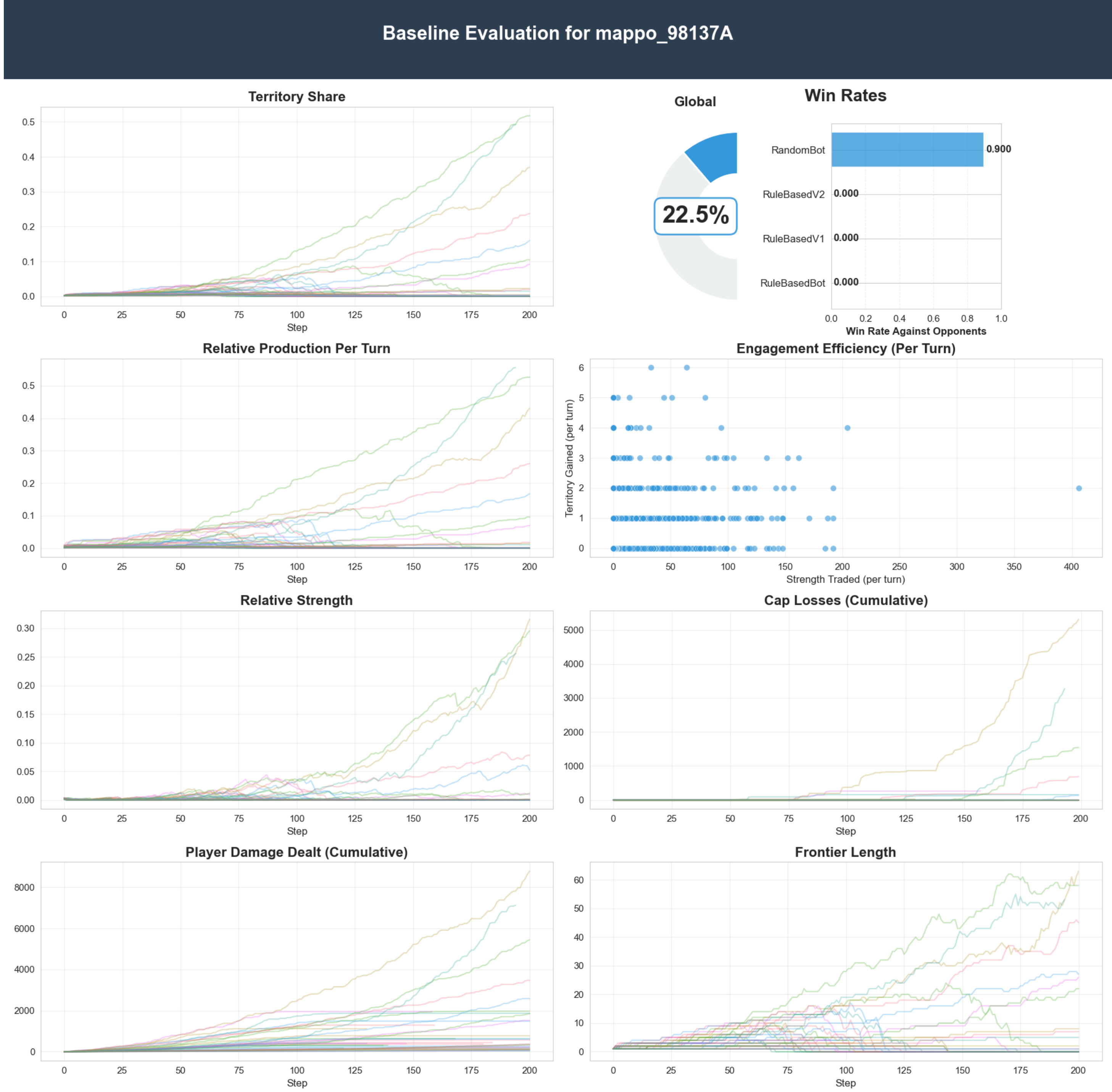
- **Learned**: Territorial expansion
- **Missing**: Attack coordination, efficient movement, inner cell combination

FINAL MAPPO MODEL

Expansion intelligence

Waits to increase strength before expanding





Baseline Evaluation for mappo_98137A
Territory Share
Step
Global
22.5%
Win Rates
RandomBot 0.900
RuleBasedV2 0.000
RuleBasedV1 0.000
RuleBasedBot 0.000
Win Rate Against Opponents
Relative Production Per Turn
Engagement Efficiency (Per Turn)
Territory Gained (per turn)
Strength Traded (per turn)
Relative Strength
Cap Losses (Cumulative)
Player Damage Dealt (Cumulative)
Frontier Length

Failure Cases

Algorithm Limitations

- **CQL/IQL**: Poor sample efficiency, training instability
- **Value-based methods**: Struggle with non-stationarity
- **All algorithms**: Cannot defeat rule-based bots (0% win rate)

Behavioral Failures

- **Inefficient movement**: Many unnecessary zero-strength moves
- **Lack of coordination**: No sophisticated attack strategies
- **Missing mechanisms**: No inner cell combination logic
- **Limited strategy**: Basic expansion only, no advanced tactics

Resource Constraints

- 6-hour training limits (MIT Engaging Cluster)
- Limited episodes (2,500 for CQL, ~18,000 for MAPPO)
- Insufficient convergence time
- Computational bottlenecks (centralized critics)

THANKS!