

Learning scalable entanglement distribution policies

Christos Chalkias,^{1,2} Jan Krzywdka,^{1,2} Patrick Emonts,^{1,3,4,5} and Evert van Nieuwenburg^{1,2,3}

¹ *$\langle aQa^L \rangle$ Applied Quantum Algorithms, Universiteit Leiden*

²*Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands*

³*Instituut-Lorentz, Universiteit Leiden, Niels Bohrweg 2, 2333 CA Leiden, Netherlands*

⁴*Institute for Complex Quantum Systems, Ulm University, 89069 Ulm, Germany*

⁵*Center for Integrated Quantum Science and Technology (IQST), Ulm-Stuttgart, Germany*

I. INTRODUCTION

A major goal in the field of quantum communication is the establishment of a (global) quantum internet via which long range entanglement can be distributed and used for many applications, ranging from quantum key distribution [1], distributed computing [2] to clock synchronization [3] and long-baseline telescropy [4], among others [5, 6]. Recent advances have made it possible to reliably entangle two distant parties [7–9], and a next step is entanglement distribution in a network. One of the core ingredients of this distribution infrastructure is its *policy*, describing the operation of the network at a high-level with the goal of achieving long range entanglement between distant nodes. It can be described as a *Markov Decision Process* (MDP) over a stochastic environment, and for that reason several groups are investigating *Reinforcement Learning* (RL) approaches for learning entanglement distribution policies [10, 11].

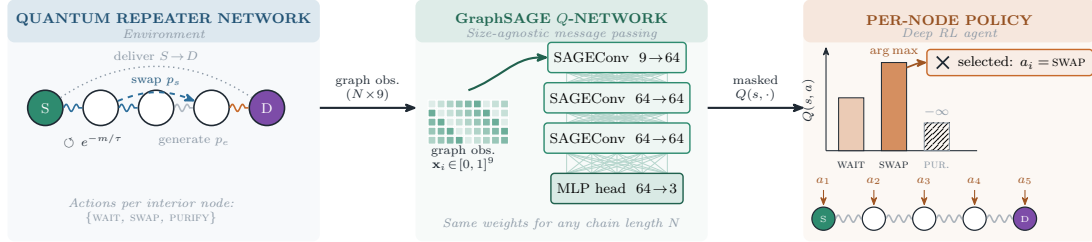


FIG. 1: Schematic of the training and inference procedure. (*Left panel*) The start (S) and destination (D) nodes in a repeater chain must be entangled, by performing repeated swap operations on entangled neighbors. The links between neighbors decay and have a cutoff time $1/\tau$, and operations on the nodes have non-unit success rates. (*Middle panel*) A graph representation of the state of the network is passed to the GNN model, which consists of several layers and produces Q -values. (*Right panel*) The policy is computed from masked Q -values and selects one action *per node*. This architecture enables training and evaluation at arbitrary amounts of nodes N .

The advantage of *learning* a policy arises from the stochastic nature of the underlying network. While entanglement swapping, the core mechanism of extending the range of entanglement enjoys a close to 1 gate fidelity, elementary (nearest-neighbor) link generation technologies still struggle to achieve near-deterministic performance [5]. This is one of the main sources of stochasticity in the environment and the reason why hand-crafted heuristics (such as the swap-as-soon-as-possible [12, 13] or the doubling protocol [14]) achieve far-from-optimal in most scenarios [13, 15].

Learned policies have been shown to outperform these heuristics [16] but suffer one major problem: lack of scalability, inherited from the fixed input/output neural network architectures used. Previous work has addressed several approaches to alleviate these issues. For restrictive scenarios (small homogeneous chains) [17] proved that the MDP can be solved exactly and one can arrive at policies that are provably optimal for the given hardware parameters. However, as system size increases, computing this optimum becomes intractable. For this reason, the authors in Ref. [18] used deep RL to examine different cutoff policies, and Ref. [19] utilized an RL agent to select the best route through a quantum network. The authors in Ref. [10] showed that the advantage of using an RL agent as the policy over set heuristics increases in scenarios where classical communication delays are considered. Later, the authors of Ref. [20] showed that “quasi-local” policies offer a more attractive alternative for multiplexed quantum repeaters with many qubits per node, achieving better connectivity rate even when the classical communication overhead of such policies is taken into account.

In this work, we achieve scalability by proposing a graph neural network (GNN) architecture for parameterizing the policy of an RL agent. We show that this architecture enables size-agnostic policies by learning

actions that are local to nodes, and can be readily applied to arbitrarily large quantum networks. Through message passing steps in the GNN we arrive at policies that take into account information from the k nearest neighbors of a given node for planning actions. This can be contrasted to Ref. [11], which used a nested policy scheme for scalability.

The combination of GNNs and RL has been used to solve routing problems in networks [21] and more recently in quantum networks as well [22, 23]. While closely related to our approach, they solve a complementary problem, that of *routing* in the network. Our approach focuses on the *policy* once a given route is selected. Hence, our agent is in charge of action scheduling in the spirit of Refs. [10, 11, 17]. This means that it can be deployed *on top* of existing routing approaches as an alternative policy for faster intra-route delivery.

II. METHOD

Setup & Actions

We consider multiplexed swap-based repeater chains of length N with n_{ch} qubits per node, with the goal of entangling the chain end-to-end. We formulate the task as a discrete-time optimization problem, where a reinforcement learning agent is provided with a state encoding: average fidelity of existing links in the repeater $\bar{F}$, average occupancy as a percentage of the total qubits of the repeater $\bar{O}$, fraction of qubits available to perform entanglement swapping $\bar{A}$, average age of links connected to the repeater $\bar{U}$, the values of operation qualities p_{gen} , p_{swap} , two binary features encoding whether the repeater has the ability to perform the swapping and purification operations C_S, C_P and a binary flag to distinguish the end-nodes. We resort to averages so that the state can be adapted to repeaters with varying numbers of qubits. At every step the agent is allowed to perform one of 3 actions per-node: do nothing (WAIT), perform entanglement swapping (SWAP) which succeeds with probability p_{swap} , or perform entanglement purification which succeeds with a link-fidelity-dependent probability p_{purify} according to the BBPSSW protocol [24]. Elementary link generation (with per-node probability p_{gen}) is handled by a background process, where each repeater attempts heralded entanglement generation with all its neighbors at the end of every time-step. Whenever the agent chooses to swap, we follow a *quasi-local* policy in which, when multiple swap candidates exist, the one that results in the longest link is chosen [20]. The networks carry homogeneous quantum memories with memory coherence times τ , under a uniform depolarizing noise model. In addition to this, the environment sets a global cutoff time similar to [13]; every episode terminates after h discrete steps have elapsed and a local cutoff in which each link expires after time $1/\tau$.

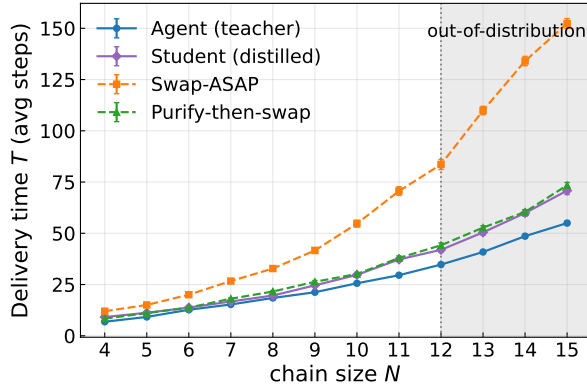
Simulation & Baselines

At every discrete time-step the state of each quantum repeater and its k immediate neighbors (k -local) is passed to the Deep Q-Network RL agent (cf. 1). A GNN [25] parses the state and the agent then outputs a Q-value *per-node* i for every possible action $\mathcal{A}_i \in \{\text{WAIT}_i, \text{SWAP}_i, \text{PURIFY}_i\}$. The action is applied ϵ -greedily followed by a background entanglement generation step. The reward of the environment is $R = -\alpha + F$, where F is the resulting fidelity of the final end-to-end link and $-\alpha$ a small per-step penalty that biases towards fast delivery. As a performance metric we track the average time to achieve end-to-end entanglement. We compare our agent with two baselines: the farthest-neighbour-swap-asap strategy and its purification counterpart distill-swap (or purify-then-swap) [20].

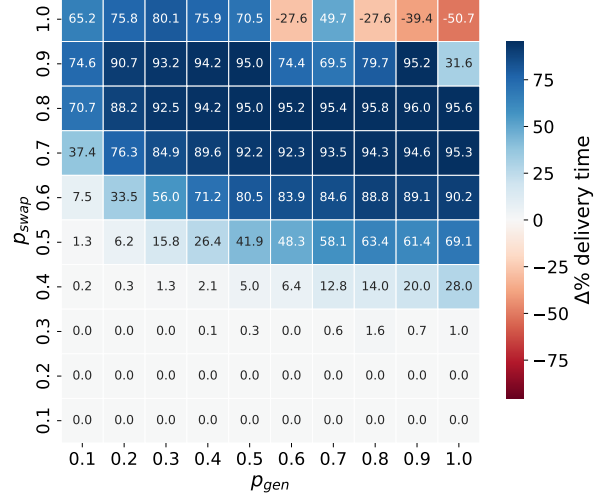
III. RESULTS

We train an agent on a (randomized) mixture of system sizes $N = 4 - 12$ and parameters $p_{\text{gen}}, p_{\text{swap}}, \tau$, and compare its policy *zero-shot*, i.e. without retraining, on larger sizes $N = 10 - 15$. We find that on average our agent outperforms both the swap-asap and purify-then-swap strategies in terms of end-to-end delivery times T on all the combinations of parameters (apart from the case of completely deterministic operations) even for system sizes unseen during training. The same weights are reused for each node, so the agent learns policies that depend only on the k -local information (we use $k = 3$) of each node and therefore can in principle generalize to any N as long as k -local information is sufficient.

We show that one can train an agent in a relatively easy setting (small number of nodes, small imperfections in the parameters) and deploy it on a larger quantum network and greater stochasticity in the environment (smaller $p_{\text{gen}}, p_{\text{swap}}$). To the best of our knowledge, this is the first quasi-local policy that is *learned* rather than hand-designed. In addition, the reinforcement learning approach lends itself well to investigating interpretability and the feasibility of extracting a hardcoded heuristic. We test this hypothesis by employing



(a) End-to-end delivery time T of the two agents (full agent and distilled student-version) and two baselines (swap-asap and purify-swap). The full agent outperforms all the other strategies with a margin increasing with N , even for sizes not seen during training ($N > 12$). The student-agent, while not on par with the full agent still manages to marginally outperform the best heuristic. Parameters for this setting are $p_{\text{gen}} = 0.4$, $p_{\text{swap}} = 0.8$ and $n_{\text{ch}} = 4$, cutoff = 20. Error bars show standard error of the mean over 1000 episodes per point (2000 for $N > 12$ on the agent and heuristic curves) and are mostly smaller than the markers (mean SE $\approx 0.6 - 1.5$ steps).



(b) Performance gap of the agent to the swap-asap (speedup percentage) breakdown for different $p_{\text{gen}}, p_{\text{swap}}$. The agent loses its advantage at the deterministic limit (upper right corner). The finite memory cutoff creates a lower envelope where both strategies fail to deliver due to rapid decoherence. Therefore near-zero cells are delivery-failure regions here and not noise. Average values collected over 200 episodes for each cell, in a different validation experiment from (a).

FIG. 2: The performance of the learned policies compared to two hand-crafted policies as a function of the system size N (a) and as a function of the $p_{\text{gen}}, p_{\text{swap}}$ (b)

a teacher-student strategy as follows: We construct a GNN with reduced number of input features ($9 \rightarrow 3$) and reduced access to local information ($k = 3 \rightarrow 1$). We find that the student model learns meaningful action scheduling, outperforming the purify-swap strategy in some instances (see Fig. 2b). Finally, the incorporation of classical communication delays is straightforward in our current setup and preliminary results show that our learned policy is still able to outperform the hand-crafted heuristic baselines. Rigorous examination of this is left for future work.

IV. CONCLUSION

In conclusion, our research shows the construction and successful learning of a system-size N independent entanglement distribution policy. We do so by training a Deep-Q reinforcement learning agent with a GNN architecture, and demonstrate that it outperforms standard baselines across a wide spectrum of parameter space. This is an important feature for future deployment in larger networks (such as in [19, 22, 23]). In addition, our architecture provides a natural framework for incorporating the effect of classical communication delays in the network as the learned policies are strictly k -local and therefore limit the amount of classical communication overhead needed.

-
- [1] C. H. Bennett and G. Brassard, Quantum cryptography: Public key distribution and coin tossing, *Theoretical Computer Science* **560**, 7 (2014).
 - [2] M. Caleffi, M. Amoretti, D. Ferrari, J. Illiano, A. Manzalini, and A. S. Cacciapuoti, Distributed quantum computing: A survey, *Computer Networks* **254**, 110672 (2024).
 - [3] P. Kómár, E. M. Kessler, M. Bishof, L. Jiang, A. S. Sørensen, J. Ye, and M. D. Lukin, A quantum network of clocks, *Nature Physics* **10**, 582 (2014).
 - [4] D. Gottesman, T. Jennewein, and S. Croke, Longer-baseline telescopes using quantum repeaters, *Physical Review Letters* **109**, 10.1103/physrevlett.109.070503 (2012).
 - [5] S. Wehner, D. Elkouss, and R. Hanson, enQuantum internet: A vision for the road ahead, *Science* **362**, 1 (2018), accepted Author Manuscript.
 - [6] H. J. Kimble, The quantum internet, *Nature* **453**, 1023 (2008).
 - [7] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert, “event-ready-detectors” bell experiment via entanglement swapping, *Phys. Rev. Lett.* **71**, 4287 (1993).
 - [8] B. Hensen, H. Bernien, A. E. Dréau, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenbergh, R. F. L. Vermeulen, R. N. Schouten, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminiau, and R. Hanson, enLoophole-free Bell inequality violation using electron spins separated by 1.3 kilometres, *Nature* **526**, 682 (2015).
 - [9] K. Azuma, S. E. Economou, D. Elkouss, P. Hilaire, L. Jiang, H.-K. Lo, and I. Tzitrin, Quantum repeaters: From quantum networks to the quantum internet, *Reviews of Modern Physics* **95**, 10.1103/revmodphys.95.045006 (2023).
 - [10] J. Li, T. Coopmans, P. Emonts, K. Goodenough, J. Tura, and E. van Nieuwenburg, Optimising entanglement distribution policies under classical communication constraints assisted by reinforcement learning (2024), arXiv:2412.06938 [quant-ph].
 - [11] S. Haldar, P. J. Barge, S. Khatri, and H. Lee, Fast and reliable entanglement distribution with quantum repeaters: Principles for improving protocols using reinforcement learning, *Physical Review Applied* **21**, 10.1103/physrevapplied.21.024041 (2024).
 - [12] L. Kamin, E. Shchukin, F. Schmidt, and P. van Loock, Exact rate analysis for quantum repeaters with imperfect memories and entanglement swapping as soon as possible (2022), arXiv:2203.10318 [quant-ph].
 - [13] K. Goodenough, T. Coopmans, and D. Towsley, On noise in swap ASAP repeater chains: exact analytics, distributions and tight approximations (2024), arXiv:2404.07146 [quant-ph].
 - [14] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, enQuantum Repeater: The Role of Imperfect Local Operations in Quantum Communication, *Physical Review Letters* **81**, 5932 (1998).
 - [15] W. Munro, K. Azuma, K. Tamaki, and K. Nemoto, Inside quantum repeaters, *IEEE Journal of Selected Topics in Quantum Electronics* **21**, 78 (2015).
 - [16] E. Shchukin and P. v. Loock, Optimal entanglement swapping in quantum repeaters, *Physical Review Letters* **128**, 150502 (2022).
 - [17] Á. G. Iñesta, G. Vardoyan, L. Scavuzzo, and S. Wehner, Optimal entanglement distribution policies in homogeneous repeater chains with cutoffs, *npj Quantum Information* **9**, 10.1038/s41534-023-00713-9 (2023).
 - [18] S. D. Reiß and P. van Loock, Deep reinforcement learning for key distribution based on quantum repeaters, *Physical Review A* **108**, 10.1103/physreva.108.012406 (2023).
 - [19] L. Le and T. N. Nguyen, DQRA: Deep Quantum Routing Agent for Entanglement Routing in Quantum Networks, *IEEE Transactions on Quantum Engineering* **3**, 1 (2022).
 - [20] S. Haldar, P. J. Barge, X. Cheng, K.-C. Chang, B. T. Kirby, S. Khatri, C. W. Wong, and H. Lee, Reducing classical communication costs in multiplexed quantum repeaters using hardware-aware quasi-local policies, *Communications Physics* **8**, 132 (2025).
 - [21] P. Almasan, J. Suárez-Varela, K. Rusek, P. Barlet-Ros, and A. Cabellos-Aparicio, Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case, *Computer Communications* **196**, 184 (2022).
 - [22] T. Meuser, J. Weil, A. Lahiri, and M. Paraschiv, RELiQ: Scalable Entanglement Routing via Reinforcement Learning in Quantum Networks, *IEEE Transactions on Communications* **74**, 1860 (2026), arXiv:2511.22321 [quant-ph].
 - [23] T. Meuser, J. Weil, A. Lahiri, and M. Paraschiv, SatQNet: Satellite-assisted Quantum Network Entanglement Routing Using Directed Line Graph Neural Networks (2026), arXiv:2604.09306 [quant-ph].
 - [24] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, Purification of noisy entanglement and faithful teleportation via noisy channels, *Physical Review Letters* **76**, 722 (1996).
 - [25] W. L. Hamilton, R. Ying, and J. Leskovec, Inductive representation learning on large graphs (2018), arXiv:1706.02216 [cs.SI].