# Curriculum reinforcement learning for optimization of variational quantum circuit architectures

**Mateusz Ostaszewski,**
Institute of Theoretical and Applied Informatics,
Polish Academy of Sciences,
Gliwice, Poland.
mm.ostaszewski@gmail.com

**Wojciech Masarczyk,**
Warsaw University of Technology,
Warsaw, Poland.
wojciech.masarczyk@gmail.com

**Lea M. Trenkwalder,**
Institute for Theoretical Physics,
University of Innsbruck
Innsbruck, Austria
lea.trenkwalder@uibk.ac.at

**Eleanor Scerri,**
Leiden University,
Leiden, The Netherlands.
scerri@lorentz.leidenuniv.nl

**Vedran Dunjko,**
Leiden University,
Leiden, The Netherlands.
v.dunjko@liacs.leidenuniv.nl

## 1 Introduction

As we are entering the so called Noisy Intermediate Scale Quantum (NISQ) [9] technology era, the search for more suitable algorithms under NISQ restrictions is becoming ever important. A truly compatible NISQ application must first be amenable to architecture constraints and size limits. Furthermore, to minimize the adverse effects of gate errors and decoherence, it is important that the circuits we run are as gate-frugal, and as shallow as possible.

Perhaps the most promising classes of such algorithms are based on variational circuit methods, applied to problems in quantum chemistry. A key problem in this field is the computing of ground state energies and low energy properties of chemical systems (the chemical structure problem). This problem is believed to be intractable in general, yet the quantum Variational Quantum Eigensolver (VQE) [8] algorithm can provide solutions in regimes which beyond the reach of classical algorithms, while maintaining NISQ-friendly properties.

VQE is a hybrid quantum-classical algorithm, where a parametrized quantum state is prepared on a quantum computer, the parameters of which are selected using classical optimization methods. The objective is to prepare the state $|\psi(\vec{\theta})\rangle$ which can be used to approximate the ground state of a given Hamiltonian $H$ by the variational principle

$$E_{\min} \leq \min_{\vec{\theta}}(\langle\psi(\vec{\theta})|H|\psi(\vec{\theta})\rangle), \tag{1}$$

where $E_{\min}$ is a ground state energy of $H$. The parametrized state is prepared by applying $U(\vec{\theta})$, which is a fixed-architecture parametrized quantum circuit, where the angles $\vec{\theta} = (\theta_1 ... \theta_n)$ specify the rotation angles of the local unitary rotations the circuit is built from. This circuit, known as the *Ansatz* is applied to the fiducial "all zero" state to prepare the state $|\psi(\vec{\theta})\rangle = U(\vec{\theta})|0\rangle$.

It is well established that the structure of the Ansatz can dramatically influence the VQE's performance [4, 6], as the closeness of the estimated ground state to the true one depends on the state

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

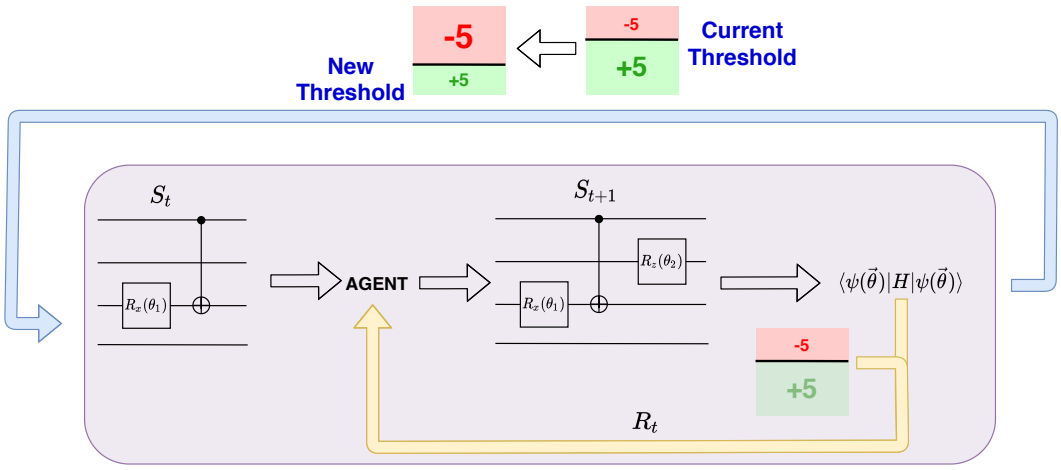


Figure 1: In the proposed method we train an agent to generate quantum circuits with basic quantum gates (rotations and CNOTs) on increasingly harder tasks defined by different thresholds. Once an agent solves the problem for a particular threshold (depicted in a purple box), the threshold is lowered and the agent starts solving harder task with knowledge obtained during the previous one.

manifold accessible by the Ansatz. Thus finding new ways to optimize the architecture could lead to breakthroughs in quantum variational methods for chemistry (e.g. for strongly-correlated systems, for which standard Ansätze might fail), but also in other domains which utilize variational circuits such as machine learning and optimization [2, 13]. Currently, the the foremost Ansätze fall primarily in two classes: chemistry-inspired (e.g. the unitary coupled-cluster Ansatz [8, 12]) and hardware inspired (e.g. the hardware efficient Ansatz in [5]). Architectures from both of these classes entail using a fixed architecture [8, 12, 5, 1] determining the unitary $U(\vec{\theta})$, and hence the corresponding Ansatz circuit which can be decomposed into two-qubit CNOT and one-qubit rotation gates parametrized by $(\theta_1, \theta_2, \ldots, \theta_n) \in [0, 2\pi]^n$ to be optimized by a classical subroutine. However, the architecture itself can also be optimized. This results in a hard structure optimization problem, as it is a combinatorial optimization problem which must balance between the expressivity of the Ansatz (guaranteeing a good approximation of the ground state energy), and an enlarged search space stemming from the depth and size of the circuit (which is incompatible with NISQ restrictions).

In this work we propose a general optimization procedure based on reinforcement learning (RL) which, despite the issues above, not only finds the correct parameters but, by the very nature of the RL approach, also results in an architecture which is both gate and depth efficient. To minimize the computational burden of training, the algorithm is trained with an approximation of minimal energy for a given circuit obtained by local, rather than global, optimization at each step. Furthermore, we propose a scheme for generating quantum circuits using so-called *curriculum learning*, the details of which are discussed in Sec. 2.

We find that utilizing the RL paradigm offers shorter, more NISQ-friendly circuits than fixed-architecture Ansätze such as UCCSD, providing good ground state energy estimates, with curriculum learning having a further edge over the basic RL approach. Thus our work suggests that RL may provide novel avenues for finding structure-optimized, NISQ-friendly Ansätze for quantum chemistry.

## 2 Methods

In this work, we express the quantum circuit design for ground state energy estimation problems as a reinforcement learning task environment. In reinforcement learning (RL), [10] a so-called learning agent (or algorithm) learns by interacting with an environment. Interactions alternate between taking actions $a \in A$ and receiving feedback in form of states $s \in S$ and rewards $r \in \mathbb{R}$. In the quantum circuit design environment, the actions are all the possible placings of a single-qubit gate (X-, Y-, or Z-rotations) and a two-qubit CNOT gate added as the next layer on the qubit wires. The state is a representation of the current circuit.

One time step corresponds to a single interaction cycle between the agent choosing gates and the environment issuing a reward, which depends on the estimate of the ground state energy obtained from the resulting circuit. This estimate of the ground energy is obtained after independently optimizing the angle of the last applied gate by the coordinate gradient descent algorithm Rotosolve [7]. An episode starts with an empty circuit and ends either when the goal condition is met or the maximum number of time steps is reached. Note, since global optimization is not performed to save resources, this means the agent is learning from imperfect data. To define a goal condition we set a hypothesized energy threshold, as soon as the agent reaches this threshold it receives a reward of +5. If the agent fails to reach the threshold within the required number of time steps, a reward of -5 is issued. Otherwise, a reward proportional to the difference between the previous and the current energy is given. The goal of the agent is to maximize its expected discounted sum of future rewards $\mathbb{E}(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1})$ with respect to the chosen discount factor $\gamma$. The agent's actions are governed by the conditional probability distribution $\pi(a|o)$, its policy. We employ a Double Deep-Q network (DDQN) [11] with an $\epsilon$-greedy policy. The energy threshold chosen is 0.001 Hartee, approximately the "chemical accuracy", which stems from typical errors encountered in thermochemical experiments (and hence ideally one would improve on this accuracy).

To obtain energies closer to chemical precision, we require larger quantum circuits. Larger quantum circuits are equivalent to a larger explorable state space which increases the difficulty of the learning problem. While this presents a formidable challenge, it also provides us with well-defined levels of difficulty that can be leveraged for curriculum learning [3], a variant of transfer learning. We employ a curriculum learning approach, where a DDQN is trained in the same environment in multiple rounds with varying complexity, as depicted in Figure 1. Each round consists of thousands of episodes and for each new round the energy threshold is lowered, which also increases the difficulty of the task.

## 3 Results

To benchmark the proposed method we carried out a series of experiments. All of the experiments consisted of 120 000 episodes, a cycle starting from an empty circuit, terminating when a limit of gates (40) is used or if the target energy is reached. All experiments are performed on the problem of finding the ground state of LiH with bond distance 1. In RL the discount factor is set to $\gamma = 0.93$, probability of random action in $\epsilon$-greedy policy is decayed by a factor $0.99995$ up to minimal value $\epsilon = 5\%$. The target network in the DQN training procedure is updated after each 500 actions. After each training episode, we included a testing phase where probability of random actions is set to $\epsilon = 0$ and experience replay procedure is turn off. Values from this test episodes are presented below.

To improve our method we propose curriculum agent (CA) training in which the agent learns in an increasingly more challenging environment. More specifically, the first threshold was set to a distance of 0.005 Hartree from the ground energy. When the behaviour of an agent becomes stable in terms of successfully solved episodes, which in our case was in less than 60 000 episodes we lower the threshold to 0.003 Hartree. In this new threshold setting, we initialize the agent's neural network with weights obtained from training on previous threshold. Additionally experiences from previous task are kept in replay buffer, with epsilon greedy parameter set to $\epsilon = 5\%$ which is a minimal value down to which we decay probability of random action. To ensure a fair comparison of all methods we run training procedure on both thresholds for 60 000 episodes.

To compare we set up two baseline methods: a random agent (RA) which is a sanity check for the reasonability of reinforcement learning methods and tabula-rasa agent (TR) to validate the reasonability of curriculum agent. In the RA setting, circuits are generated by randomly selecting gates for each layer terminating when a limit of gates (40) is used or if the target energy is reached. Tabula-rasa agent is trained from scratch on the threshold 0.003 Hartree. We compare abovementioned methods on circuits that exceeds chemical precision after one Rotosolve update with respect to minimum and average depth, number of gates and number of CX gates.

In Table 1 one can see that curriculum agent (CA) achieves best results in all three statistics. Note that CA generates at least twice as many unique circuits as other methods. What is more important, proposed circuits are significantly shorter in terms of all examined criteria.

In the next experiment we examine how well the proposed architectures perform under global optimization. We compare our method with previously introduced baselines and with standard VQE approaches i.e. hardware efficient [5] and UCCSD Ansätze [8, 12]. On each circuit examined in

Table 1: Results from unique sequences of actions which allowed to pass the threshold of 0.003. One full Rotosolve cycle was run, resulting in some circuits exceeding chemical precision. In the table below, we report the number of circuits which exceeded 0.001, average number of gates, minimal number of gates, average depth, minimal depth, average number of CX gates and minimal number of CX gates in circuits which exceeded chemical precision. First three columns of TR and ME experiments were run over 10 seeds, and the average over solved trials is reported. In the CA and TR experiment, the agent achieved threshold 0.003 in 7 out of 10 trials, i.e. independent runs. Bold results are the best.

| | #(0.1%) | avg #gates | min #gates | avg depth | min depth | avg #CX | min #CX |
|---|---|---|---|---|---|---|---|
| RA | 5 | 29.4 | 26 | 17.4 | 13 | 13.8 | 12 |
| TR | 10.5 | 30.3 | 23 | 21.38 | 13 | 22.72 | 13 |
| CA | **24.28** | **20.21** | **13** | **11.21** | **8** | **10.41** | **6** |

Table 2: Comparison of different architectures with respect to distance obtained after global optimization procedure. Third column presents number average number of unique circuits that achieved chemical precision after global optimization. Bold results are the best.

| | avg distance | min (dist) | #(0.1%) | avg #gates | min #gates | min depth |
|---|---|---|---|---|---|---|
| RA | 0.00041 | 0.00009 | 5 | 29.4 | 26 | 13 |
| HE | 0.00239 | 0.00230 | N.A. | 33 | 33 | 12 |
| UCCSD | **0.00038** | 0.00038 | N.A. | 430 | 430 | 430 |
| TR | 0.00049 | 0.00013 | 129.71 | 30.68 | 23 | 13 |
| CA | 0.00043 | **0.00007** | **846.29** | **16.21** | **13** | **6** |

previous experiments, we run the Constrained Optimization By Linear Approximation (COBYLA) optimizer with maximal number of iterations set to 1000. UCCSD consists of 430 gates, while hardware efficient (HE) consists of 33 gates i.e. three layers.

As one can see in Table 2 curriculum agents also provide competitive architectures in terms of getting as close to the ground state energy as possible. Note that only RA and UCCSD achieves lower average energies after global optimization than CA, however both of them require significantly more gates.

In the last experiment we investigate how winning action sequences from task with threshold 0.005 (left Fig. 2) correspond to winning action sequences from tasks with threshold 0.003 (right Fig. 2). Fig. 2 shows why curriculum learning is a powerful technique in this problem. The agent do reuse previously learned knowledge on 0.005 threshold and starts circuits with the same combination of rotation and CNOT gates. As one can see at the left Fig. 2 agent successfully learns to use $R_y$ gate on third qubit at the beginning of the circuit, instead of CNOTs which would have no effect on the energy of a particular circuit. Then agent combine this rotation gates with CNOT gates for suitable qubits, which is beneficial in terms of energy. Intrestingly, during the training with threshold 0.005 the 3rd gate was redundant due to the nature of Rotosolve algorithm and during the second task (0.003 threshold) agent unlearnt this behaviour (third slot on the circuits). In the second task most action are taken repeatedly, which can suggest that agent shifted from exploration toward exploitation quickly thanks to the knowledge obtained on previous tasks.

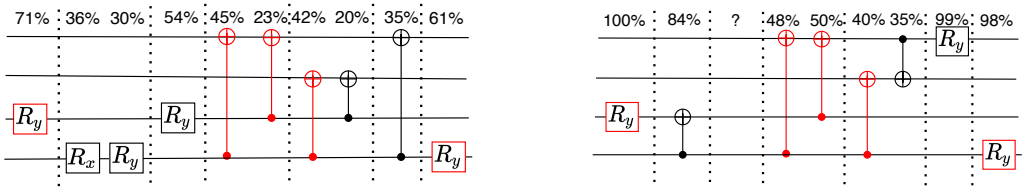


Figure 2: Comparison of the dominating gates in the initial seven actions for different thresholds. Circuits on the left and right side correspond to tasks with threshold 0.005 and 0.003 respectively. The frequency of its use is written above each gate. Question mark denotes that all gates are used almost uniformly. The gates that do not change when the threshold is changed are marked in red.

## 4 Conclusion

In this work we presented a novel approach for training the reinforcement learning agents to generate quantum circuits. The curriculum agents are trained on increasingly harder tasks to generate circuits that approximate ground state energies with lower error. The resulting circuits are short in terms of quantum gates and hence noise robust resulting in the shortest circuits obtaining chemical precision for this task. Importantly, the proposed method has a unique feature that allows to approach arbitrary energy level with an increasingly higher precision. This feature allows us to bypass the need for good estimates of ground state energy which we show improve performance. Additionally, the objective function of an agent can be further modified to explicitly promote shorter circuits e.g. number of CNOT gates can be used to penalize an agent. Last but not least, curriculum learning can be naturally extended beyond a single problem defined by a particular molecule configuration and knowledge obtained during training can be transferred to an agent solving different task.

## Broader Impact

Quantum computers may offer significant improvements in chemistry of the future, with applications in drug and materials design which could have widespread positive consequences in society (e.g. in more effective and cheaper medicine).

Our work presents novel approaches for enhancing VQE-based methods targeting quantum chemistry problems, and thus contributes to this objective. In particular, this research focuses on the use of reinforcement learning to automatically program existing quantum devices. Whilst our work mainly focuses on finding the LiH molecule ground state energy, the benefits of such a solution extends to research questions that can be reformulated as a VQE optimization problem.

We foresee no negative impact stemming from our research, no significant consequences from system failures, nor to we believe our methods leverage any bias in any data.

## Acknowledgments and Disclosure of Funding

MO acknowledge the support of the Foundation for Polish Science (FNP) under grant number POIR.04.04.00-00-17C1/18-00. LMT acknowledges support by the Austrian Science Fund FWF within the DK-ALM (W1259-N27). This work was also supported by the Dutch Research Council(NWO/OCW), as part of the Quantum Software Consortium programme (project number 024.003.037). VD and ES acknowledges the support of SURFsara through the QC4QC project. This research was partially funded by the Grant of Priority Research Domain at Warsaw University of Technology - Artificial Intelligence and Robotics.

## References

[1] Marcello Benedetti, Mattia Fiorentini, and Michael Lubasch. Hardware-efficient variational quantum algorithms for time evolution, 2020.

[2] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, nov 2019.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ACM International Conference Proceeding Series*, volume 382, pages 1–8, New York, New York, USA, 2009. ACM Press.

[4] Harper R. Grimsley, Sophia E. Economou, Edwin Barnes, and Nicholas J. Mayhall. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nature Communications*, 10(1):3007, Jul 2019.

[5] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, Sep 2017.

[6] Ho Lun Tang, V. O. Shkolnikov, George S. Barron, Harper R. Grimsley, Nicholas J. Mayhall, Edwin Barnes, and Sophia E. Economou. qubit-ADAPT-VQE: An adaptive algorithm for constructing hardware-efficient ansatze on a quantum processor. *arXiv e-prints*, page arXiv:1911.10205, November 2019.

[7] Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Quantum circuit structure learning. *arXiv preprint arXiv:1905.09692*, 2019.

[8] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5:4213, 2014.

[9] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018.

[10] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

[11] Hado van Hasselt, Arthur Guez, and David Silver. Deep Reinforcement Learning with Double Q-learning. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 2094–2100, sep 2015.

[12] M.-H. Yung, J. Casanova, A. Mezzacapo, J. McClean, L. Lamata, A. Aspuru-Guzik, and E. Solano. From transistor to trapped-ion computers for quantum chemistry. *Scientific Reports*, 4(1):3589, Jan 2014.

[13] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D. Lukin. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Phys. Rev. X*, 10:021067, Jun 2020.