Quantum Wasserstein Compilation: Unitary Compilation using the Quantum Earth Mover's Distance

Marvin Richter*
Department of Microtechnology and Nanoscience
Chalmers University of Technology
Gothenburg, Sweden

Abhishek Y. Dubey
Fraunhofer IIS
Nuremberg, Germany

Axel Plinge
Fraunhofer IIS
Nuremberg, Germany

Christopher Mutschler
Fraunhofer IIS
Nuremberg, Germany

Daniel D. Scherer
Fraunhofer IIS
Nuremberg, Germany

Michael J. Hartmann
Friedrich-Alexander University Erlangen-Nürnberg (FAU)
Erlangen, Germany

Abstract

Quantum circuit compilation (QCC) is an essential preproccessing step for quantum algorithm execution. Key challenges include translation into hardware-specific gates, reduction of circuit depth and adaptation to platform-specific noise. Variational quantum circuit compilation (VQCC) optimizes the parameters of an ansatz according to the goal of reproducing a given unitary transformation. In this work, we present a VQCC-objective function called the quantum Wasserstein compilation (QWC) cost function based on the quantum Wasserstein distance of order 1. We show that the QWC cost upper bounds the average infidelity of two circuits. An estimation method based on measurements of local Pauli-observables is utilized in a generative adversarial network to learn a given quantum circuit. We compare the efficacy of the QWC cost with other cost functions, such as the Loschmidt echo test (LET) and the Hilbert-Schmidt test (HST). Finally, our experiments demonstrate that QWC as a cost function is least affected by barren plateaus compared to other cost functions.

1 Introduction

Quantum circuit compilation (QCC) involves translating a target quantum algorithm into an executable quantum circuit compatible with real quantum hardware. One approach to QCC is based on the variational quantum computing paradigm, which focuses on optimizing the parameters of a circuit to minimize a cost function. Several cost functions have been developed for this purpose, beginning with the work by Khatri et al. [2019], where the similarity between the target circuit and the ansatz was evaluated directly on the quantum computer. This method allows for bypassing the need for exponentially many resources that arise from the increasing complexity of the Hilbert space of quantum states and their transformations. However, this approach has its drawbacks. The target circuit and the ansatz must be executed on the same joint, entangled quantum system, typically on the same quantum hardware. This means that an already compiled and executable target circuit is required. Additionally, both global and local formulations encounter the barren plateau problem, as demonstrated in Section 4.

*Work done while at Fraunhofer IIS. Corresponding author: marvin.richter@chalmers.se

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

Motivation Until now, variational compilation methods have relied on measures related to the quantum fidelity of quantum states. However, fidelity has two fundamental limitations as a cost function. First, if even a single subsystem has orthogonal local support between the target and the approximate state, the total fidelity vanishes regardless of the similarity in other subsystems. This is a consequence of the fidelity's unitary invariance. Second, the fidelity between two randomly chosen quantum states decreases exponentially with system size. This exponential vanishing of fidelity leads to an exponentially small learning signal, making measurement costs prohibitive.

Therefore, we introduce a cost function for VQCC based on a fundamentally different distance metric: the quantum Wasserstein distance of order 1. Unlike the trace distance or quantum fidelity, this metric is not unitarily invariant. Consequently, it grows linearly with the size of the quantum system (as shown by Kiani et al. [2022]). These promising properties motivate us to formulate a compilation method based on this distance.

2 Preliminaries

Unitary compilation Unitary compilation describes the process of finding a decomposition of a unitary transformation V into a specific set of parameterized unitaries available on the hardware $\{U_i(\theta_i)\}$, i.e.

$$V \overset{!}{\approx} U_1(\theta_1)U_2(\theta_2)\dots U_T(\theta_T) =: U(\boldsymbol{\theta}) \tag{1}$$

with possibly independent parameters θ_i and the number of unitaries T. The unitary compilation process is thereby twofold: a) chose an appropriate ansatz represented by the kind of parameterized unitaries U_i and b) find the optimal parameters. The technique that we developed in this work tackles the problem of finding optimal parameters for a given ansatz $U(\boldsymbol{\theta})$, such that it is close to a given target unitary operator V. Since our goal is to reproduce the complete unitary matrix and thus mimic the target evolution for all possible input states, the average fidelity emerges as a natural measure of closeness.

Definition 1. *Average Fidelity [Nielsen [2002], Khatri et al. [2019]]. Given two unitary transformations U and V, the average fidelity between them is defined as:*

$$\bar{F}(U,V) = \int d\psi |\langle\psi| V^\dagger U |\psi\rangle|^2 \tag{2}$$

Here, $d\psi$ represents the integration over the unitarily invariant Fubini-Study measure on pure states.

The Quantum Wasserstein Distance of Order 1 De Palma et al. [2021] introduced the Wasserstein distance of order 1 (or quantum W_1 distance) as a generalization of the classical Wasserstein distance for probability distributions (also called earth mover's distance) to quantum states. It has an interpretation as a continuous version of a quantum Hamming distance which could be intuitively described as the number of differing qubits. Further intuition is provided in Appendix B. The dual formulation of the quantum W_1 distance in terms of the quantum Lipschitz constant enables its practical use.

Proposition 1. *For two n-qubit quantum states $\rho, \sigma \in \mathcal{D}(\mathcal{H}_n)$, where $\mathcal{D}(\mathcal{H}_n)$ is the set of density operators on the Hilbert space $\mathcal{H}_n$, the quantum W_1 distance admits a dual formulation,*

$$W_1(\rho,\sigma) = \|\rho-\sigma\|_{W_1} := \max(\mathrm{Tr}[H(\rho-\sigma)] : H \in \mathcal{M}_n, ||H||_L \leq 1) \tag{3}$$

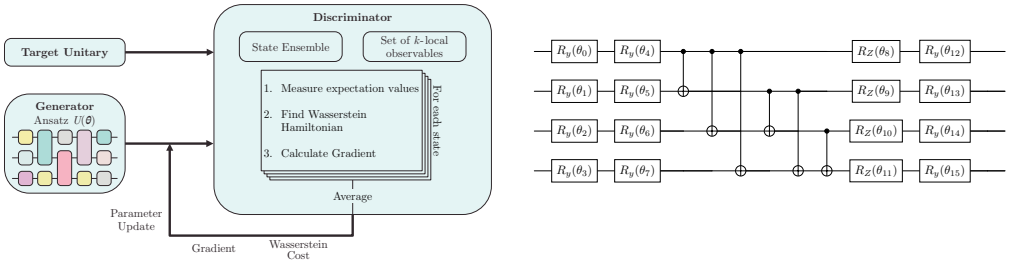
with $\mathcal{M}_n$ being the set of observables on the Hilbert space $\mathcal{H}_n$ and $||\cdot||_L$ the quantum Lipschitz constant as defined in De Palma et al. [2021].

We benchmark our method against other well known metrics for unitary compilation, such as the Hilbert-Schmidt test (HST), the Loschmidt-echo test (LET) and their corresponding local variations. We describe these in detail in Appendix A.

3 Our work

3.1 Ideal Cost

As outlined above, the quantum W_1 distance is a measure for the closeness of two quantum states. We will now extend this distance to measuring the closeness of two unitary operators, U and V, by applying the operators on (pure) quantum states and measuring the pairwise distances.



(a) Overview of the compiling algorithm

(b) HEA with full entanglement

Figure 1: (a) The target unitary and the parameterized circuit acting as the generator are assessed by the discriminator which calculates the Wasserstein compilation cost. The distance estimation requires a state ensemble acting as input states for target and generator and a set of k-local observables whose expectation values are measured. A Wasserstein Hamiltonian can be constructed from the differences of the expectation values and the gradient of the averaged cost can be used for updating the parameters of the generator. (b) A single layer of hardware efficient ansatz (HEA) with full entanglement.

Definition 2 (Quantum Wasserstein Compilation Cost)**.** *Let U, V be unitary operators on the Hilbert space $\mathcal{H}_n$ and $|\psi\rangle$ be a n-qubit quantum state in $\mathcal{H}_n$. Then the quantum Wasserstein compilation distance is defined as ($d\psi$ is the Fubini-Study metric)*

$$C_{QW}(U,V) = \int_{\psi} \mathrm{d}\psi \, W_1^2\Big(U\,|\psi\rangle\,, V\,|\psi\rangle\Big) \tag{4}$$

We chose to define the QWC cost in Eq. (4) as the squared W_1 distance since it then acts directly as an upper bound for the average infidelity:

Proposition 2. *Let U, V be unitary operators on $\mathcal{H}$. Then the following inequality holds between the QWC cost $C_{QW}(U,V)$ and the average fidelity $\overline{F}(U,V)$*

$$C_{QW}(U,V) \geq 1 - \overline{F}(U,V) \tag{5}$$

The proof can be found in Appendix E.2. Proposition 2 establishes the practical significance of C_{QW} for QCC: When used as a minimization objective for the parameters of an ansatz $U(\boldsymbol{\theta})$, it ensures the compiled circuit approximates the target circuit V with bounded average infidelity.

3.2 Empirical Cost

In order to calculate the cost in Eq. (4) we need to first estimate the quantum W_1 distance as defined in Eq. (3). For this, we begin by choosing the observables satisfying the quantum Lipschitz condition. We use the ansatz for H that is a weighted sum of locally acting Pauli observables.

$$H = \sum_m w_m H_m \quad H_m = \bigotimes_{j=1}^{n} \sigma_{P_j}^{(j)} \quad P_j \in \{I, X, Y, Z\} \tag{6}$$

This ansatz has 4^n observables which grows exponentially with the number of qubits. Following Kiani et al. [2022], we can restrict the set of observables $\mathcal{M}_n$ to $\mathcal{M}_n^{(k)}$. We define this as the set of Pauli strings that act non-trivially only on a subset of k qubits, and refer to them as k-local Pauli observables. Using local Pauli operators restricts the growth of the number of Pauli observable to $\mathcal{O}(n^k)$ for $k \ll n$. Thus we instead have the approximation

$$W_1^{(k)} = \max(\mathrm{Tr}[H(\rho - \sigma)] : H \in \mathcal{M}_n^{(k)}, ||H||_L < 1) \tag{7}$$

Moreover, the space of all quantum states is growing exponentially fast for increasing system size and even for small qubit numbers, is prohibitively large. To overcome this hurdle, we use a *state ensemble* $\mathcal{A} = \{|\psi\rangle_s\}$ of finite size and measure the empirical distance restricted to k-local observables:

$$\tilde{C}_{QW}^{(k)}(U,V,\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{\psi \in \mathcal{A}} \Big(W_1^{(k)}(U\,|\psi\rangle\,, V\,|\psi\rangle)\Big)^2 \tag{8}$$

The size of the state ensemble is a hyperparameter of the method. For efficiently implementable unitaries (those with a depth polynomial in n) and cost functions that minimize the expectation value of an observable, Caro et al. [2022] demonstrated that only a polynomial number of training samples is sufficient, meaning that $|\mathcal{A}| \in \mathcal{O}(\text{poly } n)$. This finding challenges the common expectation that an exponentially large dataset would be necessary. Our numerical experiments in Appendix F confirm this theoretical result for QWC.

4 Experiments

In the previous sections, we introduced the empirical quantum Wasserstein compilation cost Eq. (8) with its derivatives for parameterized unitaries in Appendix D, Eq.(12). Based on these, we can formulate a procedure to learn a target unitary V, see Fig. 1a. In all experiments, we use the hardware-efficient ansatz (HEA, Kandala et al. [2017]) (see Fig. 1b), since large-scale implementations for chemistry(Quantum et al. [2020]) and optimization(Harrigan et al. [2021]) applications have shown, this ansatz leads to smaller errors due to hardware noise. Furthermore, our hyperparameter search (see Appendix F) for the training state size $|\mathcal{A}|$ and the locality of Pauli observables k led us to choose $k = \lceil n/2 \rceil$ and $|\mathcal{A}| = 8$ for all the qubits.

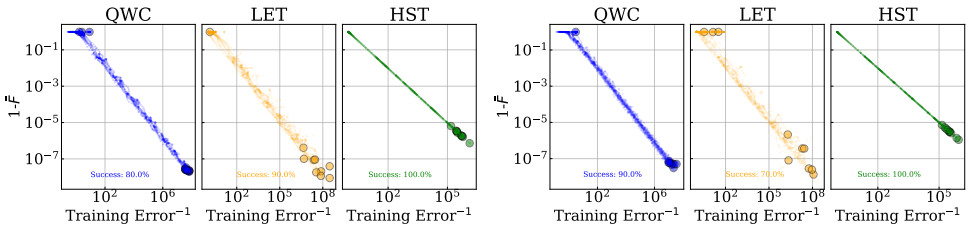
Training The compilation is in the form of a quantum Wasserstein Generative Adversarial Net (qWGAN) inspired from Kiani et al. [2022]. Quantum GAN as a quantum adversarial game was introduced by Lloyd and Weedbrook [2018]. Due to the limited scope of this study, the expressivity of the generator was not explicitly addressed and assumed to be given. The discrimination ability, on the other hand, depends on several factors that were examined in this work. The generator is a variational quantum circuit with parameters $\boldsymbol{\theta}$ which we denote as $G(\boldsymbol{\theta})$ that outputs a state $G(\boldsymbol{\theta})\,|\psi_a\rangle$, and the discriminator is the weighted Hamiltonian from Eq. (6) with k-local Pauli strings.

The first step of every optimization is to measure the expectation values of the Pauli observables $H_m \in \mathcal{M}_n^{(k)}$ for every input state $|\psi_a\rangle \in \mathcal{A}$ after evolving with the generator ansatz and the target. We denote the evolved set of states as $\{G(\boldsymbol{\theta})\,|\psi_a\rangle\}$ $(\rho(\boldsymbol{\theta}))$ and $\{V\,|\psi_a\rangle\}$ (σ) respectively. The difference in expectation value is given by $c_m = \text{Tr}(\rho(\boldsymbol{\theta})H_m) - \text{Tr}(\sigma H_m)$. The state ensemble is built from products of Haar-random single-qubit states. Caro et al. [2023] proved that this ensemble suffices to learn efficiently implementable targets when average fidelity is used as the cost function, even for entangled states. If the states and the observables are fixed, the result of the target can be cached. We then solve the linear program for the weights w_m

$$\begin{array}{ll} \text{maximize} & \sum_m w_m c_m \\ \text{constraint} & \sum_{m:i\in\mathcal{I}_m} |w_m| \leq 1/2 \quad \forall i \in [n] \end{array} \tag{9}$$

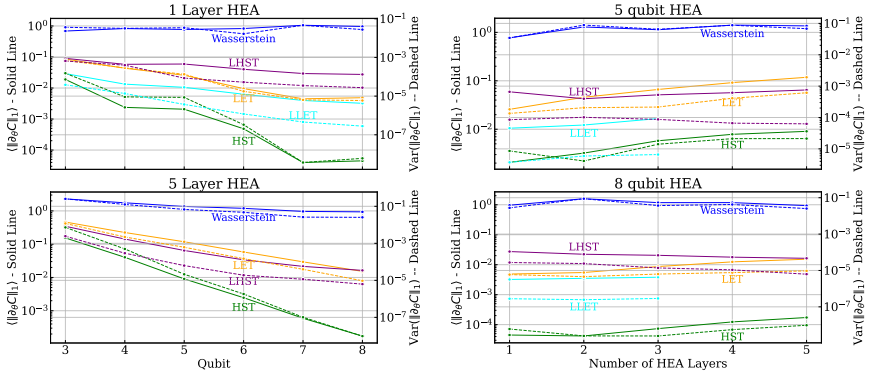
The weights w_i are sparse with only n non-zero entries and the corresponding Pauli operators are called active. The state-wise quantum W_1 distances $W_1^{(k)}$ can be measured from Eq. (7) with the Hamiltonian $H_W = \sum_{n\in\mathcal{N}} w_n^* H_n$ where $\mathcal{N}$ is the set of active Pauli operators and w_n^* are the solutions to the linear program. Finally, the gradients of the state-wise distances (see Eq. (12)) are averaged and used to update the parameters of the generator $G(\boldsymbol{\theta})$. In our experimental setup, we selected the hardware-efficient ansatz (HEA) as our target and ansatz for demonstration. We fix the parameters of the target and randomly choose a different set of parameters for the ansatz. This ensures that at least one solution exists for the compilation problem. Thus, we do not allocate resources towards addressing the issue of expressivity by attempting to learn a diverse target unitary within a given ansatz structure. We plot the infidelity vs. inverse training error for 5- and 8-qubit circuits in Fig. 2 over 10 runs and find that the success rate is comparable to both LET and HST. Further results are shown in Appendix. C.

Effects of barren plateaus To demonstrate that QWC is least affected by barren plateaus in the optimization landscape, we plot the expectation and variance of the l_1- norm of the gradient of the cost function with respect to the parameters of the ansatz as a function of (a) the number of qubits in the circuit and (b) the number of layers in the circuit (See Fig.3). We consider different numbers of layers $(1-5)$ of the HEA for both the target and ansatz. As before the number of layers is identical in both the target and ansatz. We follow the same approach as Kiani et al. [2022] and calculate the gradients at the first optimization step. We see that the gradient norms of LET and HST decrease drastically as the number of qubits increase in both one layer and five layer circuits, indicating that these cost



(a) $n = 5$ (b) $n = 8$

Figure 2: Final infidelity ($1 - \bar{F}$) vs. inverse training error (C_{QW}^{-1}) for hardware efficient ansatz (HEA) with full entanglement for $n = 5, 8$. A run is successful when the cost function is below the threshold of 10^{-3}. QWC reaches very low values with a high probability. We employ early stopping in training whenever the last 100 values of the variance of the cost function reaches 10^{-8}.



(a) Function of number of qubits (b) Function of number of layers

Figure 3: Mean and Variance of l_1-norm of the gradient of the three cost functions. The gradient is taken of the first parameter update step. Each point corresponds to the average over 100 runs.

functions are adversely affected by the barren plateaus. For QWC, we see that for circuits with one layer and five layers, the gradient and the variance saturate as the number of qubits increase. As a function of the number of layers there is no decay in the norms but the absolute values themselves have a difference of orders of magnitude. Thus, we can conclude that QWC is least affected by barren plateaus compared to LET and HST. These results are consistent with the no-go theorems formulated by Cerezo et al. [2021], since the QWC cost function is built from local observables (Appendix A).

5 Conclusion

We introduce the QWC cost function for variational quantum circuit compilation, which computes an ensemble average of the order-1 Wasserstein distance. This metric scales linearly with qubit count while preserving a meaningful learning signal throughout the optimization process. Despite this scaling behavior, it remains suitable as a variational compilation objective since it provides a bound on the average fidelity. The distance is estimated using local products of Pauli operators and their expectation values in a linear program to construct a Hamiltonian. The difference of the Hamiltonian's expectation value represents the discriminator in a quantum GAN. The locality of Pauli observables affects the discriminator's ability, where smaller locality can lead to over-estimation of similarity, with measurement effort scaling as $\mathcal{O}(n^k)$. While the scaling of measurement observables with qubit count presents a limitation, recent developments in classical estimation techniques by Angrisani et al. [2024] offer promising approaches for future improvements. Additionally, the sample state set's size should increase polynomially in n, as suggested by Caro et al. [2022]. Our numerical simulations were noise-free, and future research will address the noise resilience of QWC, expanding on findings by Sharma et al. [2020] regarding HST and LET.

Acknowledgments and Disclosure of Funding

The research is part of the Munich Quantum Valley (MQV) and was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy with funds from the Hightech Agenda Bayern via the project BayQS.

References

A. Angrisani, A. Schmidhuber, M. S. Rudolph, M. Cerezo, Z. Holmes, and H.-Y. Huang. Classically estimating observables of noiseless quantum circuits. *arXiv preprint arXiv:2409.01706*, 2024.

M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13(1):4919, Aug. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32550-3.

M. C. Caro, H.-Y. Huang, N. Ezzell, J. Gibbs, A. T. Sornborger, L. Cincio, P. J. Coles, and Z. Holmes. Out-of-distribution generalization for learning quantum dynamics. *Nature Communications*, 14(1): 3751, July 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39381-w.

M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1):1791, Mar. 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21728-w.

G. De Palma, M. Marvian, D. Trevisan, and S. Lloyd. The Quantum Wasserstein Distance of Order 1. *IEEE Transactions on Information Theory*, 67(10):6627–6643, Oct. 2021. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2021.3076442.

M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya, J. Atalaya, J. C. Bardin, R. Barends, S. Boixo, et al. Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, 17(3):332–336, 2021.

A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, B. R. Johnson, and J. M. Gambetta. Quantum computing with Qiskit, 2024.

A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *nature*, 549(7671):242–246, 2017.

S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles. Quantum-assisted quantum compiling. *Quantum*, 3:140, May 2019. ISSN 2521-327X. doi: 10.22331/q-2019-05-13-140.

B. T. Kiani, G. De Palma, M. Marvian, Z.-W. Liu, and S. Lloyd. Learning quantum data with the quantum earth mover's distance. *Quantum Science and Technology*, 7(4):045002, Oct. 2022. ISSN 2058-9565. doi: 10.1088/2058-9565/ac79c9.

S. Lloyd and C. Weedbrook. Quantum Generative Adversarial Learning. *Physical Review Letters*, 121(4):040502, July 2018. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.121.040502.

N. Meyer, C. Ufrecht, M. Periyasamy, A. Plinge, C. Mutschler, D. D. Scherer, and A. Maier. Qiskit-torch-module: Fast prototyping of quantum neural networks. *arXiv:2404.06314*, 2024. doi: 10.48550/arXiv.2404.06314.

M. A. Nielsen. A simple formula for the average gate fidelity of a quantum dynamical operation. *Physics Letters A*, 303(4):249–252, Oct. 2002. ISSN 03759601. doi: 10.1016/S0375-9601(02)01272-0.

X. Qiu, L. Chen, and L.-J. Zhao. Quantum wasserstein distance between unitary operations. *Physical Review A*, 110(1):012412, 2024.

G. A. Quantum, Collaborators*†, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, S. Boixo, M. Broughton, B. B. Buckley, et al. Hartree-fock on a superconducting qubit quantum computer. *Science*, 369(6507):1084–1089, 2020.

M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran. Evaluating analytic gradients on quantum hardware. *arXiv:1811.11184 [quant-ph]*, Nov. 2018. doi: 10.1103/PhysRevA.99.032331.

K. Sharma, S. Khatri, M. Cerezo, and P. J. Coles. Noise resilience of variational quantum compiling. *New Journal of Physics*, 22(4):043006, Apr. 2020. ISSN 1367-2630. doi: 10.1088/1367-2630/ab784c.

A Related Work

There has been considerable work in the field of variational quantum algorithms. Still, here we would like to highlight one of the studies by Cerezo et al. [2021], which establishes some conditions on the cost function under which the cost function is affected by barren plateaus. In particular, for a variational circuit approximating a 2-design locally, they show that defining the cost in terms of global observables leads to exponentially vanishing gradients even when the ansatz is shallow. On the other hand, defining the cost with local observables leads to, at worst, a polynomially vanishing gradient, as long as the depth of the ansatz is $\mathcal{O}(\log n)$. These results constrain the cost functions for training variational circuits to avoid barren plateaus.

Other work on distances of unitaries based on Wasserstein distance Another distance for unitaries based on the Wasserstein-1 distance between quantum states was proposed in Ref. Qiu et al. [2024]. Their formulation defines the distance between unitary operations by taking the maximum deviation of their effects on all possible quantum states, inspired by approaches to unitary discrimination. While this metric provides insights into local distinguishability of operations and circuit complexity, the maximization over all states makes it computationally challenging. In contrast, our approach uses the average distance, making it more amenable to practical computations while maintaining physical significance.

Cost functions for Variational compiling The Hilbert-Schmidt Test was introduced by Khatri et al. [2019]. The cost function for two unitaries U and V is defined as

$$C_{\text{HST}} = 1 - |\text{Tr}(V^\dagger U)|^2/4^n \tag{10}$$

with n as the number of qubits. C_{HST} is closely related to the average fidelity.

Another approach uses the Loschmidt echo test, which evaluates the fidelity over a set of probe states $\mathcal{A} = \{|\psi\rangle\}$. The corresponding cost function reads as

$$C_{\text{LET}} = 1 - \frac{1}{|\mathcal{A}|}\sum_{\psi\in\mathcal{A}} |\langle\psi|V^\dagger U|\psi\rangle|^2 \tag{11}$$

Further details on this method can be found in Sharma et al. [2020]. Both HST and LET as described above require measuring all the n qubits and the cost functions suffer from a vanishing gradient problem if evaluated on a quantum computer. To address this, local HST (LHST) and local LET (LLET) were introduced. Here, instead of measuring all $2n$ qubits in case of HST and n qubits in case of LET, two qubits and one qubit respectively are measured at a time, and a mean is taken over all possible measurements.

B Quantum Wasserstein Distance of Order 1 for VQCC

In the context of VQCC, the quantum W_1 distance has an intriguing property of not being unitarily invariant. While this might not seem advantageous, it makes the quantum W_1 distance fundamentally different from better-known quantum state metrics like the trace distance or the quantum fidelity. As Kiani et al. [2022] pointed out, this property facilitates the learning of quantum states: consider we want to learn and reproduce a state $|\text{GHZ}_2\rangle\,|1\rangle$ and our initial state is $|000\rangle$. If we change the state during learning from the initial state to $|\text{GHZ}_2\rangle\,|0\rangle$ then this significant improvement towards the target should be admitted by the cost function. No unitarily invariant distance can discriminate between the three pair-wise orthogonal states and hence indicate the improvement. Additionally, this distance is additive rather than multiplicative with respect to subsystems, preventing any single subsystem from dominating the distance.

(a) $n = 3$ (b) $n = 4$

(c) $n = 6$ (d) $n = 7$

Figure 4: Final infidelity ($1 - \bar{F}$) vs. inverse training error (C_{QW}^{-1}) for hardware efficient ansatz (HEA) with full entanglement. (a),(b) The training is carried out for 1000 steps. A run is successful when the cost function is below the threshold of 10^{-3}. QWC reaches extremely low values with a high probability. (c),(d) Since most applications do not require infidelity values of order 10^{-15}, we employ early stopping in training when the variance of last 100 cost values reaches 10^{-8}.

C Training results

We show the infidelity vs. inverse training error C_{QW}^{-1} for 3- and 4-qubit circuits in Fig. 4a, 4b. We let the training run for 1000 steps and see that our cost function can reach values down to 10^{-16} in infidelity, which is comparable to both LET and HST. We also show the infidelity vs. inverse training error C_{QW}^{-1} for $n = 6, 7$ qubit circuits in Fig. 4c, 4d. Training is carried out with early-stopping. The early-stopping condition is invoked whenever the variance of the cost function in the last 100 steps is less than 10^{-8}. Both LET and HST reach convergence faster than our cost function and they also have higher success rates compared to our method.

D Empirical Cost

The choice and size of the set of probe states $\mathcal{A}$ are decisive for the practical use of $\tilde{C}_{QW}^{(k)}$ as an optimization objective in VQCC. In the limit of infinitely many states that are sampled according to the Fubini-Study metric and no restricting on the locality of Pauli operators, the empirical quantum Wasserstein compilation distance becomes equivalent to the ideal distance from Eq. (4). The derivatives of the cost function with respect to a parameter $\theta \in \boldsymbol{\theta}$ can be directly calculated from the respective derivative of the EM distance (Kiani et al. [2022]), i.e. around a value t:

$$\left(\frac{\partial}{\partial\theta}\tilde{C}_{QW}^{(k)}(U(\theta),V,\mathcal{A})\right)_t = \sum_{|\psi_a\rangle\in\mathcal{A}} \frac{2W_1^{(k)}}{|\mathcal{A}|}\left(|\psi_a(t)\rangle, V|\psi_a\rangle\right)\cdot\left(\frac{\partial}{\partial\theta}W_1^{(k)}\left(|\psi_a(\theta)\rangle, V|\psi_a\rangle\right)\right)_t \tag{12}$$

where we use the short-hand notation $|\psi_a(\theta)\rangle = U(\theta)|\psi_a\rangle$.

The term $\left(\frac{\partial}{\partial\theta}W_1^{(k)}\left(U(\theta)|\psi\rangle, V|\psi\rangle\right)\right)_t$ in the derivative can be evaluated using standard techniques like the parameter-shift rule introduced by Schuld et al. [2018]. Since we now have the cost function and its gradients, the only missing building block for learning unitaries is the choice of the state ensemble.

E Deferred Proofs

E.1 Motivation of Definition 2: Quantum Wasserstein distance upper bounds infidelity

The starting point for our derivation is Proposition 2 of [De Palma et al., 2021] that gives upper and lower bounds for the quantum W_1 norm in terms of the trace norm $\|\cdot\|_1$.

$$\frac{1}{2}\|\rho-\sigma\|_1 \leq \|\rho-\sigma\|_{W_1} \leq \frac{n}{2}\|\rho-\sigma\|_1 \tag{13}$$

Additionally, the trace norm is bounded by the fidelity $F(\rho,\sigma)$:

$$1-\sqrt{F(\rho,\sigma)} \leq \frac{1}{2}\|\rho-\sigma\|_1 \leq \sqrt{1-F(\rho,\sigma)}. \tag{14}$$

Hence, we can find a lower bound for the fidelity in terms of the quantum W_1 norm:

$$1-\|\rho-\sigma\|_{W_1} \leq \sqrt{F(\rho,\sigma)}. \tag{15}$$

Since the fidelity is bounded, $0 \leq F(\rho,\sigma)\,\forall\,\rho,\sigma \in \mathcal{S}(\mathcal{H})$, the same holds for $\sqrt{F(\rho,\sigma)}$. We will now constrain the quantum W_1 norm to small values, $0 \leq \|\rho-\sigma\|_{W_1} \leq 1$. This domain is of particular interest as we formulate the VQC problem as a minimization of the quantum W_1 norm. With this constraint, we can square the inequality and make use of Bernoulli's inequality:

$$F(\rho,\sigma) \geq (1-\|\rho-\sigma\|_{W_1})^2 \geq 1-2\|\rho-\sigma\|_{W_1}. \tag{16}$$

By this bound, we now know that a vanishing W_1 distance between two mixed states translates to high fidelity of the states. But this result for mixed states only holds for small distances, e.g. $\|\rho-\sigma\|_{W_1} \leq 1$.

A more general result can be found for pure states. Note that quantum Wasserstein compilation actually uses pure states. For two pure states $\rho = |\psi\rangle\langle\psi|, \sigma = |\phi\rangle\langle\phi|$, the following equality between trace norm and fidelity $F(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$ holds:

$$\big\| |\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| \big\|_1 = \sqrt{1-F(|\psi\rangle, |\phi\rangle)}. \tag{17}$$

Using again Eq. (13), we bound the fidelity by the quantum W_1 norm

$$\big\| |\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| \big\|_{W_1} \geq \sqrt{1-F(|\psi\rangle, |\phi\rangle)} \tag{18}$$

and square without further constraints

$$\big\| |\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| \big\|_{W_1}^2 \geq 1-F(|\psi\rangle, |\phi\rangle). \tag{19}$$

This upper bound for the infidelity of pure states in terms of the quantum W_1 norm motivates our definition of the Wasserstein compilation cost as the squared W_1 distance.

E.2 Proof of Proposition 2

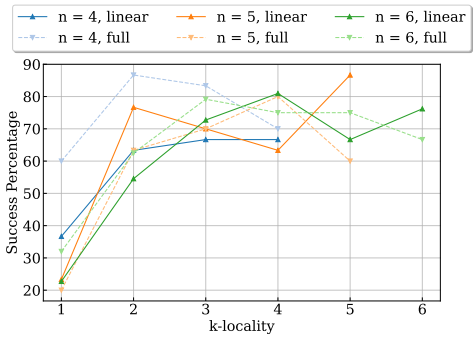
Proposition 2. *Let U, V be unitary operators on $\mathcal{H}$. Then the following inequality holds between the QWC cost $C_{QW}(U,V)$ and the average fidelity $\overline{F}(U,V)$*

$$C_{QW}(U,V) \geq 1-\overline{F}(U,V). \tag{20}$$

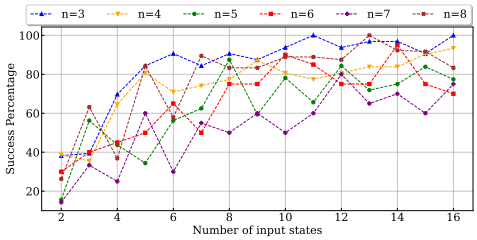
Proof. We use that the quantum W_1 norm is an upper bound for the infidelity that we derive in the Appendix E.1, Eq. (19). Starting from the definition of the QWC cost in Eq. (4), we can directly upper bound the average fidelity:

$$\begin{aligned} C_{QW}(U,V) &= \int_\psi \mathrm{d}\psi\, W_1^2\Big(U|\psi\rangle, V|\psi\rangle\Big) && (21)\\ &\geq \int_\psi \mathrm{d}\psi\, \left(1-F(U|\psi\rangle, V|\psi\rangle)\right) && (22)\\ &= 1-\overline{F}(U,V). && (23) \end{aligned}$$

□



(a) Success percentage out of a total of 30 runs for different k-locality.



(b) Success percentage out of a total of 10 runs for different number of states used as input.

Figure 5: Experimental results for determining the k-locality and the amount of data (number of input states) for successful compilation (a) The number of k-local Pauli observables required to distinguish between the different types of entanglement. We take the 4-,5-, and 6-qubit single layer HEA with linear and full entanglement and run the compilation routine for each $k \in \{1, .., n\}$, where n is the number of qubits under consideration, with 30 experiments each. The solid line shows the trend for linear entanglement, and the dashed line for full entanglement. (b) We fix $k = \lceil n/2 \rceil$ and use single layer HEA with linear entanglement. Successful compilation for any n-qubit can be achieved by using a number of states which gives the highest success probability according to the plot.

F Training Details

Hyperparameters In our experimental setup, the primary goal is to showcase the viability of our chosen approach. We specifically selected the hardware-efficient ansatz (HEA) as our target and ansatz for demonstration. We fix the parameters of the target and randomly choose a different set of parameters for the ansatz. This ensures that at least one solution exists for the compilation problem. Additionally, we compare two distinct entanglement procedures to assess the amount of Pauli data necessary for the learning process. In all experiments, we utilized the ADAM optimizer with a learning rate of 0.1 for QWC and 0.04 for LET(HST), and exponential decay rates for the first and second moment estimates set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively.

We also define successful compilation in terms of the cost function, whenever the cost function is below 10^{-3}. We found from our initial experiments that using a fixed set of states already gives successful training curves. For the discriminator, we mentioned that the expectation value of the Hamiltonian Eq. (6) needs to be evaluated for a k-local Pauli string. k is another hyper-parameter which needs to be tuned according to the problem. We show in Fig. 5a the success percentage over 30 experiments of compilation of a $4, 5$ and 6-qubit single layer HEA target ansatz pair, against the k-locality used to detect the entanglement in the target for two cases, linear and full entanglement. We see a general trend of higher k having higher success probability. Yet, a larger k also means many observables for computation. We choose to scale k with n as $k = \lceil n/2 \rceil$. After choosing the k-locality for the discriminator and choosing a fixed state set $\mathcal{A}$, we conducted experiments to determine the number of states needed to achieve successful compilation. For number of qubits $n \in \{3, ..., 8\}$ we ran the training for $|\mathcal{A}| \in \{2, ..., 16\}$ and calculated the fraction of runs which were successful out of a total of 10 runs for each state. We show the results in Fig. 5b. Our hyper-parameter search for determining the balance between optimal number of states required and the computational resources led us to use $|\mathcal{A}| = 8$ for the rest of the experiments.

F.1 Computation Details

We make use of Qiskit v1.0 Javadi-Abhari et al. [2024], qiskit-aer v0.13.3, qiskit-algorithms v0.3 and qiskit-torch-module v0.1 Meyer et al. [2024] with Python 3.10 for all our simulations. The hardware leverages AMD Ryzen Threadripper PRO 5965WX 24-Cores with 2 threads per core. The simulations make use of parallel processing of 8 cores by distributing the compilation for each of the $|\mathcal{A}|$ states.