

Quantum circuit optimization with deep reinforcement learning

Thomas Fösel,^{1,2,3} Murphy Yuezhen Niu,⁴ Florian Marquardt,^{1,2} and Li Li (李力)³

¹*Max Planck Institute for the Science of Light, Staudtstr. 2, 91058 Erlangen, Germany*

²*Physics Department, University of Erlangen-Nuremberg, Staudtstr. 5, 91058 Erlangen, Germany*

³*Google Research, Mountain View, CA 94043, USA*

⁴*Google Research, Venice Beach, CA 90291, USA*

A central aspect for operating future quantum computers is quantum circuit optimization, i.e., the search for efficient realizations of quantum algorithms given the device capabilities. In recent years, powerful approaches have been developed which focus on optimizing the high-level circuit structure. However, these approaches do not consider and thus cannot optimize for the hardware details of the quantum architecture, which is especially important for near-term devices. To address this point, we present an approach to quantum circuit optimization based on reinforcement learning. We demonstrate how an agent, realized by a deep convolutional neural network, can autonomously learn generic strategies to optimize arbitrary circuits on a specific architecture, where the optimization target can be chosen freely by the user. We demonstrate the feasibility of this approach by training agents on 12-qubit random circuits, where we find on average a depth reduction by 27% and a gate count reduction by 15%. We examine the extrapolation to larger circuits than used for training, and envision how this approach can be utilized for near-term quantum devices.

I. INTRODUCTION

The long-term goal of fault-tolerant large-scale quantum computing promises disruptive progress in multiple important areas of science and technology [1], such as decryption [2], database search [3], quantum simulation [4], and for optimization problems [5]. However, the considerable overhead required for fault-tolerant operations implies a strong incentive in the near term to explore applications that can already work on “noisy intermediate-scale quantum” (NISQ [6]) devices. With the help of suitable algorithms [7, 8], such devices will be able to produce results that outperform classical computers, despite a moderately high rate of errors (“noise resilience”). Recently, a quantum advantage has been demonstrated experimentally for some first benchmark problems on NISQ devices [9, 10].

Research on NISQ applications aims to deliver useful results while keeping the overall resources (number of qubits, computation time) required at a manageable level. Quantum circuit optimization (QCO) constitutes an essential step in addressing this challenge. Starting from a given quantum circuit, i.e., a sequence of quantum operations (gates) acting on a set of qubits, the goal of QCO is to find a logically equivalent circuit that is likely to reduce the probability of errors, by having an overall shorter runtime and using fewer gates (possibly with emphasis on gates of a certain type). This can be achieved mainly by compression (via combination or cancellation) and efficient parallelization of the quantum gates.

There has been important progress in QCO over the years, giving rise to a set of approaches like the family of T-par [11, 12], TOpt [13] and T-Optimizer [14], and QCO based on ZX calculus [15]. These approaches are inspired by the requirements of large-scale fault-tolerant quantum computation and therefore work on scenarios which are quite removed from what will be encountered in NISQ hardware devices. In fact, they have been formulated in a

hardware-independent way and rather operate on a global level.

What is especially needed for QCO on NISQ devices is, however, an approach that is able to take into account the optimization opportunities afforded by a more detailed consideration of the actual hardware (connectivity, gate set, etc.) [16]. Finding the right strategy for this task is a challenging problem. The benefits of any change in the quantum circuit are very context-dependent, especially if the circuit structure is irregular, and a greedy strategy (to always aim for the immediate improvement) is often not ideal.

Optimizing circuits by hand is very time-consuming, requires revising the strategy for every new platform, and can be prone to human errors. Hard-coded algorithmic strategies require considerable human implementation effort specific to the given architecture. In situations with complex decision-making, hard-coded strategies are also prone to unconsidered special cases impeding their ability to generate high-quality results. Optimization techniques like genetic algorithms or simulated annealing improve on these aspects, but lack computational efficiency because they do not select actions in a deliberate way, and they are unable to transfer the experience from earlier solved problem instances.

Optimizing strategies for complex decision-making problems is the goal of reinforcement learning (RL) [17], an important subfield of machine learning that goes beyond straightforward supervised learning. Based on repeated attempts, strategies which have been successful in the past are reinforced over time (exploitation), while alternatives are still occasionally tried out in a trial-and-error fashion (exploration). Due to the generality of this approach, RL is being employed in a broad and growing spectrum of applications, such as robotics [18], video games [19, 20], board games like chess and Go [21, 22], chemistry [23–25], neural architecture search [26], and chip placement [27]. Because RL discovers strategies autonomously and

therefore does not rely on a teacher, it often achieves super-human performance in situations where such a comparison can be made.

In recent years, RL has also been proposed for several problems in the field of quantum computing. Examples include quantum phase estimation [28], the design of quantum experiments [29], quantum control [30–33], quantum error correction [34–37] (alongside other machine learning approaches [38–41]), and quantum metrology [42, 43].

In this work, we introduce deep reinforcement learning for quantum circuit optimization. Our approach enables the computer to autonomously discover strategies for reducing the depth and gate count of quantum circuits, for arbitrary gate sets and connectivity. It allows to choose the optimization target at will and permits extrapolation of the discovered strategy to larger circuits. Due to its flexibility and generality, the RL approach proposed here has the potential to become a valuable component of the toolbox needed to unlock the power of NISQ devices in the near future.

II. TECHNIQUE

A. Quantum circuit optimization as reinforcement learning problem

The goal of RL is to discover strategies for decision-making problems. This is described by an “agent” interacting with the rest of the world, the “environment”. In several rounds, the agent receives information from the environment and, in response to this observation, chooses an action which alters the state of the environment. The agent is supposed to adapt its strategy so as to maximize a success measure, the “reward”. More information is provided in Sec. II C.

In the spirit of previous RL applications to quantum problems [28, 30–34, 42, 43], the obvious approach seems to let the agent build a circuit gate by gate to implement a certain target operation. However, this would come with two central problems here. First, it is extremely unlikely to find a suitable circuit by chance, so an untrained agent would in practice probably never see a positive reward signal. This problem is exacerbated by the fact that the gate set is typically not discrete, but gates can depend on continuous parameters. Second, in the particularly interesting quantum supremacy regime where the circuit cannot be simulated on a classical computer, there is the problem that even if one had found a valid circuit, verifying its correctness would be very hard and computationally expensive. Note that some tools like ZX calculus promise to arrive at a statement in polynomial time, but with the two possible results being positive or inconclusive whether two circuits are equivalent.

Therefore, we follow a different strategy that appears more promising: In QCO, it is common to start from a complete and correct, but typically inefficient circuit, and to progressively optimize it by applying a sequence of

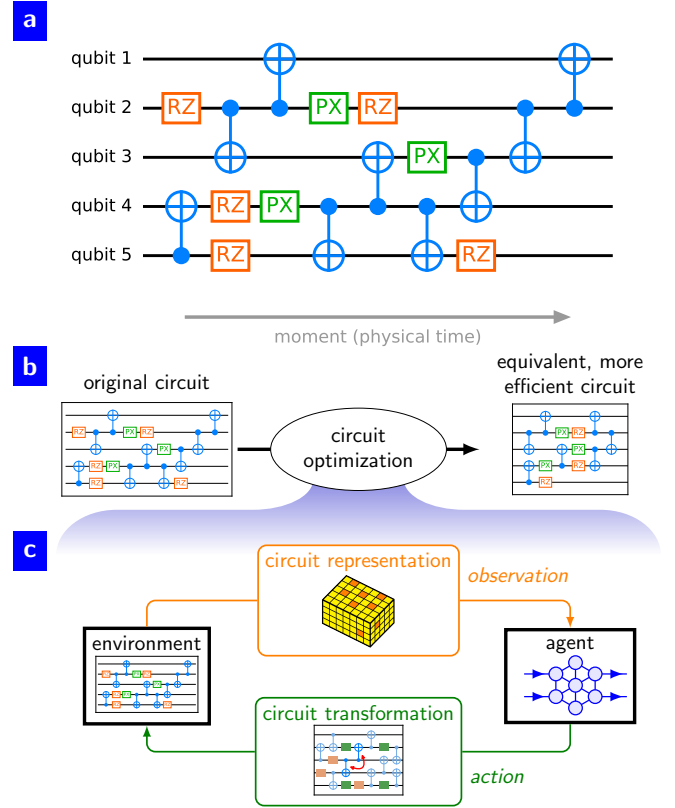


Figure 1. Overview. a) Diagram representation for quantum circuits. Each qubit is indicated with one line. The colored symbols represent operations (gates) on these qubits, with time increasing to the right. b) Quantum circuit optimization. For a given circuit, we aim to find a logically equivalent, but more efficient representation. c) Our reinforcement learning approach to quantum circuit optimization. Based on a diagram-like representation of the circuit, the agent, realized by a neural network, can choose between several circuit transformations to generate another, logically equivalent circuit; this process is repeated multiple times.

circuit transformations. However, it can be a formidable challenge to appropriately choose these transformations, and we make this decision the task of our agent. From the RL perspective, this means that *the states are the circuits* and *the actions are the circuit transformations*. By design, this approach immediately solves the challenge to finish with a correct, i.e., logically equivalent, circuit: we can preserve this property for the full process by allowing in each step only equivalence transformations. In addition, our approach is also scalable, i.e., it allows us to operate in the quantum supremacy regime: it is sufficient to verify equivalence for the few operations directly involved in an elementary circuit transformation, which is relatively cheap as long as all operations act only on a limited number of qubits.

Our general goal is to use RL to train a multi-purpose agent which afterwards will be able to optimize a wide class of circuits based on one given hardware architecture, without going through the RL procedure again in each

instance. Sometimes, though, it can be helpful to fall back to the more restricted approach, where an agent is trained to optimize one specific circuit only.

For our approach, we need to carefully distinguish between different notions of time: the physical time which refers to the execution order of the operations in the circuit, and the agent time in which the actions, here to modify the circuit, are applied. In contrast to comparable previous works on RL for quantum problems [28, 30–34, 42, 43], these two time notions are here independent. From the agent’s perspective, physical time is treated as an extra spatial dimension of the observation (cmp. Sec. II E). In addition, there is the training time along which the weights and thus the behavior of the agent changes.

B. Circuit transformations

Our goal is to optimize a circuit by applying a suitable sequence of circuit transformations. These transformations are based on local transformation rules. Such a rule could be to remove two subsequent operations that cancel each other; or, to reverse the order of two operations that commute (or anti-commute, as a global phase does not matter). Note that one given rule might be applicable to multiple locations in the circuit, yielding an independent transformation for each of them.

In our approach, we distinguish two kinds of transformation rules, which we refer to as “hard” and “soft” rules, and likewise call the transformations they induce hard and soft. Hard rules are those which are always advantageous, like the example of cancelling operations (if the goal is to shrink the circuit, cmp. Sec. II D). Soft rules, on the other hand, are those where the benefit of the transformation depends on the context. For example, for two commuting operations there might be a preferable order contingent on the surrounding operations, and dependent on the current configuration, it can make sense to exchange them or not. Also, it is not uncommon for soft transformations to be beneficial only in combination with several other hard or soft transformations.

In our approach, the decision of the agent is restricted to which soft transformation to apply next. This is always followed by a “pruning” step where automatically all permissible hard transformations are applied. In this sense, the decision of the agent can also be interpreted as not only selecting the soft, but implicitly also all ensuing hard transformations.

Currently, we determine and implement the small set of acceptable transformation rules by hand. The effort for this is manageable because only a moderate number of real-world quantum architectures exists, each with only a limited number of rules according to its specific gate set. As soon as we have arrived at a proper set of rules for one architecture, we can deal with any circuit on it.

C. Learning algorithm

The full RL training process is divided into several episodes. In our case, an episode corresponds to one optimization run for a specific circuit (which might change between episodes). Each episode itself is a sequence of T subsequent steps, in which the agent changes the state s_t with its action a_t into the new state s_{t+1} . For this, the agent receives a reward r_t ; in our case, the reward indicates the improvement of the circuit compared to the previous one. The objective in RL is to maximize the cumulative reward $\sum_{t=0}^{T-1} r_t$. This makes the return

$$R_t = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'}, \quad (1)$$

where γ is a discount rate, an important quantity: For $\gamma = 1$, maximizing $\mathbb{E}[\sum_{t=0}^{T-1} r_t]$ would be equivalent to greedily maximizing $\mathbb{E}[R_t]$ in each time step; in practice, however, a value $\gamma < 1$ often makes learning more stable and efficient.

The agent is typically implemented by a neural network, or in general by an arbitrary parameterized ansatz. The interpretation of its output is defined by the RL algorithm, which also specifies how to train the agent’s parameters θ . In this work, we use proximal policy optimization (PPO [44]), from the family of advantage actor-critic (AAC) methods. An AAC agent computes two quantities: the policy $\pi_\theta(a|s)$, according to which the actions are probabilistically chosen; and an estimator for the state value $V_\theta(s)$, that is the expectation value for the return R when starting from state s and following policy π_θ . In our case, the meaning of the state value $V_*(s)$ for the best policy π_* is the further potential for optimizations of the circuit representing state s .

AAC methods cooptimize policy π_θ and state value V_θ during training [17]: Actions a_t are reinforced according to their advantage $r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)$, whose expectation value tells how much the average return $\mathbb{E}_{\pi_\theta}[R_t|s_t, a_t] = \mathbb{E}_{\pi_\theta}[r_t + \gamma V_\theta(s_{t+1})|s_t, a_t]$ for the chosen action a_t exceeds the average over all actions, that is $V_\theta(s_t)$. Simultaneously, the state-value estimator $V_\theta(s)$ is refined (and adjusted to changes in the policy π_θ) so as to satisfy the Bellman equation $V_\theta(s) = \mathbb{E}_{\pi_\theta}[r + \gamma V_\theta(s')|s]$. In the simplest case, this can be achieved by updating θ according to the learning gradient

$$g = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{T-1} (r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)) \frac{\partial}{\partial \theta} [\ln \pi_\theta(a_t|s_t) - \lambda V_\theta(s_t)] \right] \quad (2)$$

where the coefficient λ weights the cost term for the state value. PPO extends this basic AAC scheme in several ways for more powerful and efficient learning.

D. Reward

The central motive behind QCO is often to decrease the probability for potential errors to occur. More generally speaking, we always want to improve some arbitrary desirable property of a circuit, which is defined by the user. The reward, which decides towards which behavior RL steers the agent, needs to be constructed according to this goal. Let $q(s)$ be a function which quantifies the desirable property for a given circuit s , and is to be minimized. A successful agent is expected to progressively improve the circuit over the course of an episode, so the goal is to minimize $q(s_T)$ at the final step T . We drive the agent towards this behavior with an immediate reward of the form

$$r_t = -(q(s_{t+1}) - q(s_t)). \quad (3)$$

This provides better time resolution than a final reward, such as $r_t \sim -q(s_T) \cdot \delta_{t,T-1}$ or $r_t \sim -(q(s_T) - q(s_0)) \cdot \delta_{t,T-1}$, and therefore promises more efficient learning.

Now, we will derive a simple form for $q(s)$ which estimates the error probability for circuit s , or – to be precise – is a monotonously increasing function thereof. The quite simplistic, but still reasonable model behind it is to assume that each qubit loses information with a certain rate Γ while being idle, and that every gate induces additional errors. Let u_k indicate how much the k -th gate underperforms compared to decay with rate Γ during idle times. Then, the probability to execute a circuit error-free is given by

$$P_{\text{success}} = e^{-m\Gamma T} \cdot \prod_{k=1}^n u_k \quad (4)$$

where m is the number of qubits, n is the number of gates in the circuit, and T denotes the total runtime. Our goal to maximize P_{success} is equivalent to minimizing $q = -\ln P_{\text{success}} = m\Gamma T - \sum_{k=1}^n \ln u_k$.

During application of our RL approach, we will have to make concrete choices for the parameters, but we want to remain as general as possible. The least special assumption we can make is to attribute unit cost to all operations, i.e., an equal execution time $\alpha/(m\Gamma)$ and an equal underperformance factor $u_k = e^{-\beta}$. In this case, q can be expressed via the circuit depth d and gate count n :

$$q(s) = \alpha d(s) + \beta n(s). \quad (5)$$

Explicitly, we will set $\alpha = 1$ and $\beta = 0.2$ in Sec. III. However, we emphasize that $q(s)$ in Eq. (5) can be replaced with any function of a circuit, and because we employ model-free RL, nothing in the learning algorithm would need to be changed.

E. Representation of observations and actions

With perfect information about the environment state, i.e., the circuit, the agent can make the best decisions.

Therefore, we provide a complete description of the circuit as observation to the agent. In quantum computing, circuit diagrams are widely used to visualize circuits for humans. Our format is inspired by them, but adjusted to the conditions of a neural network: We reserve one input neuron for each possible operation in the circuit, and activate neurons depending on whether the corresponding operation is actually present. These input neurons are arranged on a 3D grid, whose axes correspond to qubit index, moment (time step) and gate class. For gate types with a discrete number of instances, each of them forms its own gate class. For gate types which depend on (continuous) control parameters, we distinguish separate classes for special parameter values (e.g., for Z rotations by $\pi/2$, π and $3\pi/2$), and group all remaining instances of this gate type into one class. This allows to represent quantum circuits for arbitrary choices of their continuous parameters, but to distinguish gates for which additional transformation rules apply. To characterize multi-qubit gates, in principle all qubit indices would have to be indicated. In a 1D chain of qubits where gates are allowed only between nearest neighbors, as in our examples in Sec. III, we can characterize two-qubit gates simply by their lowest-index qubit, and specify by the gate class which roles the qubits play (e.g., control vs. target qubit for a CNOT gate).

A particular challenge in this project was to design the output format to represent the policy of the agent. The basic problem is that the actions, i.e., the transformations of the circuit, are defined by the underlying transformation rule and all the gates affected by this transformation (which are characterized by their location in the circuit). Due to combinatorial explosion, the usual approach for discrete action spaces, to reserve one output neuron for every possible action, is not feasible here. One alternative would be to sort the transformations in a certain way and assign them consecutively to a queue of output neurons. However, this approach is not very promising either since the meaning for each output neuron would change for every circuit, potentially depending on aspects of very distant locations of the circuit, and therefore be very hard for the network to understand. Another alternative would be Q-learning where the agent, instead of computing the action value $Q(s, a)$ for state s and action a , computes $Q(s, s')$ for input state s and output state s' [24]; while conceptually this should work, this RL variant is less widely used and therefore not as well explored, and is restricted to Q-learning.

Instead, inspired by the representation of chess moves in [22], we follow a different approach where we associate transformations with certain locations in the circuit, in a way that uniquely characterizes which gates they affect. In our case, we use the following mapping: for a transformation involving a single gate, we indicate the moment and the smallest qubit index of the gate; for a transformation involving two gates, we indicate the smallest qubit index shared between both gates, and the moment of the first one. By providing multiple neurons per location to

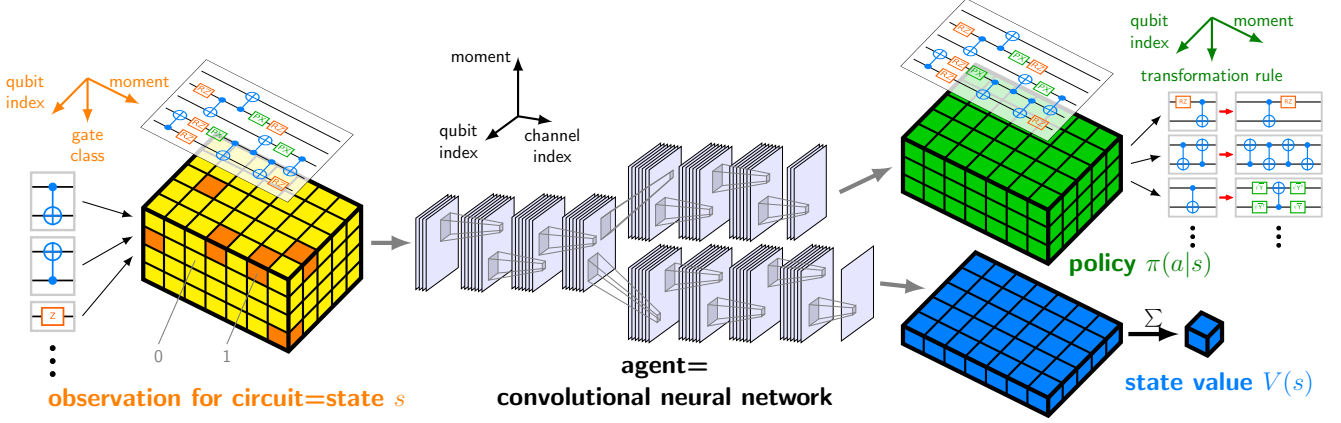


Figure 2. Deep convolutional network architecture of our RL agent. As observation, the agent receives a complete description of the state s , i.e., the quantum circuit. The input neurons are arranged on a 3D grid, whose axes correspond to qubit index, moment and gate class. This information is processed through a stack of multiple convolutional layers, where qubit index and moment are treated as spatial dimensions and the gate classes as input color channels. For the output, the agent computes two quantities: (i) The policy $\pi(s|a)$, according to which the actions a in state s are probabilistically chosen. Every action, i.e., circuit transformation, is mapped uniquely to one policy output neuron; the remaining neurons are disabled with an action mask. And (ii), the state value $V(s)$, which helps to update the policy $\pi(s|a)$ more efficiently during training. For us, $V(s)$ has the meaning of the optimization potential for the circuit.

indicate the underlying rule, we can achieve an injective mapping from transformations to output neurons (for the policy). Therefore, also these neurons are arranged on a 3D grid, whose axes correspond to qubit index, moment and transformation rule. There can be neurons to which no transformation is associated; we disable them with an action mask, whose value changes with the input circuit.

Besides solving the problem to keep the total number of output neurons at a moderate level, another central advantage of this format is that we can exclusively use convolutional layers [45] to process the observation into the policy, treating qubit and moment as spatial dimensions, and the remaining grid axis (gate class and transformation rule, respectively) as input “color” channels. Also to compute the state value, we use convolutional layers (with one output channel), and eventually average over the spatial dimensions. Fig. 2 illustrates the architecture of our deep convolutional network. The weight sharing in the convolutional layers contributes to efficient and robust learning, and a *fully* convolutional architecture will allow us to directly extrapolate to different circuit sizes (see Sec. III B).

F. RL problem classification

The RL problem in this article can be classified as a Markov decision process (MDP) with perfect information (since the circuit representation, which is given to the agent as its input, completely describes the state of the environment). The set of all circuits comprises the state space. The set of possible circuit transformations repre-

sented the action space, which is therefore discrete, and its size depends on the state, i.e., the circuit. The environment is deterministic: a fixed transformation (action) on a fixed circuit (state) always leads to the same outcome. Because the goal is to optimize a property which can be evaluated for any single given circuit, we can construct an immediate reward scheme (as opposed to situations where the reward is given only at the end of an episode).

III. RESULTS

In exploring the power of our RL approach, we need to select both a specific architecture (available gate set, processor layout, qubit connectivity) as well as the family of quantum circuits on whose optimization we want to focus.

Gate set For our simulations, we consider the gate set consisting of Z-Rotation, Phased-X and Controlled-Not (CNOT) gates. Together, they form a universal gate set. Whereas Z-Rotation and Phased-X are actually gate classes parameterized by 1 and 2 continuous variables, respectively, the CNOT is one fixed gate.

For our purposes, Z-Rotation, Phased-X and CNOT is a decent gate set because on the one hand, it induces a rich set of relatively simple transformation rules. On the other hand, they are quite similar to current real-world quantum hardware, such as Google’s Bristlecone (Z-Rotation, Phased-X and Controlled-Z gate [46]) and Sycamore (Z-Rotation, Phased-X and fermionic simulation gate [9, 47, 48]) processor: The CNOT gate differs from the Controlled-Z gate only by local gates (e.g., Hadamard or

Sqrt-Y) on the target qubit, and the fermionic simulation gate is a generalization of the Controlled-Z gate.

A. Training on random circuits

Dataset While a quite large number of quantum circuits for various applications and architectures can be found in publications of the recent years, to the knowledge of the authors a joint dataset of such circuits does not exist, and collecting them by hand would have been very tedious. At the stage of developing the technique and verifying its feasibility, we therefore decided to employ randomly generated circuits. Although the majority of them is likely not to be useful for actual quantum applications, they have the big advantage that they can be easily generated in large numbers. Due to the irregular structure of such circuits, the benefits of transformations are strongly context-dependent, and therefore random circuits constitute a case where formulating a generally applicable strategy would be enormously complex.

We now briefly describe our method to generate these random circuits. We consider 12 qubits arranged in a 1D line; two-qubit gates, here CNOTs, are only allowed between nearest neighbors. First, we build a sequence of 150 gates whose type (either CNOT or local) and the qubits it acts on (constrained by the connectivity) are chosen randomly. Because local gates might be decomposed into a Z-Rotation and a Phased-X gate, we obtain a slightly higher value for the mean number of hardware gates, $\langle n_1 \rangle \approx 159$, and for the mean depth we find $\langle d_1 \rangle \approx 48$. For many of these circuits, some hard rules (cmp. Sec. II B) can immediately be applied, i.e., there are some “trivial” optimizations; for example, two subsequent CNOTs on the same qubits directly cancel. To apply these transformations, which we refer to as pruning, reduces the mean depth to $\langle d_2 \rangle = 37.15 \pm 0.01$ and the mean gate count to $\langle n_2 \rangle = 115.27 \pm 0.03$. From here, we continue by applying random transformations (based on soft rules, each time followed by a pruning step), which turns out to make the circuit much larger and hence very inefficient. After 500 of these transformations, we find a mean depth of $\langle d_3 \rangle \approx 248$ and a mean gate count of $\langle n_3 \rangle \approx 508$.

We use these randomly expanded circuits as the training tasks for the RL agent, where each of them defines the starting point for one episode, i.e., for one optimization run. The random expansion step enhances the ability of the fully trained agent to generalize, by demanding it to deal well with a broad spectrum of both inefficient and efficient circuits, which will be encountered during every successful optimization run.

Simulated annealing As a benchmark for the results by the RL agent, we consider simulated annealing as an alternative approach, using an adapted Metropolis rule and an exponential cooling schedule. We apply simulated annealing to the randomly expanded circuits, i.e., on the same dataset as later for the agent. The results depend

strongly on the number of steps: Within 10000 steps, simulated annealing is able to merely reach the level after pruning (on average). Within 200000 steps, out of which roughly 15...20% are actually accepted, a mean depth $\langle d \rangle = 27.35 \pm 0.08$ and gate count $\langle n \rangle = 105.15 \pm 0.32$ can be achieved.

RL Now, we apply the RL approach as described in Sec. II on our dataset of expanded random circuits. The hyperparameters have been optimized in a systematic search. One full learning process takes 6 to 7 days, on a node with 32 CPU cores.

We find that the learning process can be divided roughly into two phases (cmp. Fig. 3c): In the first ca. 50 epochs, the agent rapidly improves in minimizing the mean depth and gate count, to a level of $\langle d \rangle \approx 29$ and $\langle n \rangle \approx 100$ which is already considerably better than after pruning. In the remaining course of the training, the agent can slowly, but steadily improve towards $\langle d \rangle = 27.20 \pm 0.07$ and $\langle n \rangle \approx 97.86 \pm 0.33$ at around epoch 1000. Hence, the agent can clearly beat the results from simulated annealing, especially regarding the gate count. Because we use a new circuit in each training episode, no difference is to be expected between the performance on training and validation samples.

In Fig. 3d, we examine closer how a trained agent proceeds in optimizing a circuit. For this purpose, we choose 3200 circuits and plot how d and n evolve with every transformation. As Fig. 3d shows, the agent can rapidly improve the circuits during the first 150 to 200 steps. Afterwards, the curve is mostly flat, except for two effects: First, there are occasional excursions into lower-quality states, but the agent usually returns quickly to the previous level. Second, the agent sometimes manages to further optimize the circuit after some time, partially via passing some intermediate lower-quality states, i.e., to cross a reward barrier. Note that we did not give the agent the ability to terminate an episode.

Compared to the pruning level, we have seen that the agent achieves much better values for $\langle d \rangle$ and $\langle n \rangle$ on average. How effective the agent has actually become can be seen by comparing d and n for individual circuits between the agent and after pruning. In Fig. 3e, we plot this comparison for epochs 900 – 1000, i.e., for 3200 circuits in total. The agent achieves a reduction of both d and n for the vast majority of circuits, partially up to even more than 50%. Only for 12 out of 3200 circuits, i.e., in around 0.4% of the cases, the agent fails to achieve a better $q(s)$ than after pruning.

B. Extrapolation to larger random circuits

In Sec. III A, the agent has been trained and evaluated on 12-qubit random circuits. However, due to its fully convolutional architecture, this very same agent can – at least in principle – be directly used to optimize also circuits of a different size, in particular much larger ones. Now, we will test how well the agent can actually generalize its

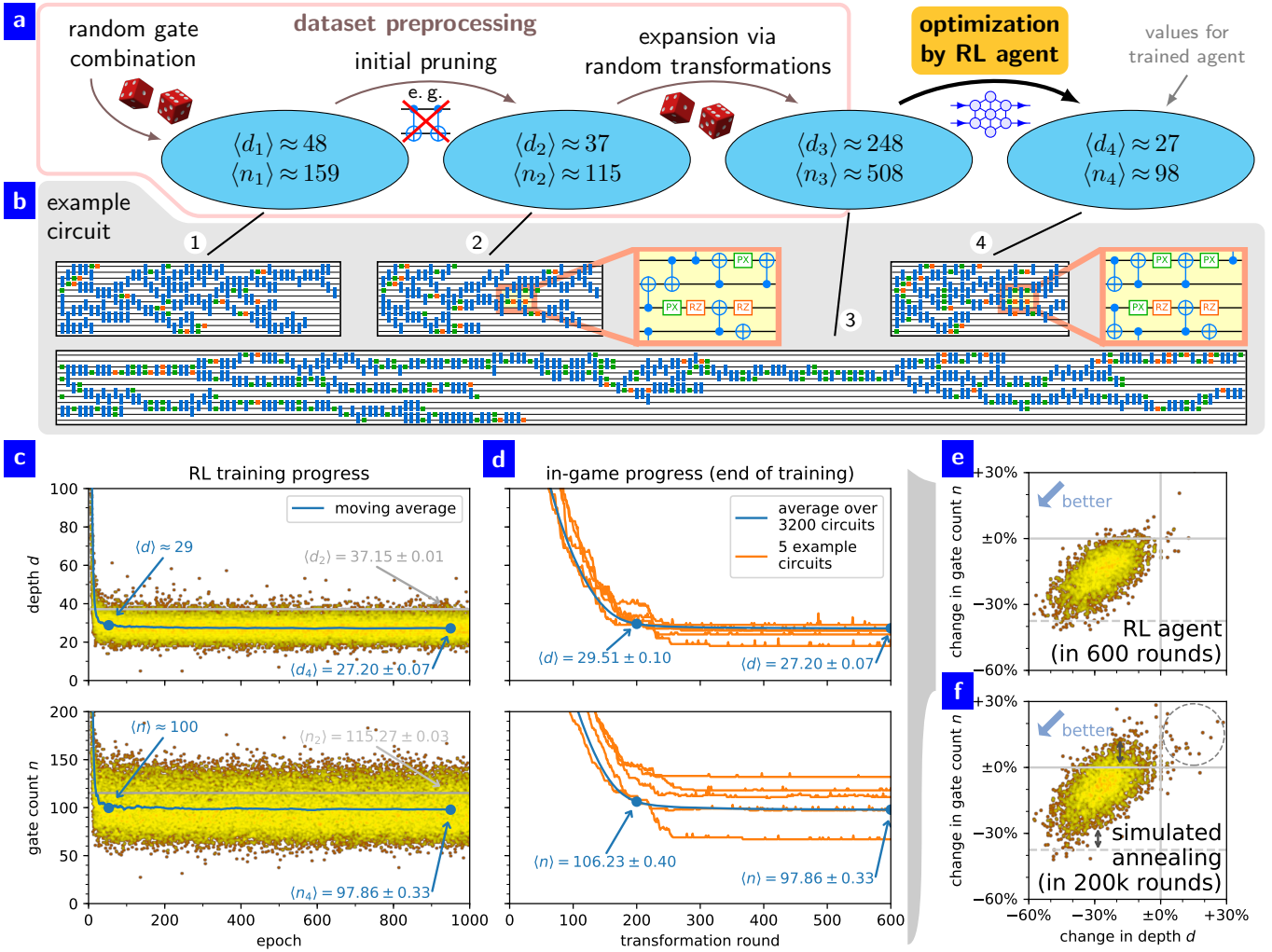


Figure 3. Training on random circuits. a) Circuit processing pipeline (see Sec. III A for details). After choosing an initial circuit by randomly combining gates, a pruning step follows where all “trivial” optimizations are applied. Afterwards, 500 random transformations are performed on this circuit, which turns out to significantly increase their depth d and gate count n . These expanded circuits are then used as the starting point of the episodes to train and evaluate the RL agent. b) Diagrams illustrating the evolution of one example circuit through this pipeline. c) Learning progress during training, demonstrating how the agent improves in reducing both the depth d (top) and the gate count n (bottom) of the circuits. The point cloud indicates, for all episodes during training, the corresponding quantity in the final time step. The blue curve shows the moving average over the latest 10% of epochs. For comparison, the gray line indicates the corresponding averages after pruning; already early in the training, the agent falls below this level for both quantities. d) In-game progress at the end of the learning process, showing for 5 episodes during the last epoch (orange) how the agent progressively optimizes (d, n) during an episode. The blue curve indicates the average over all episodes in the last 100 epochs of training. e) Relative improvement achieved by the RL agent, in reference to the corresponding circuit size after pruning. Each point corresponds to one episode during the last 100 epochs. f) Comparison with circuit optimization by simulated annealing (see Sec. III A for details). The graphical depiction and the considered circuits are equivalent to (e), which makes them directly comparable.

knowledge in this situation.

For this purpose, we reuse the scheme to generate random circuits as described in Sec. III A, except for changing two parameters: we increase the number of qubits from 12 to 50, and the number of initial gates from 150 to 2500. We find $\langle d \rangle = 199.25 \pm 0.08$ and $\langle n \rangle = 2655.3 \pm 1.2$ before pruning, and $\langle d \rangle = 156.67 \pm 0.07$ and $\langle n \rangle = 1940.2 \pm 1.6$ after pruning. Because we will not

use the circuits here to train the agent, we can skip the step to expand them by random transformations, whose purpose it was to feed to the agent also very inefficient circuits during training. Instead, the optimization by the agent starts here directly from the pruned circuits. As shown in Fig. 4b, the agent achieves to reduce $\langle d \rangle$ to 110.84 ± 0.07 and $\langle n \rangle$ to 1616.3 ± 2.0 within 2500 transformations. Remarkably, the reduction ratio in these

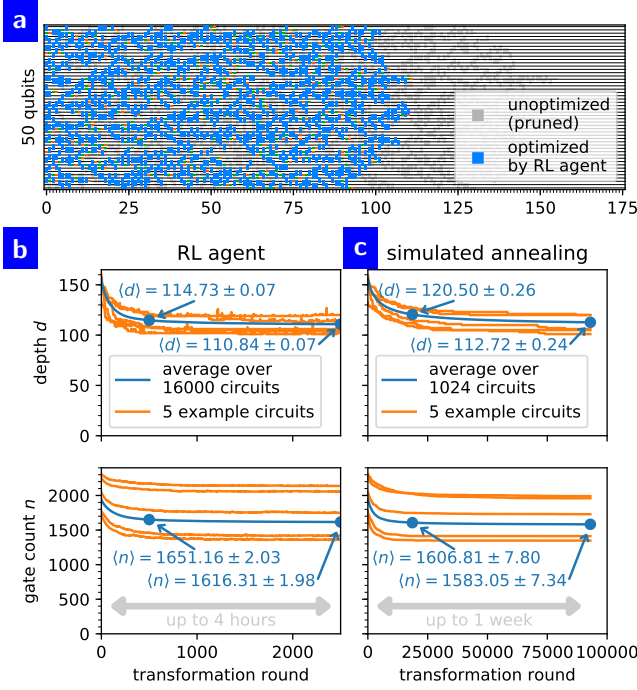


Figure 4. Extrapolation to 50-qubit random circuits. The agent has been trained on 12-qubit circuits (cmp. Fig. 3), no further learning updates are performed here. (a) shows the comparison between an unoptimized example circuit (after pruning) and the result of the optimization by the RL agent. (b) shows the progress of the agent in reducing depth d and gate count n over the course of 2500 transformations. (c) shows the corresponding curves for simulated annealing, which requires almost 100000 transformations to achieve a comparable degree of optimization (the computation was terminated after 1 week, at transformation 93000).

two quantities is comparable to the smaller circuits it has been trained on (cmp. Fig. 3).

Simulated annealing arrives at similar values, $\langle d \rangle = 112.72 \pm 0.24$ and $\langle n \rangle = 1583.0 \pm 7.3$, within 93000 transformations. These are much fewer transformations than required to optimize the smaller random circuits in Sec. III A, probably because here the random expansion step has been skipped. Nevertheless, 93000 transformations for each larger random circuits here have already taken one week (our termination criterion), which is comparable to the time needed to train an RL agent. Afterwards, this agent can optimize arbitrary circuits, in a relatively short time (3...5 h in this case).

Our results show that an agent can actually extrapolate its knowledge to larger circuits. More generally, they demonstrate that our approach, both with RL and simulated annealing, works deep in the quantum supremacy regime. Furthermore, this also highlights a situation where optimizing even a single circuit with simulated annealing needs already a runtime comparable to the full training of an RL agent and subsequently optimizing the particular circuit.

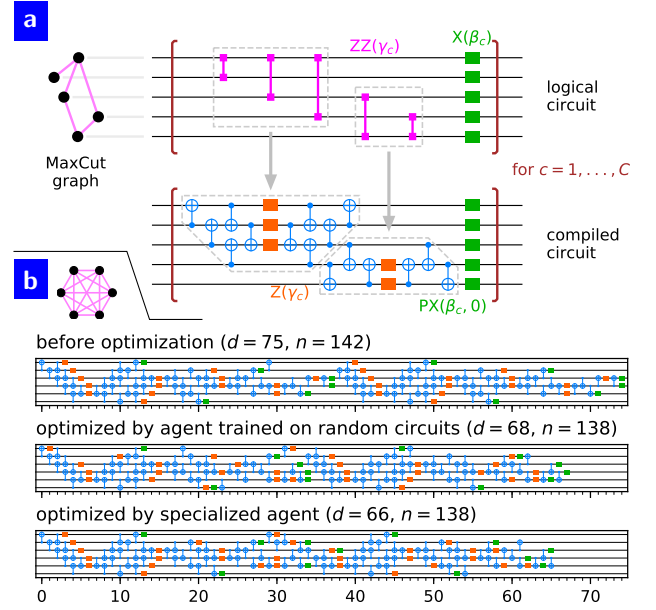


Figure 5. Optimization of QAOA-MaxCut circuits. (a) indicates how to translate the MaxCut problem for a graph into a quantum circuit following QAOA, and how to efficiently compile this logical circuit into our gate set. We display one of M cycles which form the full circuit, each with a different set of parameters (γ_c, β_c) whose values are refined during the QAOA algorithm. (b) shows the compiled circuit for $C = 2$ cycles and an all-to-all-connected graph with 6 nodes, which has depth $d = 75$ and gate count $n = 142$ (top). Using a generic agent trained on random circuits as in Fig. 3, we find (by postselection) improved circuits with $d = 68$ and $n = 138$ (middle). A specialized agent trained on this particular circuit can further optimize it to $d = 66$ and $n = 138$ (bottom).

C. QAOA-MaxCut circuit

As an example for a real-world quantum algorithm, we now consider the MaxCut problem. The goal is to arrange the nodes of an undirected, non-weighted graph into two groups such that the amount of cut edges is maximized. Finding the exact solution is an NP-hard problem. Following the quantum approximate optimization algorithm (QAOA [7]), approximate solutions can be found with the help of a quantum circuit consisting of repeated cycles of ZZ gates and local X rotations with variable angles [48] (cmp. Fig. 5a). We consider the same gate set as in the examples above, such that we can reuse the previously trained agent. Also, this covers the realistic situation where the native gates of the quantum algorithm do not necessarily match the native gates of the hardware. Fig. 5a shows an efficient compilation of this circuit into our gate set, where ZZ gates need to be decomposed into CNOTs and local Z rotations (local X rotations are a special case of Phased-X gates). Note that the variable angles of the gates do not affect the optimization strategy, as long as we assume these angles to be generic (i. e., not set to special values which would

allow additional optimizations).

Every graph corresponds to a different circuit. We restrict our analysis here to all-to-all connected graphs. While any such graph has trivial MaxCut solutions (nodes evenly distributed), the corresponding circuit is still useful for us to consider because the circuit for any other graph layout can be derived from it by removing some of the ZZ gates; therefore, it also provides a worst-case estimate. In Fig. 5b, we specifically consider a graph with 6 nodes and 2 cycles in the QAOA circuit.

First, we try to optimize this QAOA circuit with a generic agent trained on 12-qubit *random* circuits, as in Fig. 3. We find that on average, the agent manages to slightly reduce the gate count, although the depth is slightly increased. Even though this agent is, thus, not able to reliably optimize the circuit, it is still able to sometimes achieve an improvement: By postselecting over all time steps in multiple optimization runs, we find that the best circuit reduces the depth from 75 to 68 and, simultaneously, the gate count from 142 to 138 (cmp. Fig. 5b); note that different runs deviate because the policy is to some degree stochastic, and that the circuit quality fluctuates also with time. We observe that the optimizations happen at the interface between subsequent cycles, and at the end of the last cycle.

If we, however, train an agent via RL directly on this specific circuit, it does not only learn to reliably improve the circuit, but it also finds two further optimizations, each reducing the depth by 1. These optimizations are very hard to discover, as they require a lookahead by 7 and 9 transformations, respectively, where the search breadth is around 350. The resulting circuit, as shown in Fig. 5b, has a depth of 66 and a gate count of 138. To our knowledge, this is the most efficient realization on the gate set we provided (when including additional gates, one can arrive at an even shorter circuit [48]).

For these QAOA-MaxCut circuits, we start from an already quite efficient compilation, so the optimization potential is not very large here. Nevertheless, the agent is still able to find optimizations, including some which require complex transformation sequences and would thus be very hard to find by hand. On the general level, this example demonstrates that our approach can also be used to optimize real-world circuits, where the optimization potential might very well be larger than here. However, we have got reliable improvement and the best optimization quality only for agents trained on this circuit, which emphasizes the importance of a proper training dataset.

During early simulations, we made a noteworthy observation: Sometimes, the agent trained on these MaxCut circuits achieved optimizations far below a depth of 66 and a gate count of 138. It turned out that this happened only if the rotation angle for one cycle of ZZ gates, which was chosen at random, was relatively close to a multiple of π . In this case, the agent found a creative way to exploit that the closeness criterion for one of the transformation rules was too loose, which it used to partially or completely remove this cycle from the circuit. Although in

this particular case these transformations were not valid, the agent’s behavior shows that it would readily exploit the simplification potential in other complex circuits.

D. Discussion

Dataset We anticipate our RL agents to be considerably more powerful and reliable on real-world quantum circuits when trained on a dataset with similar properties, which is also indicated by our results in Sec. III C. In contrast to currently existing collections of quantum circuits, like [49], we need for our purposes first the circuits being compiled for specific hardware architectures, and second, also a significantly larger dataset size to fully train agents on them (based on Sec. III A, this can require in the order of 10000 to 100000 episodes).

We think of our future dataset to be generated as follows. We consider a set of various quantum algorithms, like QAOA [7], variational quantum eigensolver [8], Shor’s algorithm [2] (for small integers), and many more [50]. Many of these algorithms imply a large number of different circuits; for example, QAOA can be applied to multiple combinatorial optimization problems, and for each of them, the circuit changes with the concrete problem instance (e.g., graph layout) and the number of cycles. These circuits are described in a high-level, hardware-independent way. For every specific architecture (gate set, layout, connectivity), we would then generate an own dataset by compiling the circuits for them, which can be done using existing quantum libraries like Cirq [51].

Eventually, each of our agents will be trained on one of these architecture-specific datasets. If the dataset size is not yet sufficient for this, we can use data augmentation techniques (like mirroring the circuit in space and/or time), and to some degree also repeat circuits during training.

RL vs. simulated annealing In Sec. III A, simulated annealing achieved $98.48 \pm 1.06\%$ of the reduction in gate count n by the RL agent, but only $58.06 \pm 2.15\%$ in the depth d ; Sec. III B is more ambiguous, as simulated annealing achieved $109.67 \pm 2.44\%$ of the reduction in n by the RL agent, but only $95.88 \pm 0.57\%$ in d (all numbers in reference to the pruning level). Therefore, in these scenarios, RL tends to give overall slightly better results than simulated annealing.

Another aspect to consider in this comparison is run-time. After having been trained once on a certain class of circuits, an agent is considerably faster than simulated annealing in optimizing arbitrary circuits of this class (ca. 2 min vs. 1...3 d in Sec. III A, 3...5 h vs. 7 d in Sec. III B). Of course, for a fair comparison of the computational effort, it is necessary to also take into account the time needed for training the agent in the first place (up to 1 week), and eventually also the time to optimize the hyperparameters for both approaches. Therefore, if we take all those aspects into account, the question which method is less computationally expensive cannot be an-

swered in general, but depends on the class of circuits and previously solved problem instances – the only safe statement is that for a sufficient number of circuits to be optimized, the RL approach will be computationally cheaper.

Symbolic parameters In our current implementation, gates are represented numerically, which requires to consider fixed choices for the gate parameters. This is not directly a property of the RL approach presented in this work, but of our underlying quantum circuit framework [52]. With manageable effort, this framework can be extended to allow also gates parameterized by symbolic variables. This would simplify the optimization of entire classes of parameterized circuits, like for QAOA with its varying angles β_c and γ_c . So far, this is only indirectly possible by repeating the same sequence of transformations for differently parameterized circuits.

IV. CONCLUSION

In this article, we have introduced RL as a powerful and flexible approach to QCO. In particular, we have focused on the optimization of the local circuit structure to the available hardware resources, which is especially relevant for the NISQ era. The main challenge here is to discover suitable sequences of circuit transformations; the size of the search space and the lack of simple strategies in many situations, especially for circuits with an irregular structure, make this a difficult optimization problem. RL is well-suited to problems of this kind, and promises to simultaneously fulfill the following key features: *hardware-efficient*, i.e., that it can find the optimal solution given the available resources; *cross-platform*, i.e., that the approach works for different quantum architectures (gate sets, chip layouts, connectivities, etc.), without large migration costs; *autonomous*, i.e., that the only human effort is to specify the problem, and the optimization is done by the computer completely on its own, without requiring further interaction; and *reliable*, i.e., that the results are always correct and for a wide spectrum of circuits (close-to-)optimum.

For our approach to be successful, we had to address several key challenges: First, we have formulated QCO as a problem which can be efficiently addressed using RL (cmp. Secs. II A and II B). Second, we have developed efficient schemes to represent circuits for the network input and transformations for the network output (cmp. Sec. II E). Third, as prerequisite for applying techniques like RL or simulated annealing, we have implemented a Python framework to automate the process of identifying

all transformations within a circuit [52].

Both RL and simulated annealing turn out to be valid optimization techniques for the QCO problems considered in this work. In terms of optimization quality, the difference between both approaches in reducing depth and gate count were not larger than 10% (in reference to the pruning level), except for Sec. III A where simulated annealing achieved only $58.06 \pm 2.15\%$ of the depth reduction by the RL agent (cmp. Sec. III D). In terms of computational effort, our RL agents required up to 1 week of training time, but were afterwards considerably faster than simulated annealing in optimizing arbitrary circuits from the class they have been trained on (ca. 2 min vs. 1...3 d in Sec. III A, 3...5 h vs. 7 d in Sec. III B). We expect that RL can deal better with situations where decision-making becomes even harder than here. Therefore, RL seems to be, especially in the long run, the more promising approach.

So far, we have mostly considered randomly generated circuits to circumvent the lack of a sufficient dataset of realistic circuits; one central future direction will be to collect such a dataset. Several direct extensions of our work are straightforward to implement, yet potentially very fruitful for future applications, such as implementations for different gate sets, 2D geometries, etc. A (semi-)automated procedure to identify transformation rules, where machine learning could be beneficial, would further reduce the effort for including new gate sets. The optimization of parameterized circuit classes can be simplified by implementing symbolic variables as gate parameters in the underlying quantum circuit framework. Furthermore, the existing complementary QCO techniques which focus on optimizing the global circuit structure [11–15] could be combined in a two-stage process with our technique to optimize the local circuit structure.

We expect our approach to soon become a standard tool for optimizing circuits in NISQ experiments. Central practical advantages are the negligible additional effort to consider non-ideal circumstances, like inhomogeneous error rates or defect qubits, and the ability to customize the optimization criterion.

Data and code availability The code is publicly available under [52]. Other findings of this study are available from the corresponding author on reasonable request.

Acknowledgements T.F. thanks Google Research for hosting during his internship.

Competing interests The authors declare no competing interests.

Corresponding author Correspondence should be addressed to T.F.

-
- [1] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.
 - [2] Peter W Shor. Polynomial-time algorithms for prime fac-

torization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.

- [3] Lov K Grover. Quantum mechanics helps in searching for

- a needle in a haystack. *Physical review letters*, 79(2):325, 1997.
- [4] Seth Lloyd. Universal quantum simulators. *Science*, pages 1073–1078, 1996.
 - [5] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, and Michael Sipser. Quantum computation by adiabatic evolution. *arXiv preprint quant-ph/0001106*, 2000.
 - [6] John Preskill. Quantum computing in the nisy era and beyond. *Quantum*, 2:79, 2018.
 - [7] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.
 - [8] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.
 - [9] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.
 - [10] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. Quantum computational advantage using photons. *Science*, 370(6523):1460–1463, 2020.
 - [11] Matthew Amy, Dmitri Maslov, and Michele Mosca. Polynomial-time t -depth optimization of clifford+ t circuits via matroid partitioning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(10):1476–1489, 2014.
 - [12] Matthew Amy and Michele Mosca. T -count optimization and reed-muller codes. *IEEE Transactions on Information Theory*, 65(8):4771–4784, 2019.
 - [13] Luke E Heyfron and Earl T Campbell. An efficient quantum compiler that reduces t count. *Quantum Science and Technology*, 4(1):015004, 2018.
 - [14] Fang Zhang and Jianxin Chen. Optimizing t gates in clifford+ t circuit as $\pi/4$ rotations around paulis. *arXiv preprint arXiv:1903.12456*, 2019.
 - [15] Aleks Kissinger and John van de Wetering. Reducing the number of non-clifford gates in quantum circuits. *Physical Review A*, 102(2):022406, 2020.
 - [16] Yuri Alexeev, Dave Bacon, Kenneth R. Brown, Robert Calderbank, Lincoln D. Carr, Frederic T. Chong, Brian DeMarco, Dirk Englund, Edward Farhi, Bill Fefferman, Alexey V. Gorshkov, Andrew Houck, Jungsang Kim, Shelby Kimmel, Michael Lange, Seth Lloyd, Mikhail D. Lukin, Dmitri Maslov, Peter Maunz, Christopher Monroe, John Preskill, Martin Roetteler, Martin J. Savage, and Jeff Thompson. Quantum computer systems for scientific discovery. *PRX Quantum*, 2:017001, Feb 2021.
 - [17] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
 - [18] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
 - [19] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
 - [20] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.
 - [21] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
 - [22] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmashan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
 - [23] Zhenpeng Zhou, Xiaocheng Li, and Richard N Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS central science*, 3(12):1337–1344, 2017.
 - [24] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.
 - [25] Steven Kearnes, Li Li, and Patrick Riley. Decoding molecular graph embeddings with reinforcement learning. *arXiv preprint arXiv:1904.08915*, 2019.
 - [26] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *International Conference on Learning Representations*, 2017.
 - [27] Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Sungmin Bae, et al. Chip placement with deep reinforcement learning. *arXiv preprint arXiv:2004.10746*, 2020.
 - [28] Pantita Palittapongarnpim, Peter Wittek, Ehsan Zahedinejad, Shakib Vedaie, and Barry C Sanders. Learning in quantum control: High-dimensional global optimization for noisy quantum dynamics. *Neurocomputing*, 268:116–126, 2017.
 - [29] Alexey A Melnikov, Hendrik Poulsen Nautrup, Mario Krenn, Vedran Dunjko, Markus Tiersch, Anton Zeilinger, and Hans J Briegel. Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences*, 115(6):1221–1226, 2018.
 - [30] Marin Bukov, Alexandre GR Day, Dries Sels, Phillip Weinberg, Anatoli Polkovnikov, and Pankaj Mehta. Reinforcement learning in different phases of quantum control. *Physical Review X*, 8(3):031086, 2018.
 - [31] Moritz August and José Miguel Hernández-Lobato. Taking gradients through experiments: Lstms and memory proximal policy optimization for black-box quantum control. In *International Conference on High Performance Computing*, pages 591–613. Springer, 2018.
 - [32] Murphy Yuezheng Niu, Sergio Boixo, Vadim N Smelyanskiy, and Hartmut Neven. Universal quantum control through deep reinforcement learning. *npj Quantum Information*, 5(1):1–8, 2019.
 - [33] Riccardo Porotti, Dario Tamascelli, Marcello Restelli, and Enrico Prati. Coherent transport of quantum states by deep reinforcement learning. *Communications Physics*, 2(1):1–9, 2019.
 - [34] Thomas Fösel, Petru Tighineanu, Talitha Weiss, and Florian Marquardt. Reinforcement learning with neural networks for quantum feedback. *Physical Review X*, 8(3):031084, 2018.
 - [35] Ryan Sweke, Markus S Kesselring, Evert PL van Nieuwenburg, and Jens Eisert. Reinforcement learning decoders for

- fault-tolerant quantum computation. *Machine Learning: Science and Technology*, 2(2):025005, 2020.
- [36] Philip Andreasson, Joel Johansson, Simon Liljestrand, and Mats Granath. Quantum error correction for the toric code using deep reinforcement learning. *Quantum*, 3:183, 2019.
 - [37] Hendrik Poulsen Nautrup, Nicolas Delfosse, Vedran Dunjko, Hans J Briegel, and Nicolai Friis. Optimizing quantum error correction codes with reinforcement learning. *Quantum*, 3:215, 2019.
 - [38] Giacomo Torlai and Roger G Melko. Neural decoder for topological codes. *Physical review letters*, 119(3):030501, 2017.
 - [39] Paul Baireuther, Thomas E O’Brien, Brian Tarasinski, and Carlo WJ Beenakker. Machine-learning-assisted correction of correlated qubit errors in a topological code. *Quantum*, 2:48, 2018.
 - [40] Stefan Krastanov and Liang Jiang. Deep neural network probabilistic decoder for stabilizer codes. *Scientific reports*, 7(1):1–7, 2017.
 - [41] Peter D Johnson, Jonathan Romero, Jonathan Olson, Yudong Cao, and Alán Aspuru-Guzik. Qvector: an algorithm for device-tailored quantum error correction. *arXiv preprint arXiv:1711.02249*, 2017.
 - [42] Han Xu, Junning Li, Liqiang Liu, Yu Wang, Haidong Yuan, and Xin Wang. Generalizable control for quantum parameter estimation through reinforcement learning. *npj Quantum Information*, 5(1):1–8, 2019.
 - [43] Jonas Schuff, Lukas J Fiderer, and Daniel Braun. Improving the dynamics of quantum sensors with reinforcement learning. *New Journal of Physics*, 22(3):035001, 2020.
 - [44] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
 - [45] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
 - [46] Julian Kelly. Engineering superconducting qubit arrays for quantum supremacy. In *APS March Meeting Abstracts*, volume 2018, pages A33–001, 2018.
 - [47] Google AI Quantum et al. Hartree-fock on a superconducting qubit quantum computer. *Science*, 369(6507):1084–1089, 2020.
 - [48] Matthew P Harrigan, Kevin J Sung, Matthew Neeley, Kevin J Satzinger, Frank Arute, Kunal Arya, Juan Atalaya, Joseph C Bardin, Rami Barends, Sergio Boixo, et al. Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, pages 1–5, 2021.
 - [49] Pyzx quantum circuit dataset. <https://github.com/Quantomatic/pyzx/tree/master/circuits>. Accessed: 2021-03-12.
 - [50] Quantum algorithm zoo. <https://quantumalgorithmzoo.org/>. Accessed: 2021-03-12.
 - [51] Cirq. <https://github.com/quantumlib/Cirq>. Accessed: 2021-03-12.
 - [52] rl4circopt, our quantum circuit framework. <https://github.com/google-research/google-research/tree/master/rl4circopt>. Accessed: 2021-03-12.