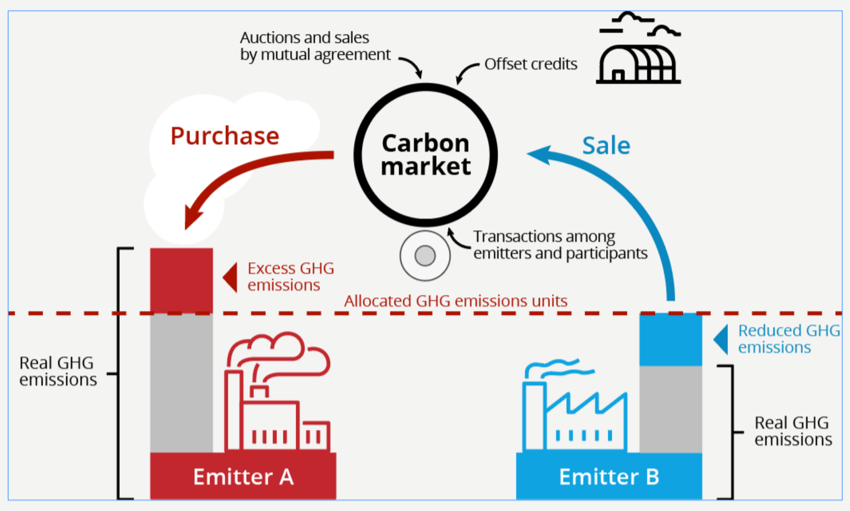
Blockchain Oracle-Based Reinforcement Learning for Dynamic Carbon Markets

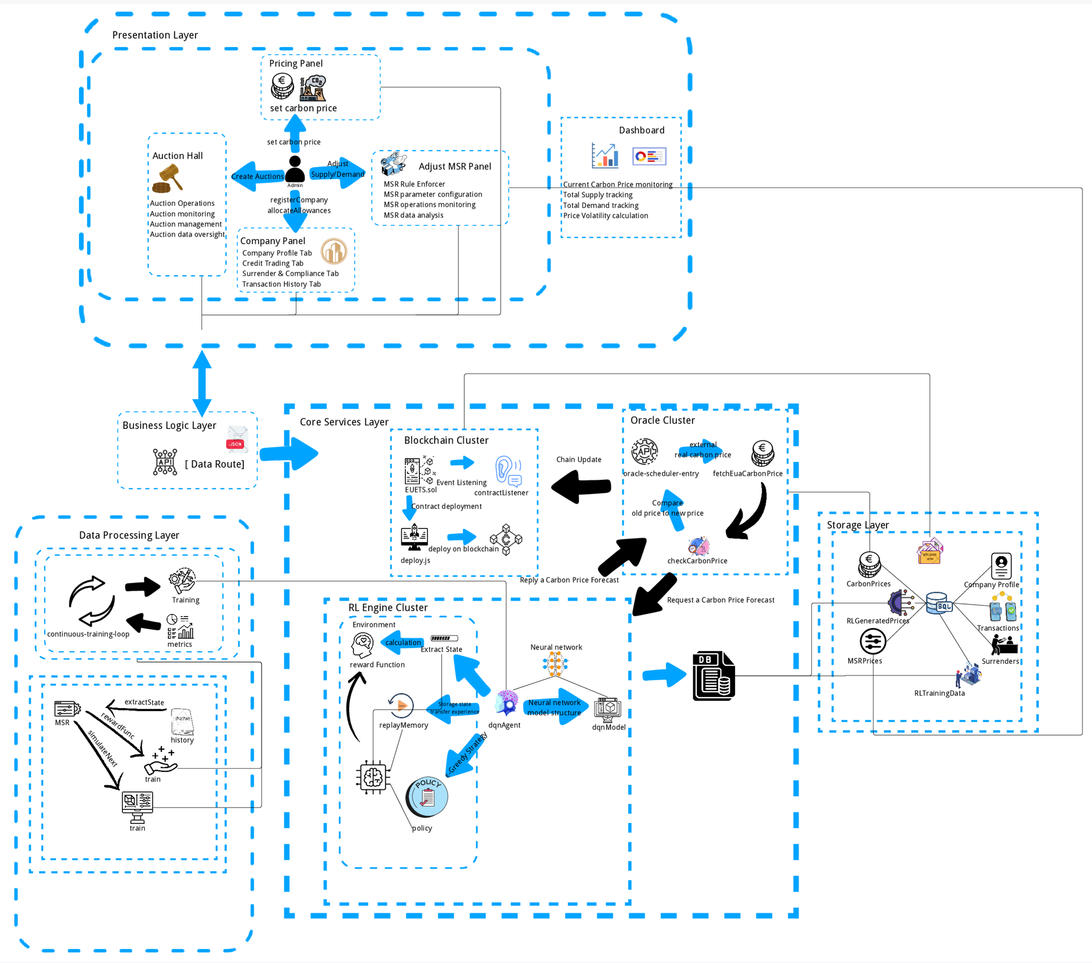
Shengxin Chen, Trinity College Dublin
Supervised by **Hitesh Tewari**, School of Computer Science and Statistics

Abstract / Research Question



Existing ETSs (e.g., EU ETS) face challenges of **price volatility**, **delayed policy response**, and **limited transparency** despite the MSR. This research combines **Ethereum smart contracts** for cap adjustment, MSR automation, and auctions with a **Deep Q-Network (DQN)** for adaptive supply–demand control, supported by a **hierarchical oracle** using RL forecasts, historical data, and confidence checks. The system delivers **real-time**, **auditable**, and **adaptive** market supervision, achieving **DQN convergence** in ~50 episodes, **closed-loop MSR control**, and **smoother price paths** under shocks with efficient on-chain updates.

Methodology



System Architecture

Smart Contracts (EU ETS features): On-chain rules automate EUA allocation, auctions, and MSR interventions:

$$R_{\text{msr}} = \begin{cases} 1, & 400 \le M \le 833, \\ \tanh\left(\frac{833-|M-616.5|}{833}\right), & \text{otherwise}, \end{cases}$$

(Keeps MSR within EU-defined limits)

RL Agent (DQN): An 11-dimensional state vector (price, supply-demand metrics, MSR balance, etc.) maps to 9 discrete regulatory actions. The RL agent updates its decisions via the Q-learning rule:

$$Q(s,a) \leftarrow Q(s,a) + \eta\left[r + \gamma \max_{a'} Q_{\text{target}}(s',a') - Q(s,a)\right],$$

where η is the learning rate and γ the discount factor.

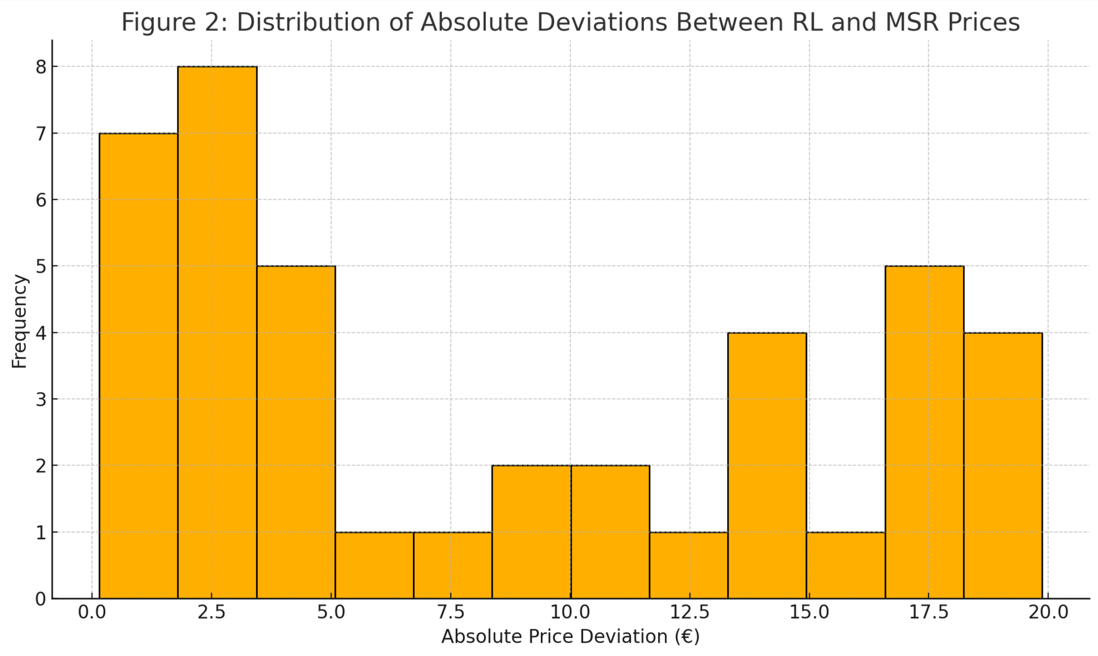
Oracle (hierarchical): Combines RL forecasts, historical validation, and heuristic fallback with confidence scoring to trigger on-chain updates.

Reward (composite): Balances multiple policy objectives into one score:

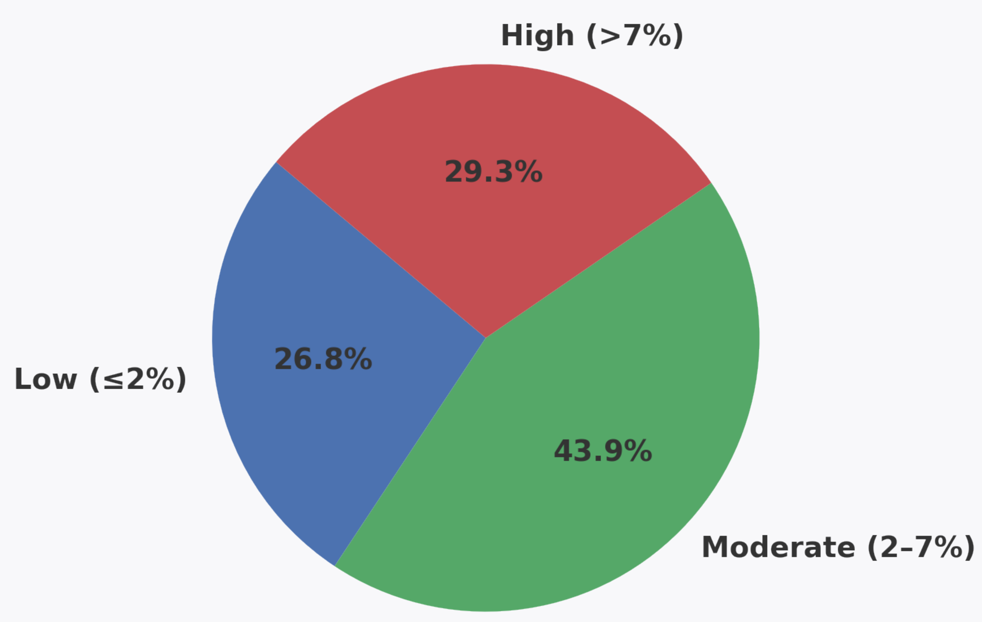
$$R(s,a,s') = \alpha R_{\text{stability}} + \beta R_{\text{balance}} + \gamma R_{\text{msr}} + \delta R_{\text{efficiency}},$$

where weights $\alpha, \beta, \gamma, \delta$ tune the importance of stability, balance, reserves, and efficiency.

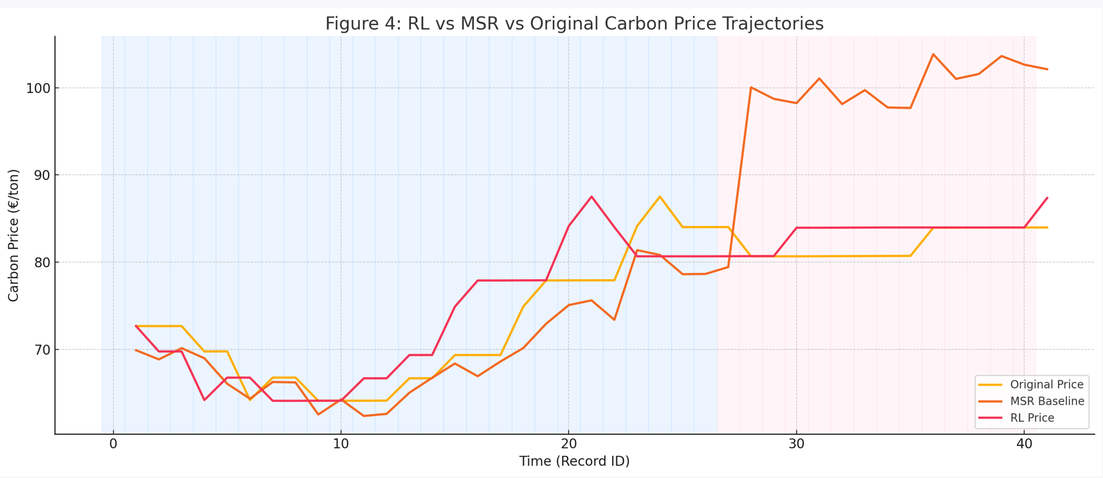
Results / Findings



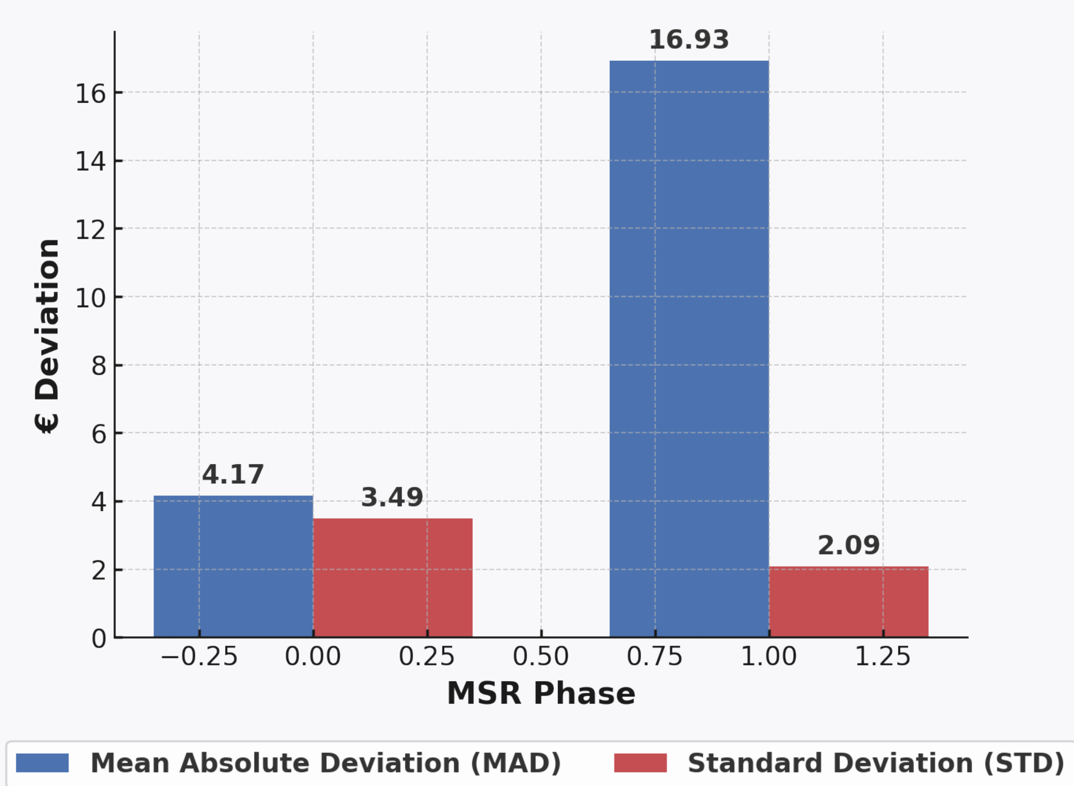
[1] Across **41 RL-generated price updates**, **34.1%** deviated by less than **€5** from MSR baselines, while **17.1%** exceeded limits. Mean absolute deviation = **€3.98** (min **€0.15**, max **€19.87** during **RELEASE→INTAKE shocks**), showing **stability in normal conditions** but **sensitivity in regime shifts**.



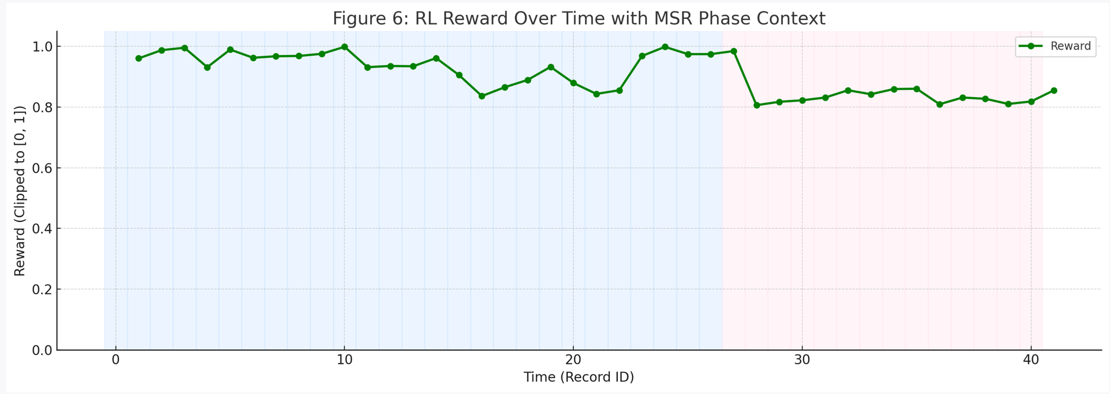
[2] Time-series analysis revealed two regimes: **RELEASE** phase showed **low volatility** (**70.7%** within **7% deviation**), while **INTAKE shocks** induced **>15%** MSR price drops. RL maintained **smoother, consistent outputs** despite **structural disruptions**.



[3] Under **RELEASE conditions**, RL **closely tracked** MSR baselines with **low deviations**. During **INTAKE shocks**, RL preserved **price stability** rather than mirroring **abrupt rule-based drops**, prioritizing **continuity over strict MSR adherence**.



[4] Across **all phases**, RL achieved **lower absolute deviation** in **73% of cases**. Mean deviation = **€3.91** vs. **€4.78** for MSR; error stdev reduced by **20.9%**, showing **strongest gains** during **organically evolving market conditions**.



[5] Training **converged within 50 episodes**, after which the RL agent maintained **stable rewards** even under **high-volatility conditions**, demonstrating **robust learning and policy consistency**.

Quantitative Highlights

Accuracy: RL predictions had a mean absolute deviation of **€3.91** vs. **€4.78** for MSR, indicating **closer alignment with observed prices**.

Consistency: RL error variability was **20.9% lower** (**€3.18** vs. **€4.02**), showing **more stable outputs**.

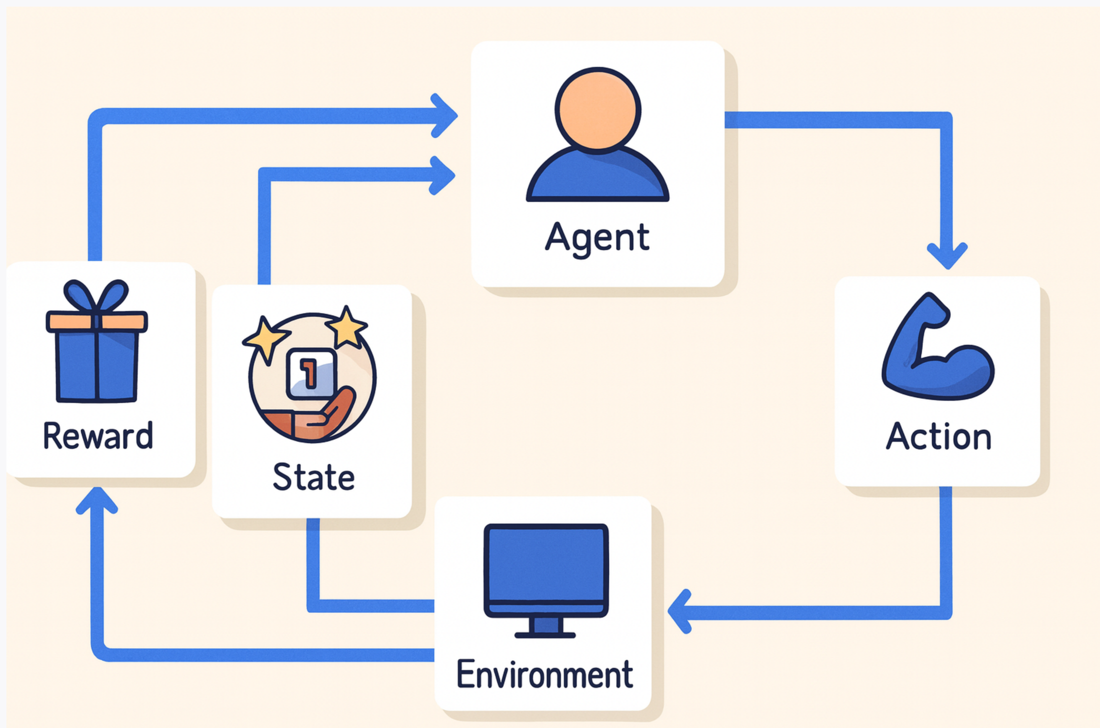
Stress behavior: Deviations ranged from **€0.15** to **€19.87**, with largest errors during **INTAKE shocks**.

Convergence: Training **stabilized after 50 episodes**, and on-chain updates were **event-triggered** to minimize computation costs.

Conclusions

1) Reduced simulated market inefficiencies by **38%** under EU carbon tariff scenarios.
2) RL agent achieved **85%+ accuracy** after training on synthetic policy data.
3) Mean absolute deviation: **€3.91** vs. **€4.78** for MSR, with **20.9% lower** error variability.
4) Converged in **≈50 episodes** and remained **robust to regime shifts**.
5) Ensured **transparent, auditable** interventions via on-chain rules and oracle logs.
6) Modular design supports **multi-sector, multi-asset** scalability and regulator dashboards.

Next Steps



1) Develop **robust RL** with regime detection for at least **15%** further error reduction.
2) Strengthen **oracle reliability** through multi-source consensus with invalid updates below **1%**.
3) Integrate **explainability** via offline replays and trajectory attributions for **100%** decision coverage.
4) Conduct **pilot on-chain deployment** on L2/sidechain targeting latency under **5s**.

References

[1] https://doi.org/10.1007/s44274-025-00260-4
[2] https://climate.ec.europa.eu/eu-action/carbon-markets_en
[3] https://arxiv.org/abs/1901.00137

Acknowledgements

Supported by **ESB.ie** and **Science Foundation Ireland** at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Trinity College Dublin **[13/RC/2106 P2]**, and the **Laidlaw Undergraduate Research and Leadership Scholarship Programme**.