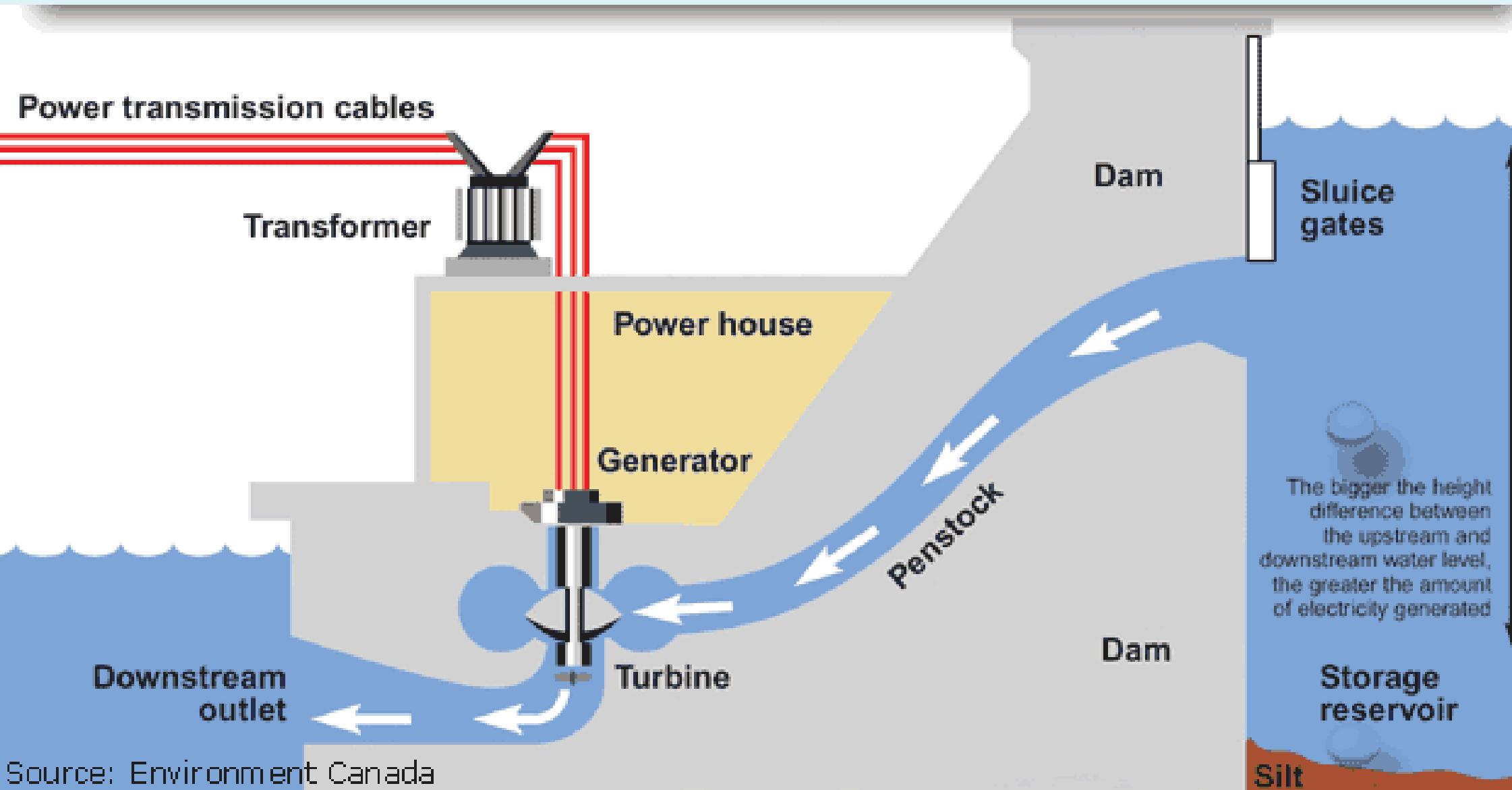


PROJECT REINFORCEMENT LEARNING

V.N. Mehta (Vaishanavi)

PROBLEM RECAP

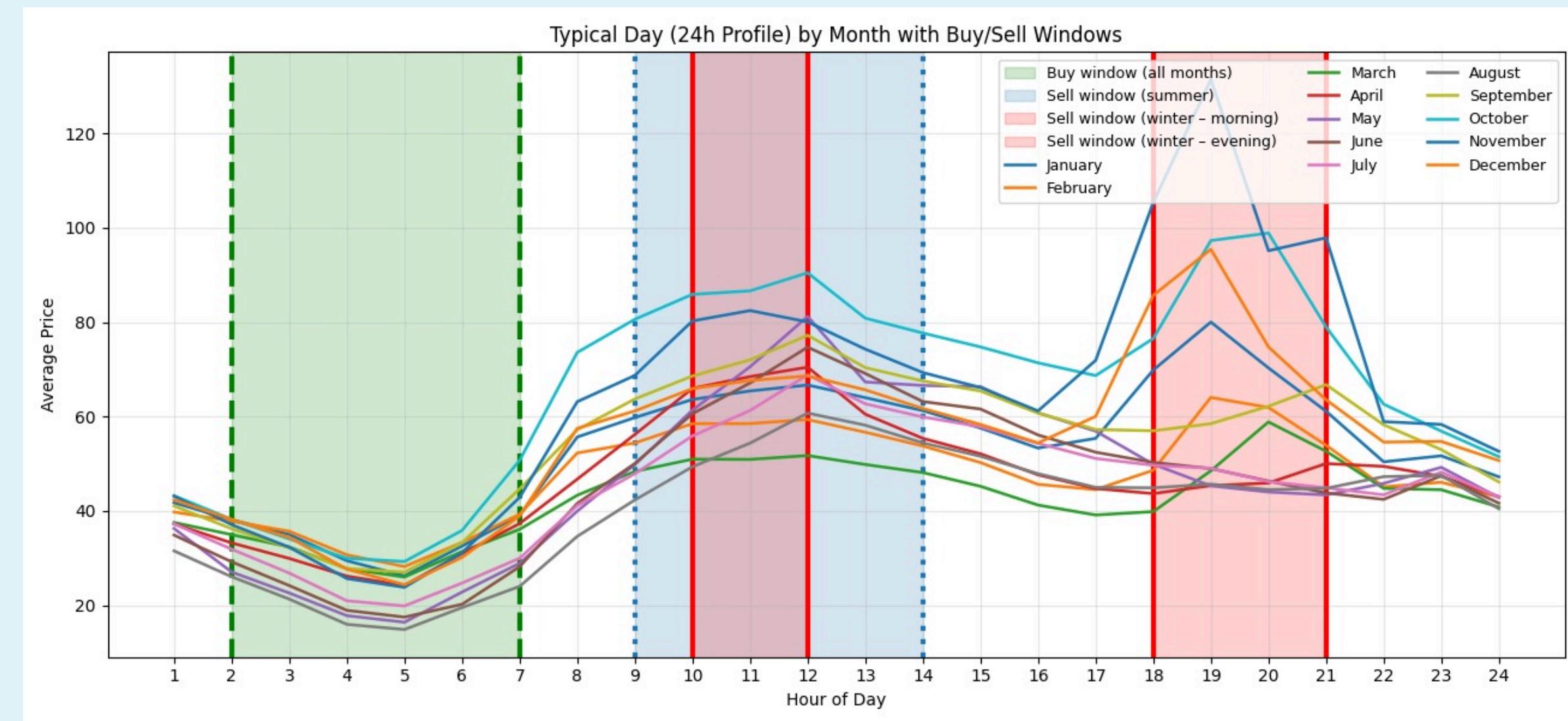
Hourly control of a pumped-hydro dam to maximize cumulative electricity market profit under stochastic prices and physical constraints.



State space (7 dimensions):
volume, price, hour, day_of_week,
day_of_year, month, year

Action space:
Continuous [-1, 1]
Negative = generate (sell),
Positive = pump (buy)

BASELINE HUERISTIC



Winter months (Oct-Feb):

Pump: 2-7am if price < avg_24h

Sell: 10-12h OR 18-21h if price > avg_24h

Summer months (Mar-Sept):

Pump: 2-7am if price < avg_24h

Sell: 9-14h if price > avg_24h

BASELINE RESULT

Total profit: €72,886

Daily average: €100

Average hourly profit: €4.11

Days profitable: 609/730

Random Heuristic Performance over 100 independent runs:

Mean total reward: -€96,355

RL SOLUTION

Algorithm: Tabular Q-learning

- Features
 - Volume (5 bins)
 - Price_below_average (2 bins)
 - Hour (24 bins)
 - is_cold_month(2_bins)
- 480 discrete states ($5 \times 2 \times 24 \times 2$ bins)
- 3 actions: {-1, 0, 1}

SHAPED REWARD FUNCTION

Aligning short term action with long term profit

Intuition

Stored Energy (E_t),
Cumulative Cost (C_t) are tracked

Pumping updates inventory,

- no immediate reward

Generating realizes profit

Profit is delayed until energy is sold

Potential change added

Reward design

Profit on generation:

- Profit = $0.9 \cdot p_t \cdot \text{energy sold} - \text{avg stored cost}$

Potential (future inventory value):

- $\Phi_t = 0.9 \cdot p_t \cdot E_t - C_t$

Training reward:

- $r_t = \text{profit} + (\gamma \Phi_{t+1} - \Phi_t) / 1000 \quad (1)$

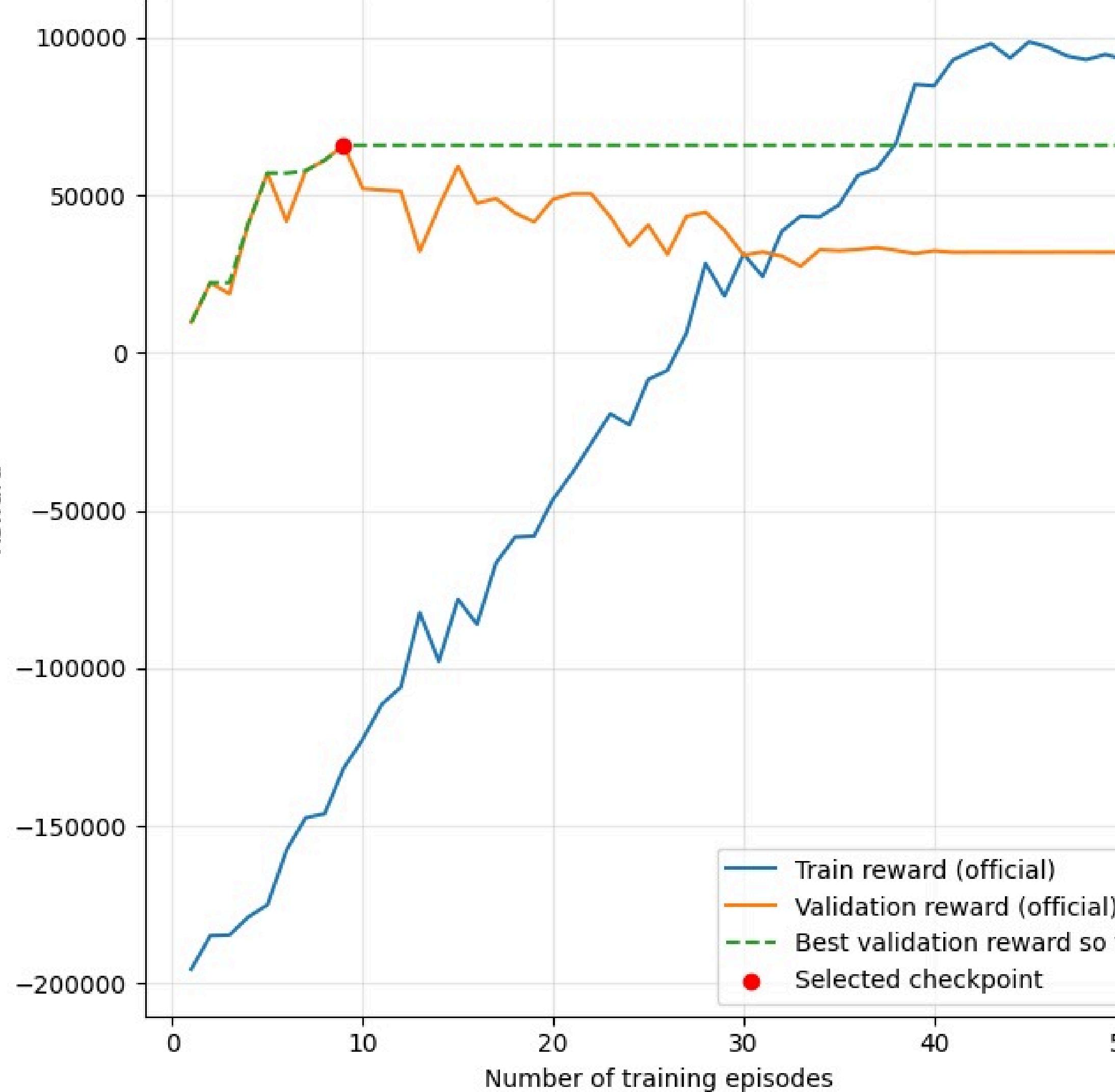
(1) Ng, Harada & Russell, Policy Invariance under Reward Transformations, ICML 1999

RL TRAINING

- Trained on 3 years of data, 50 episodes
- Best policy after ~8 episodes (checkpointed)
- γ : 0.95 → 0.99 (better long-term stability)
- ϵ : linearly decayed from 1.0 over 80% of training
- α : decayed 0.03 → 0.003 to stabilize final policy

VALIDATION

The validation curve reaches its maximum around **episode 8** (highlighted by the red dot), after which performance degrades. This indicates that additional training episodes do not improve generalization.



RL RESULTS

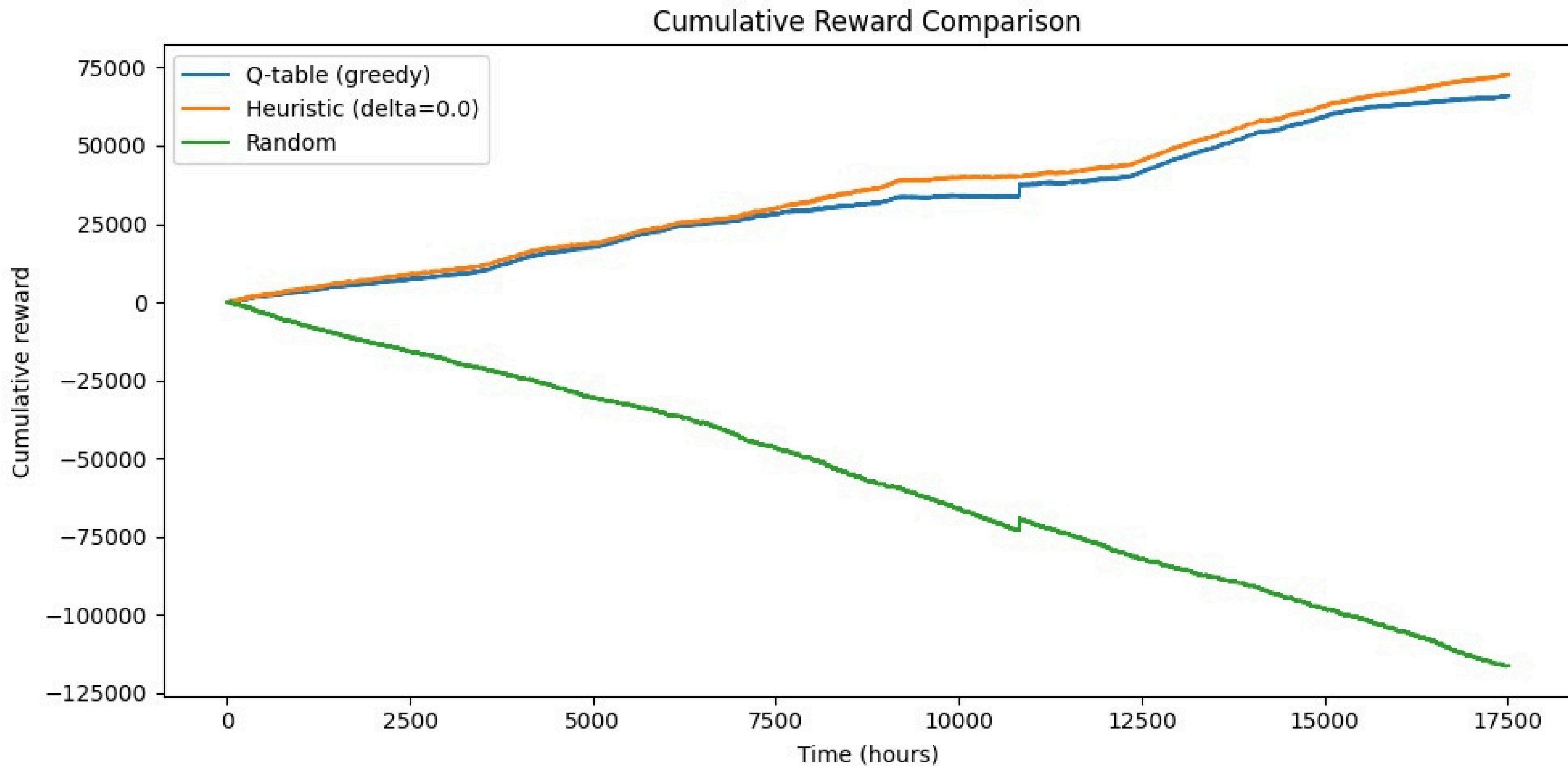
Long-term Performance (730 days validation set)

Total profit: €65,902

Daily average: €89

Average hourly profit: €3.7

PERFORMANCE COMPARISON



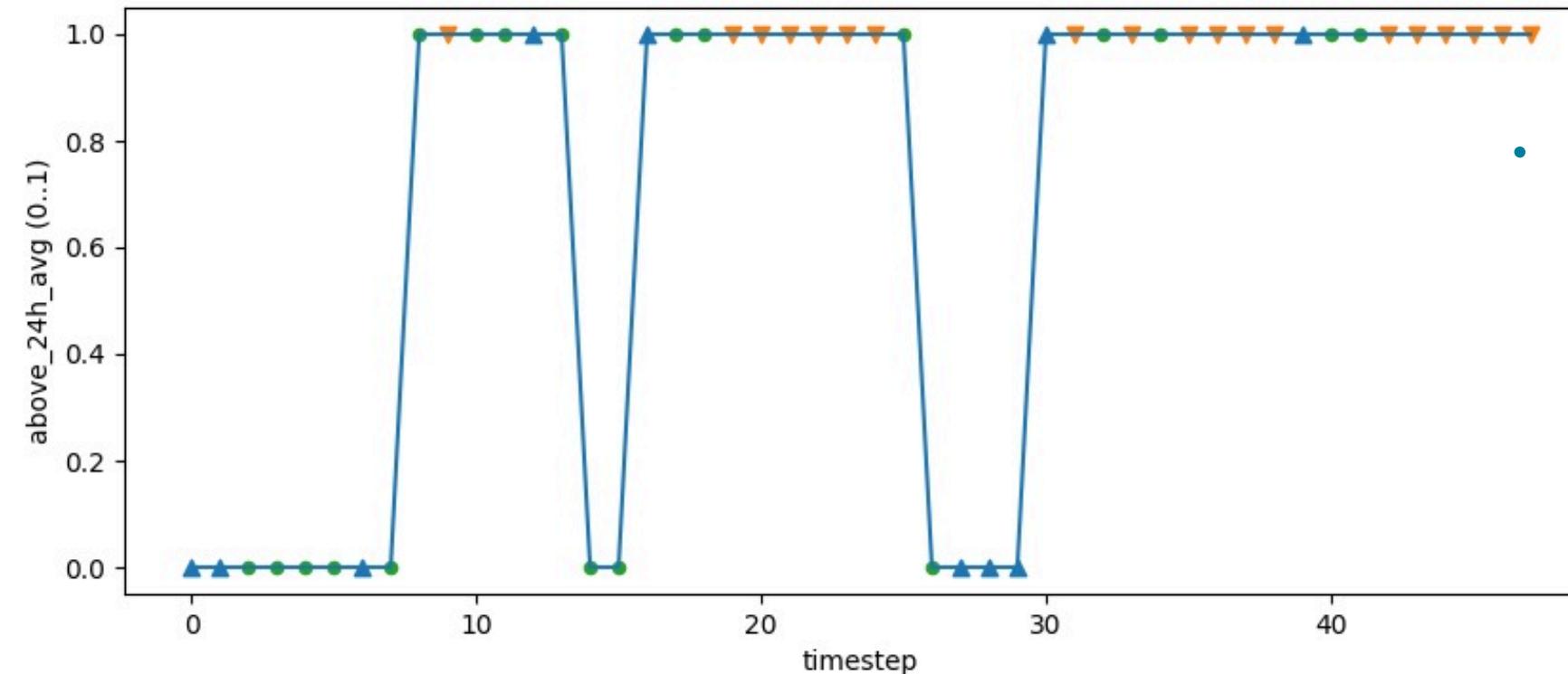
PERFORMANCE COMPARISON

Method	Random	Baseline	RL
Total profit	-€96,355	€72,886	€65,902
Daily avg	-€131	€100	€89
Utilization	100%	100%	100%

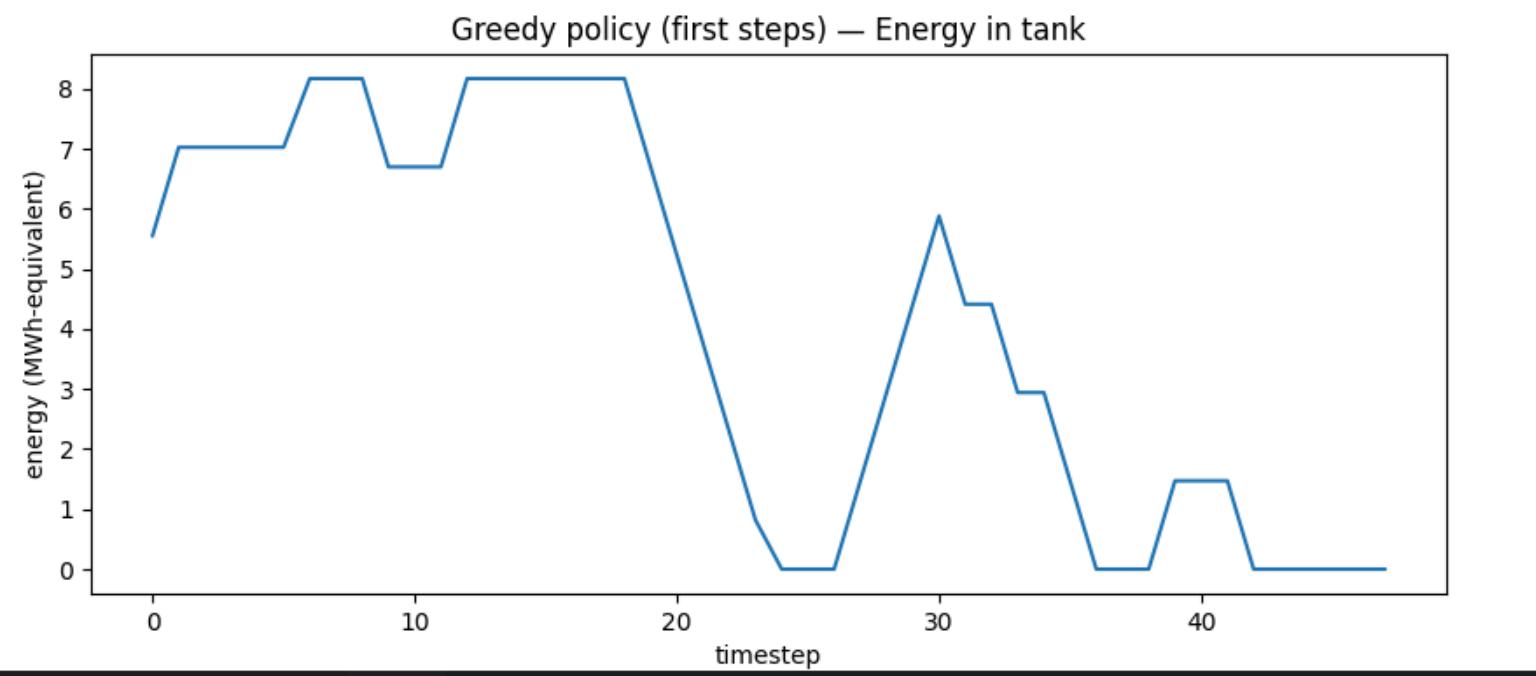
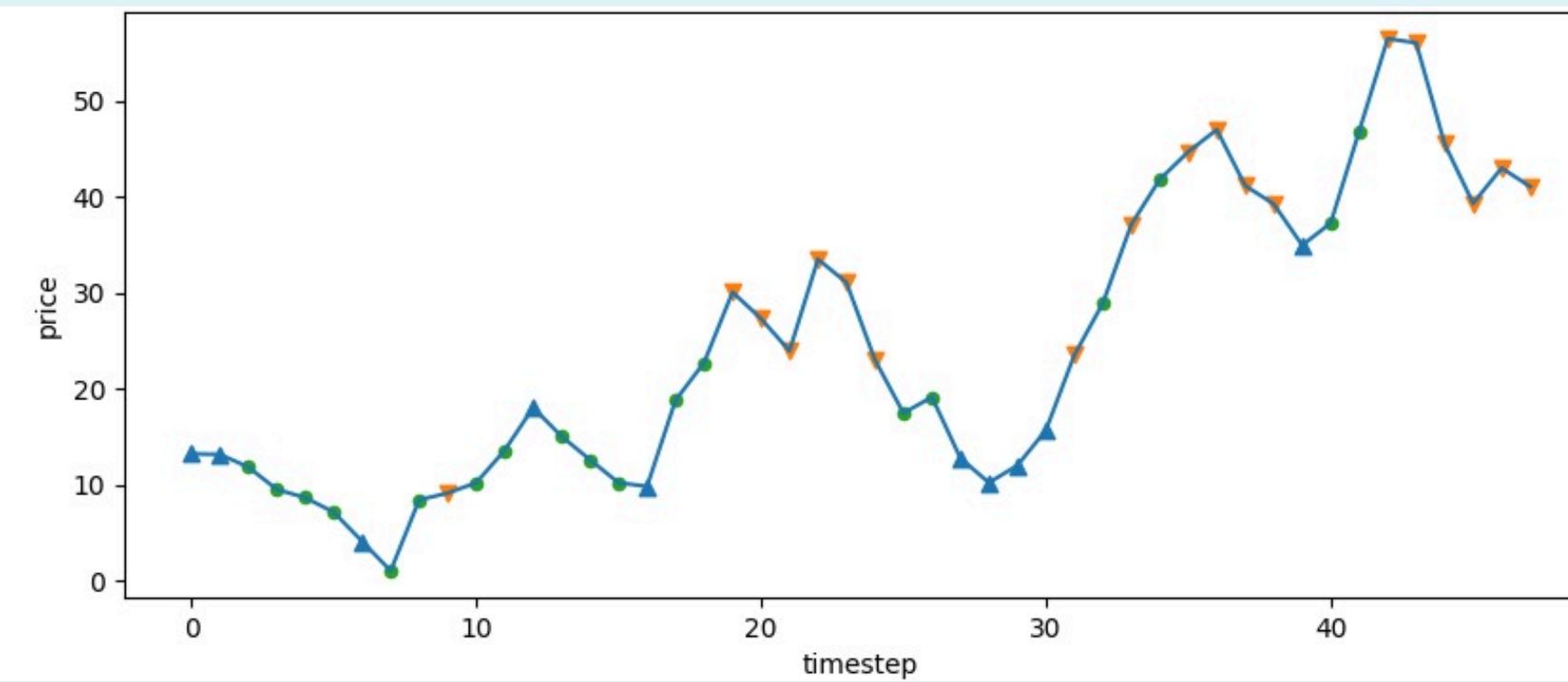
Q-Learning substantially outperforms the random baseline and is a bit lower than the rule-based heuristic (-9.5%). The result confirms that the learned policy captures exploitable price patterns.

BEHAVIOR ON VALIDATION

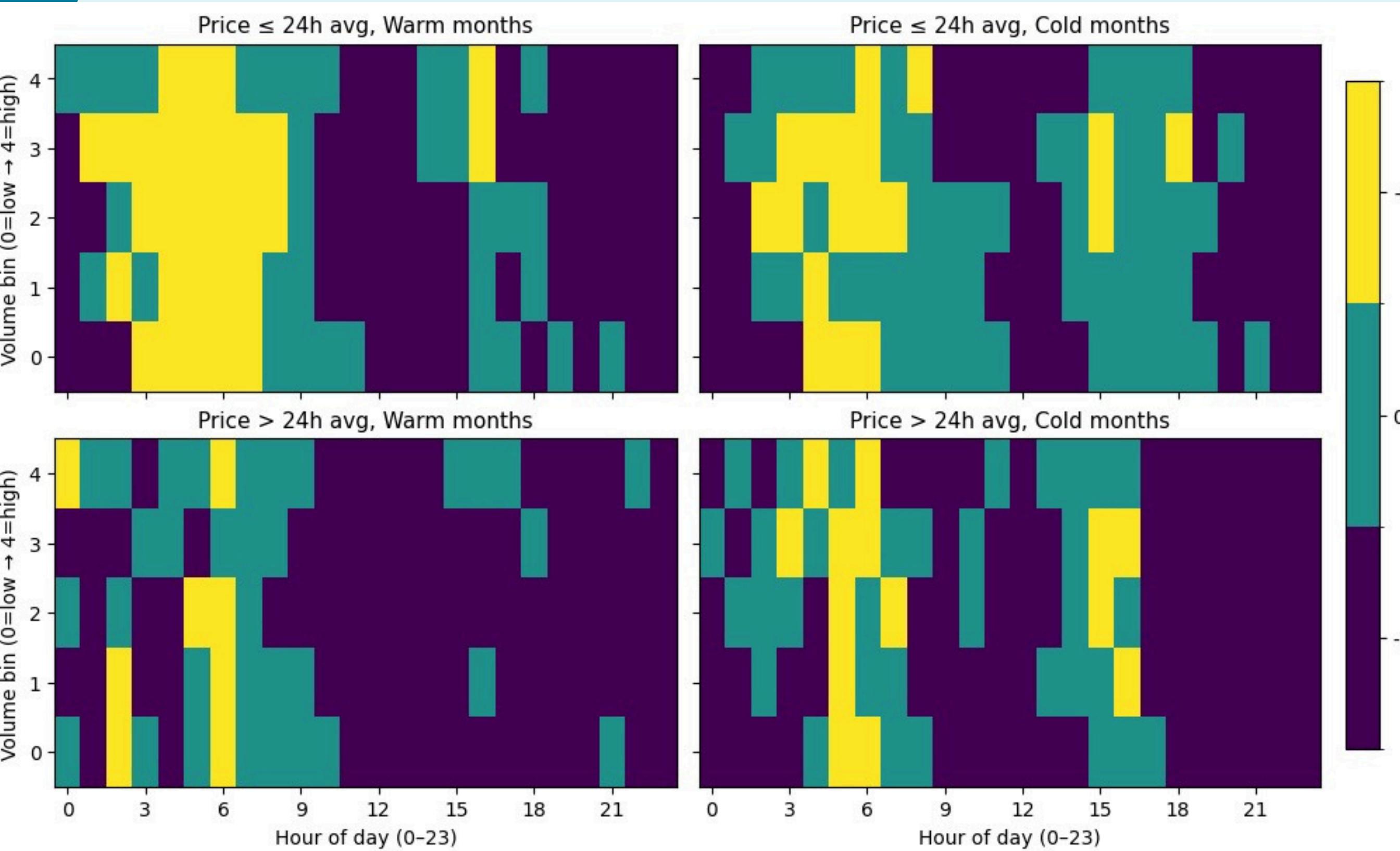
Difficult day!



▲ Pump ▼ Generate ● Hold

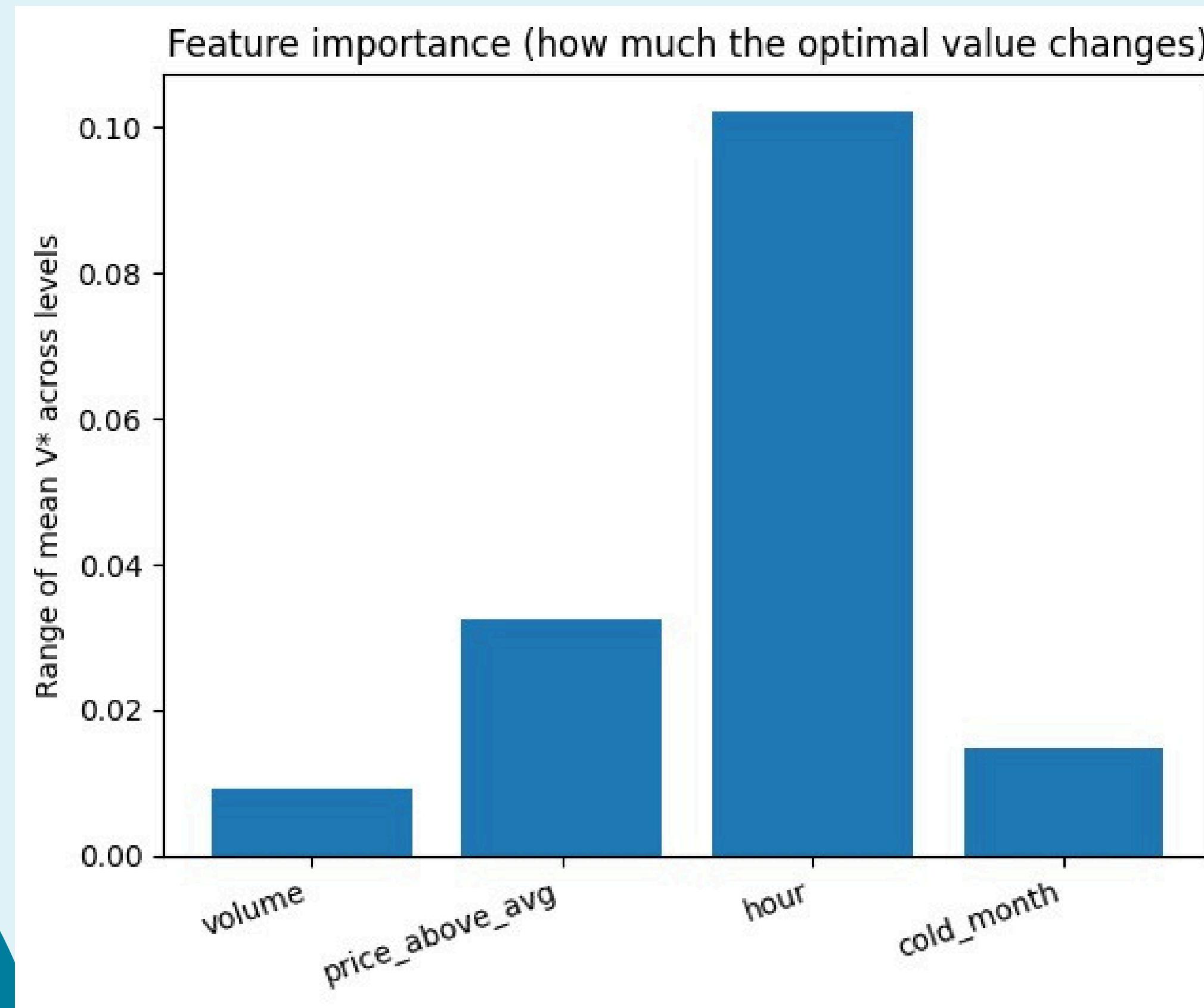


WHAT DID THE AGENT LEARN?



Performance improved most from including **hour of day** and **price relative to 24h avg**, capturing intraday arbitrage structure.

WHAT DID THE AGENT LEARN?



The cold-month indicator gave minor gains.

Hour feature most important

Volume gives the least change in value

ABLATION STUDIES

Feature experiments:

Relative price: Critical (+40% validation)

Volume: Required (constraint handling)

Hour/day of the week bins

Finer seasons: Minimal gain

Weekday/weekend: No improvement

Hyperparameters:

$\alpha=0.03$: Stability vs speed

$\gamma=0.99$: Long-term planning essential

ϵ decay 80%: Best exploration-exploitation

CONCLUSION

Main Results

RL works, Intuitive behavior, Good validation performance

Comparison with Heuristics

Complexity, Predictability, Rule-based still strong

Future Work

Features, Realism, Scaling, DDQ, Deep RL

**THANK
YOU**