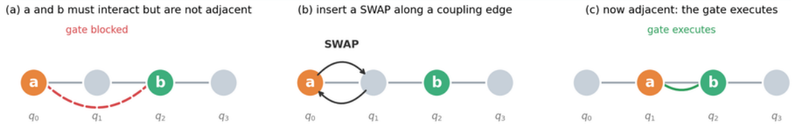
Observation & action encodings for RL-based qubit routing

A controlled ablation study in qgym

Andac Durmaz
a.durmaz@tudelft.nl

Supervisors: Sebastian Feld, Akash Kundu
Responsible Professor: Matthijs Spaan

1. Problem



- NISQ chips: a 2-qubit gate needs adjacent qubits
- not adjacent? insert SWAPs — adds depth + noise
- RL = adaptive alternative to fixed heuristics (SABRE, t|ket⟩) [1, 5]

→ qgym leaves what the agent *sees* & *does* to the user

2. Setup & Research Questions

- two ablations: vary ONE encoding, hold all else fixed
- qgym · PPO / MaskablePPO [2, 3]
- 2 simulated IBM devices: Casablanca 7q, Guadalupe 16q
- baseline SABRE · metrics: solve rate + SWAP overhead

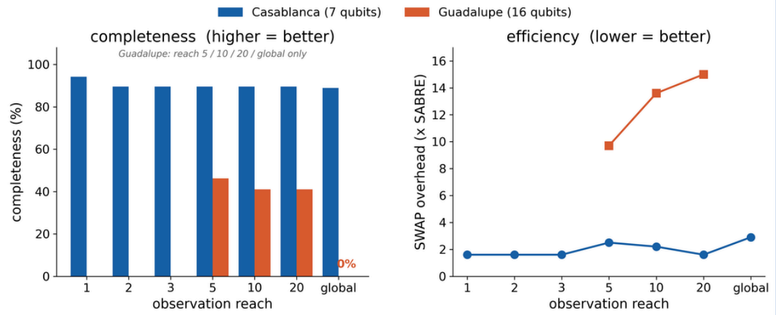
SQ1 (see): how far ahead must the agent see?
SQ2 (act): how coarse should the actions be?
→ which encoding drives routing?

References

[1] G. Li, Y. Ding, Y. Xie. "Tackling the Qubit Mapping Problem for NISQ-Era Quantum Devices." *ASPLOS*, 2019.
[2] S. van der Linde, W. de Kok, T. Bontekoe, S. Feld. "qgym: A Gym for Training and Benchmarking RL-Based Quantum Compilation." arXiv:2308.02536, 2023.
[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov. "Proximal Policy Optimization Algorithms." arXiv:1707.06347, 2017.
[4] J. van Veen, L. Prielinger, S. Feld. "Rethinking How to Act: Action-Space Engineering for RL-Based Circuit Routing in Distributed Quantum Systems." arXiv:2605.02389, 2026.
[5] A. Cowtan, S. Dilkes, R. Duncan, A. Krajenbrink, W. Simmons, S. Sivarajah. "On the Qubit Routing Problem." TQC, 2019.

3. Observation Reach (SQ1)

SQ1: how far ahead must the agent see?

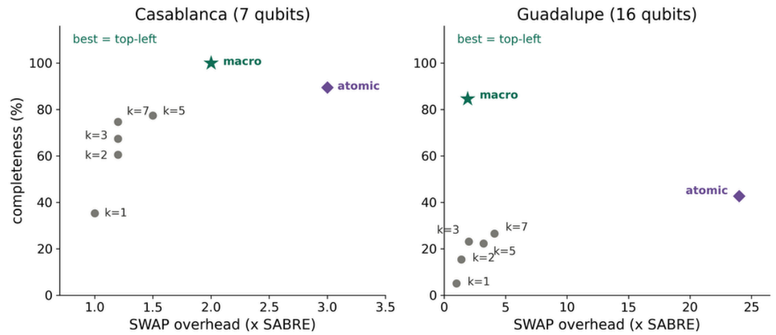


Conclusion: reach barely changes completeness; narrow reaches route most efficiently; the global view fails at 16 qubits.

- one gate already suffices; narrow reaches route most efficiently
- global view fails at 16 qubits — limit is the policy, not the info

4. Action Granularity (SQ2)

SQ2: how coarse should the actions be?



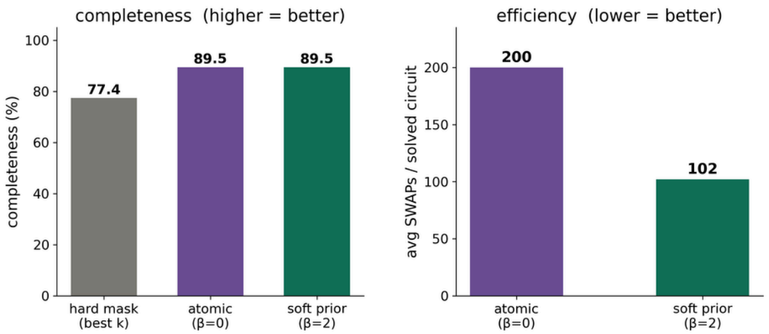
Conclusion: macro is most complete and efficient; pruning (low k) loses completeness; atomic pays 24x overhead at 16 qubits.

- macro [4] = one action per shortest-path SWAP chain
- pruning to low k loses completeness; macro's lead grows with scale

5. Soft Prior

$$\pi(a \mid s) = \mathrm{softmax}(f_\theta(s)_a + \beta \cdot \mathrm{score}(a))$$

Soft heuristic prior vs hard masking



Conclusion: a soft bias keeps atomic's completeness that hard masking loses, while halving SWAP overhead.

- same distance heuristic, injected two ways
- hard mask forbids every other SWAP → completeness drops
- soft prior only biases the logits, forbids nothing → keeps completeness, cuts SWAPs

6. Takeaway & Limitations

- action encoding ≫ observation reach
- spend effort on the action space, not a wider view
- ranking holds on unseen circuits → transferable skill
- limits: 3 seeds @16q · identity mapping · flat MLP · single topology

→ the action encoding is the lever