

VQE-generated quantum circuit dataset for machine learning

Akimoto Nakayama^{1,*}, Kosuke Mitarai^{1,2,†}, Leonardo Placidi^{1,2}, Takanori Sugimoto^{2,3,4}, and Keisuke Fujii^{1,2,5,‡}¹Graduate School of Engineering Science, *Osaka University*, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan²Center for Quantum Information and Quantum Biology, *Osaka University*, 1-2 Machikaneyama, Toyonaka 560-0043, Japan³Computational Materials Science Research Team, *RIKEN Center for Computational Science (R-CCS)*, Kobe, Hyogo, 650-0047, Japan⁴Advanced Science Research Center, *Japan Atomic Energy Agency*, Tokai, Ibaraki 319-1195, Japan⁵*RIKEN Center for Quantum Computing (RQC)*, Hirosawa 2-1, Wako, Saitama 351-0198, Japan

(Received 15 June 2023; accepted 17 June 2025; published 11 July 2025)

Quantum machine learning has the potential to computationally outperform classical machine learning, but it is not yet clear whether it will actually be valuable for practical problems. While some artificial scenarios have shown that certain quantum machine learning techniques may be advantageous compared to their classical counterpart, evidence does not yet suggest that quantum machine learning has surpassed conventional approaches in dealing with standard classical datasets, such as the MNIST dataset. In contrast, dealing with quantum data, such as quantum states or circuits, may be the task where we can benefit from quantum methods. Therefore, it is important to develop practically meaningful quantum datasets for which we expect quantum methods to be superior. In this paper, we propose a machine learning task that is likely to soon arise in the real world: clustering and classification of quantum circuits. We provide a dataset of quantum circuits optimized by the variational quantum eigensolver. We utilized six common types of Hamiltonians in condensed matter physics, with a range of 4–20 qubits, and applied ten different ansatz with varying depths (ranging from 3 to 32) to generate a quantum circuit dataset of six distinct classes, each containing 300 samples. We show that this dataset can be easily learned using quantum methods. In particular, we demonstrate a successful classification of our dataset using real four-qubit devices available through IBMQ. By providing a setting and an elementary dataset where quantum machine learning is expected to be beneficial, we hope to encourage and ease the advancement of the field.

DOI: [10.1103/c43x-9866](https://doi.org/10.1103/c43x-9866)

I. INTRODUCTION

Quantum machine learning has attracted much attention in recent years as a promising application of quantum computers [1,2]. Many techniques, such as quantum neural networks [3–5], quantum generative models [6,7], quantum kernel methods [8], and so on have been developed for achieving possible quantum speedups in machine learning tasks. They have also been realized experimentally [7–11]. While some artificial, carefully designed scenarios have demonstrated that certain quantum machine learning techniques may be advantageous compared to classical methods [12–16], it is not yet clear whether quantum techniques would be beneficial for practical applications.

In traditional machine learning, standard datasets, such as MNIST handwritten digits [17], are used to evaluate the performance and thus the practicality of new models. As of now, there is no evidence to suggest that quantum machine learning methods have surpassed the state-of-the-art classical

machine learning techniques on these datasets. On the other hand, it has been demonstrated that classical machine learning can efficiently learn from these datasets. For example, with a large-scale numerical experiment involving up to 30 qubits, Huang *et al.* [13] have shown that the Fashion-MNIST dataset [18] is better learned by classical models. In contrast, when working with “quantum data,” such as quantum states or circuits, there is a good reason to believe that quantum computers may provide a significant advantage. In another work by Huang *et al.* [19], it has been rigorously shown that quantum machine learning is beneficial when learning unknown quantum states or processes provided from physical experiments. It is therefore important to develop a practical quantum dataset in which we can expect quantum methods to be superior.

Several works have made efforts in this direction. Schatski *et al.* [20] proposed a dataset consisting of parametrized quantum circuits with different structure whose parameters are optimized to give output states with certain values of entanglement. However, this dataset lacks data involving higher qubit counts, which is crucial for evaluating scalable quantum machine learning methods. Also, Huang *et al.* [13] proposed to relabel a classical dataset by outputs of quantum circuits so that the relabeled one is difficult to be learned by classical computers. These examples, while giving datasets that are possibly hard to learn by classical computers, are not plausible “real-world” quantum data. Consequently, it remains unclear how insights or results from these datasets could translate

*Contact author: akimoto.nakayama@gmail.com†Contact author: mitarai.kosuke.es@osaka-u.ac.jp‡Contact author: fujii@qc.ee.es.osaka-u.ac.jp

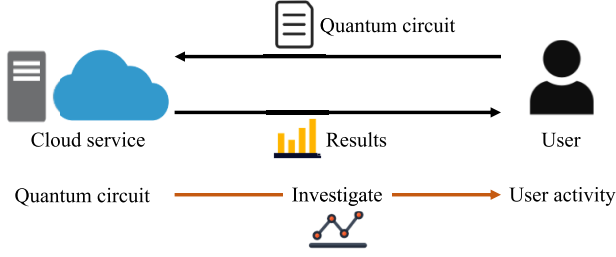


FIG. 1. Scenario of the proposed machine learning task. Cloud quantum computer providers get descriptions of quantum circuits constantly from their users and return execution results to them. The providers wish to investigate the user activities, e.g., whether two users are interested in solving a similar computational task or not, from the circuit data.

into realistic scenarios or tangible real-world benefits. Perrier, Youssry, and Ferrie [21] provide a quantum dataset called QDataSet, but its aim is to provide a benchmark for classical machine learning applied to quantum physics, and therefore it does not fit to the context of this paper. An example of a quantum-classical dataset in which quantum and classical methods have relatively good performance is instead MNISQ [22], which was introduced after the present work. In our work, we apply machine learning techniques, such as classification or clustering, directly to quantum circuits as input data. Moreover, these quantum circuits originate from practically important quantum computations and are of sufficient scale to meaningfully evaluate their usefulness on current or near-future quantum hardware.

In this work, we propose a more practical machine learning task that we expect to naturally arise in near the future and in the real-world: a clustering and classification of many quantum circuits, and provide an elementary dataset for this task. A successful model that can perform such a task could be beneficial to providers of cloud quantum computers as depicted in Fig. 1, it would allow them to understand the preferences of their users by analyzing circuits submitted by them. While there are many possible ways to analyze the circuits, in this work, we focus on a setting where the providers want to cluster or classify circuits based on their similarity of output states. This task is easy when we have access to quantum computers because similarity, which can be measured in terms of overlaps between output states, can be readily estimated on them. Such estimation is likely to be hard for classical computers when a quantum circuit is large enough [23].

The quantum dataset provided in this work is a set of quantum circuits optimized by the variational quantum eigensolver (VQE) [24,25], which is an algorithm to find circuits that output approximate ground states of quantum systems. More specifically, we use six model Hamiltonians that are famous in condensed matter physics, and optimize 300 different parametrized circuits with varying structure and depth to output each of their ground states. The dataset includes circuits with up to 20 qubits, but it can be easily extended to larger numbers of qubits with access to quantum hardware. To demonstrate the potential of the dataset, we perform a proof-of-principle experiment using quantum circuit simulators and show that the circuits can be accurately clustered. We also

demonstrate successful classification of our four-qubit dataset using real quantum hardware.

The dataset is freely accessible on the GitHub repository [26], therefore the reader will be able to freely use it for research or benchmarking. For each number of qubit, data are stored in QASM [27] format and are publicly accessible. By providing a natural setting and an elementary dataset where quantum machine learning is expected to be beneficial, we hope to support the thoughtful and grounded advancement of the field.

II. DATASET CONSTRUCTION

A. Idea overview

The machine learning task that we consider in this work is clustering and classification of quantum circuits based on the similarity of their output states. More specifically, the task is to classify M quantum circuits $\{U_m \mid m = 1, 2, \dots, M\}$ based on the fidelities of output states $|\langle 0|U_m^\dagger U_{m'}|0\rangle|^2$. We expect this task to naturally arise when quantum computer providers wish to analyze what their users do on their hardware. Also, we believe this task to be hard in general when we have access only to classical computers, since estimation of $|\langle 0|U|0\rangle|^2$ to an accuracy of ϵ for a general polynomial-sized quantum circuit U in polynomial time is clearly a BQP-complete task.

To construct an elementary dataset for this task, we use the VQE [24,25]. It is a method to generate a quantum circuit which outputs an approximate ground state of a target Hamiltonian H . This is usually done by using a parametrized quantum circuit, also referred to as an ansatz, $U(\theta)$ whose parameter θ is optimized to minimize the energy expectation value $\langle H(\theta) \rangle := \langle 0|U^\dagger(\theta) H U(\theta)|0\rangle$. The dataset is constructed by optimizing various ansatz to generate ground states of different Hamiltonians $\{H_l \mid l = 0, 1, \dots, L-1\}$ which have ground states $|g_l\rangle$ that are mutually almost orthogonal, that is, $\langle g_l | g_{l'} \rangle \approx 0$ for $l \neq l'$. Labeling each optimized ansatz U_m based on the Hamiltonian to which it is optimized, we define a dataset $\{(U_m, l_m) \mid m = 1, 2, \dots, M\}$ and $l_m \in \{0, 1, \dots, L-1\}$ as a set of pairs of a quantum circuit and its label.

We can expect this dataset to have a nice property that $|\langle 0|U_m^\dagger U_{m'}|0\rangle|^2 \approx 1$ when $l_m = l_{m'}$ and $|\langle 0|U_m^\dagger U_{m'}|0\rangle|^2 \ll 1$ otherwise. Note that this property persists even if the optimization is imperfect. Suppose two quantum circuits $U_1(\theta_1)$ and $U_2(\theta_2)$ are respectively optimized to output nondegenerate ground states $|g_{l_1}\rangle$ and $|g_{l_2}\rangle$ of Hamiltonians H_{l_1} and H_{l_2} . Furthermore, consider a scenario where the optimization is imperfect, such that, for instance, $|\langle g_{l_m} | U_m(\theta_m) | 0 \rangle|^2 \geq \frac{3}{4}$ for $m = 1, 2$. When $H_{l_1} = H_{l_2}$, $|\langle 0|U_1^\dagger(\theta_1) U_2(\theta_2) | 0 \rangle|^2 \geq 1/4$. On the contrary, assuming $H_{l_1} \neq H_{l_2}$ and $\langle g_{l_1} | g_{l_2} \rangle = 0$, $|\langle 0|U_1^\dagger(\theta_1) U_2(\theta_2) | 0 \rangle|^2 \leq 1/16$. We expect this property makes

TABLE I. Overview of the dataset.

Numbers of qubits	N	4, 8, 12, 16, 20
Number of labels	L	6 (5 for $N = 4$)
Number of circuits for each label	M/L	300
Total number of circuits	M	1800 (1500 for $N = 4$)

TABLE II. Hamiltonians used to generate the dataset.

Label	Name	Hamiltonian
0	1D transverse-field Ising model	$\sum_{n=1}^{N-1} Z_n Z_{n+1} + 2 \sum_{n=1}^N X_n$
1	1D Heisenberg model	$\sum_{n=1}^{N-1} (X_n X_{n+1} + Y_n Y_{n+1} + Z_n Z_{n+1}) + 2 \sum_{n=1}^N Z_n$
2	Su-Schrieffer-Heeger model	$\sum_{n=1}^{N-1} (1 + \frac{3}{2}(-1)^n)(X_n X_{n+1} + Y_n Y_{n+1} + Z_n Z_{n+1})$
3	J_1 - J_2 model	$\sum_{n=1}^{N-1} [(X_n X_{n+1} + Y_n Y_{n+1} + Z_n Z_{n+1}) + 3(X_n X_{n+2} + Y_n Y_{n+2} + Z_n Z_{n+2})]$
4	1D Hubbard model	$-\sum_{j=1}^{N/2-1} \sum_{\sigma \in \{\uparrow, \downarrow\}} (a_{j,\sigma}^\dagger a_{j+1,\sigma} + \text{H.c.}) + \sum_{j=1}^{N/2} (a_{j,\uparrow}^\dagger a_{j,\uparrow} - \frac{1}{2})(a_{j,\downarrow}^\dagger a_{j,\downarrow} - \frac{1}{2})$
5	2D Hubbard model	$-\sum_{\sigma \in \{\uparrow, \downarrow\}} (\sum_{j_x=1}^{N/4-1} \sum_{j_y=1}^2 a_{j_x,j_y,\sigma}^\dagger a_{j_x+1,j_y,\sigma} + \sum_{j_x=1}^{N/4} a_{j_x,1,\sigma}^\dagger a_{j_x,2,\sigma} + \text{H.c.})$ $+ \sum_{j_x=1}^{N/4} \sum_{j_y=1}^2 (a_{j_x,j_y,\uparrow}^\dagger a_{j_x,j_y,\uparrow} - \frac{1}{2})(a_{j_x,j_y,\downarrow}^\dagger a_{j_x,j_y,\downarrow} - \frac{1}{2})$

it easier to extend the dataset by actual experiments using quantum hardware.

B. Dataset construction details

Table I shows the overview of the dataset that we provide in this work. To define the $L = 6$ dataset, we use the Hamiltonians in Table II. X_n , Y_n , and Z_n ($n = 1, 2, \dots, N$) are, respectively, Pauli X , Y , and Z operators acting on the n th qubit. $a_{i,\sigma}^\dagger$ and $a_{i,\sigma}$ is fermion creation and annihilation operators acting on the i th site with the spin σ . For the Hubbard models, we map the Hamiltonians to qubit ones by Jordan-Wigner transformation; specifically, defining $\tilde{a}_i$ ($i = 1, 2, \dots, N$) as $\tilde{a}_{2j-1} = a_{j,\uparrow}$, and $\tilde{a}_{2j} = a_{j,\downarrow}$ for $j = 1, 2, \dots, N/2$, we map them as

$$\tilde{a}_i \rightarrow \frac{1}{2}(X_i + iY_i)Z_{i-1} \cdots Z_1. \quad (1)$$

The sites of 2D Hubbard model are defined on 1×2 , 2×2 , 3×2 , 4×2 , and 5×2 square grids for $N = 4, 8, 12, 16$, and 20 , respectively, and a 2D site index (j_x, j_y) is mapped to

the one-dimensional index j as $j = 2(j_x - 1) + j_y$ for $j_x = 1, 2, \dots, N/4$ and $j_y = 1, 2$.

For ansatz, we use the ones listed in Table III. The one which we refer to as hardware-efficient ansatz has been a popular choice for the VQE ansatz to perform proof-of-principle demonstration of ideas [28–32]. The brick-block ansatz where we sequentially apply parametrized two-qubit gates are also a popular choice [33–36]. The last type of ansatz used in this work is the Hamiltonian ansatz where we sequentially apply Pauli rotations $R_\sigma(\theta) = \exp(-i\theta\sigma/2)$ for all Pauli operators $\sigma = X, Y, Z$ appearing in the problem Hamiltonian. This ansatz is physically motivated by the adiabatic evolution [37,38]. For each ansatz, we vary the circuit depth D from 3 to 32 so that we obtain 30 optimized circuits from each ansatz type. For each D , we sample initial parameters for optimization from the uniform distribution on $[-2\pi, 2\pi)$, except for the Hamiltonian ansatz where we sample from $[0, 0.1)$, for ten times, and adopt the one which achieves the lowest energy expectation value after the optimization to the dataset. The optimization of parameters is performed by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method

TABLE III. Ansate used to generate the dataset. Angles in every rotation gate are treated as independent parameters of circuit. R_σ is a rotation gate defined as $R_\sigma(\theta) = \exp(-i\theta\sigma/2)$ for Pauli operators $\sigma \in \{I, X, Y, Z\}^{\otimes N}$. $V_{n,n'}$ is a four-parameter unitary gate defined as $V_{n,n'} = C_{X_n,n'} R_{Y_n} R_{Y_{n'}} \text{CNOT}_{n,n'} R_{Y_n} R_{Y_{n'}}$, where $\text{CNOT}_{n,n'}$ is a NOT gate on the qubit n' controlled with the qubit n , which is taken from [33]. In addition, the following $\mathcal{P}$ s represent various ordered sets of pairs of sites for the two-qubit gate. $\mathcal{P}_{\text{chain}} = \{(N-2j-1, N-2j) \mid j = 1, 2, \dots, N/2-1\} \cup \{(N-2j, N-2j+1) \mid j = 1, 2, \dots, N/2\}$, $\mathcal{P}_{\text{stair}} = \{(N-n, N-n+1) \mid n = 1, 2, \dots, N-1\}$, $\mathcal{P}_{\text{complete}} = \{(N-n, N-n') \mid n' = 0, 1, \dots, n-1; n = 1, 2, \dots, N-1\}$, $\mathcal{P}_{\text{ladder}} = \{(N-2j-1, N-2j+1) \mid j = 1, 2, \dots, N/2-1\} \cup \{(N-2j, N-2j+2) \mid j = 1, 2, \dots, N/2-1\} \cup \{(2j-1, 2j) \mid j = 1, 2, \dots, N/2\}$, $\mathcal{P}_{\text{cross-ladder}} = \mathcal{P}_{\text{ladder}} \cup \{(N-2j, N-2j+1), (N-2j-1, N-2j+2) \mid j = 1, 2, \dots, N/2-1\}$, where N is the number of qubits.

Name	Ansatz definition
Hamiltonian ansatz	$\prod_{d=1}^D (\prod_{\sigma \in H} R_\sigma \prod_{n=1}^N R_{X_n} R_{Z_n})$
Hardware-efficient ansatz (HE)	$\prod_{d=1}^D [\prod_{n=1}^N R_{Z_n} R_{Y_n} [\prod_{p \in \mathcal{P}_{\text{chain}}} \text{CZ}_p]] \prod_{n=1}^N R_{Z_n} R_{Y_n}$
Complete Hardware-efficient ansatz (Complete-HE)	$\prod_{d=1}^D [\prod_{n=1}^N R_{Z_n} R_{Y_n} [\prod_{p \in \mathcal{P}_{\text{complete}}} \text{CZ}_p]] \prod_{n=1}^N R_{Z_n} R_{Y_n}$
Ladder Hardware-efficient ansatz (Ladder-HE)	$\prod_{d=1}^D [\prod_{n=1}^N R_{Z_n} R_{Y_n} [\prod_{p \in \mathcal{P}_{\text{ladder}}} \text{CZ}_p]] \prod_{n=1}^N R_{Z_n} R_{Y_n}$
Cross-Ladder Hardware-efficient ansatz (Cross-ladder-HE)	$\prod_{d=1}^D [\prod_{n=1}^N R_{Z_n} R_{Y_n} [\prod_{p \in \mathcal{P}_{\text{cross-ladder}}} \text{CZ}_p]] \prod_{n=1}^N R_{Z_n} R_{Y_n}$
1D brick-block ansatz (1D-BB)	$\prod_{n=1}^N R_{Y_n} \prod_{d=1}^D [\prod_{p \in \mathcal{P}_{\text{chain}}} V_p]$
Stair brick-block ansatz (Stair-BB)	$\prod_{n=1}^N R_{Y_n} \prod_{d=1}^D [\prod_{p \in \mathcal{P}_{\text{stair}}} V_p]$
Complete brick-block ansatz (Complete-BB)	$\prod_{n=1}^N R_{Y_n} \prod_{d=1}^D [\prod_{p \in \mathcal{P}_{\text{complete}}} V_p]$
Ladder brick-block ansatz (Ladder-BB)	$\prod_{n=1}^N R_{Y_n} \prod_{d=1}^D [\prod_{p \in \mathcal{P}_{\text{ladder}}} V_p]$
Cross-Ladder brick-block ansatz (Cross-ladder-BB)	$\prod_{n=1}^N R_{Y_n} \prod_{d=1}^D [\prod_{p \in \mathcal{P}_{\text{cross-ladder}}} V_p]$

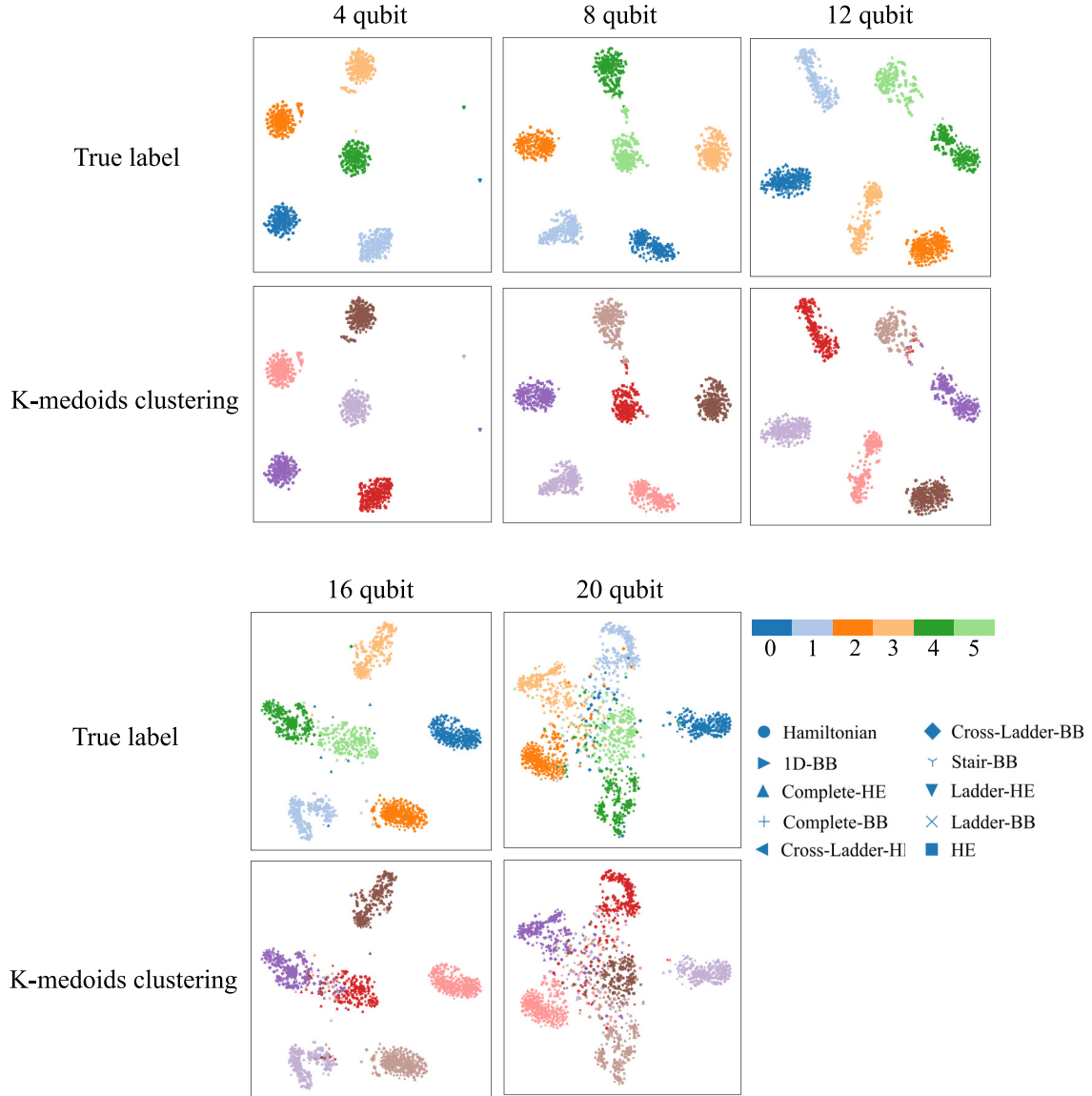


FIG. 2. Visualization and clustering result of the dataset. (Top panels) Visualization of the dataset using t-SNE. The points are colored depending on their true labels. (Bottom panels) Clustering result by k -medoids algorithm. The points are colored based on the clusters. Both plots are generated using exact values of distance matrix $d_{m,m'}$. Different markers correspond to different ansatzes.

implemented on SciPy [39]. Using the exact expectation values without statistical noise computed by Qulacs [40], we optimize the parameters until the norm of gradient becomes less than 10^{-5} or the number of iterations exceeds 1000.

III. DATASET PROPERTIES

A. Visualization of dataset

First, we visualize the constructed dataset using t-stochastic neighbor embedding (t-SNE) [41] to understand the distribution of data intuitively. t-SNE is a visualization method which takes a distance matrix $d_{m,m'}$ of a dataset $\{\mathbf{x}_m\}$ consisting of high-dimensional vectors as its input, and generate low-dimensional points which maintain the similarities among the data points. Here, we adopt $d_{m,m'} = 1 - |\langle 0|U_m^\dagger U_{m'}|0\rangle|^2$ as the distance matrix of our dataset consisting

of quantum circuits $\{U_m\}$. We use the exact values for the inner product of $\langle 0|U_m^\dagger U_{m'}|0\rangle$ calculated by Qulacs [40].

The visualization result is shown in Fig. 2 (top panels). Each point corresponds to a circuit U_m , and it is colored depending on its label. We observe that the 4-, 8-, 12-, and 16-qubit datasets are well clustered, while a portion of the data from the 20-qubit dataset appear to be somewhat intermingled. We find it harder to solve Hamiltonians corresponding to 20 qubits than the others with some ansatz employed in this work. These outcomes depend on the chosen distance metric. Alternative metrics may yield disparate results. Furthermore, the inability to achieve impeccable clustering of the dataset is not regarded as a deficiency because it allows for the comparison of different clustering algorithms. We present in the Appendix the fidelity between the output state and the true ground state of the Hamiltonian for each ansatz. Another interesting feature is the existence of multiple clusters in a

label. We believe this is an artificial effect caused by the t-SNE visualization. This feature cannot be observed by another visualization technique called multidimensional scaling (MDS) [42], which is shown in the Appendix.

To demonstrate the ease of clustering the proposed dataset for ideal quantum computers, we perform clustering based on the exact value of d_{ij} . We note that this is not the only approach for solving the proposed learning task. We employ the k -medoids algorithm implemented in PyClustering [43]. The performance of the clustering is evaluated by adjusted Rand index (ARI) [44], which takes a value between 0 (random clustering) and 1 (perfect clustering). The ARI is evaluated as the mean of ten trials of clustering with different random seeds.

The result is visualized in Fig. 2. The ARI of the 4-, 8-, 12-, 16-, and 20-qubit dataset is, respectively, 0.992, 0.968, 0.927, 0.883, and 0.692. The relatively low ARI for the 20-qubit dataset is caused by the difficulty of producing quantum circuits to output the ground state of label 5. This result indicates that it is possible to cluster this dataset, even in the unsupervised setting, with an ideal quantum computer.

B. Clustering using real quantum hardware and noise model simulator

Here, we show our four-qubit dataset can be reliably learned by using real quantum computers that are presently

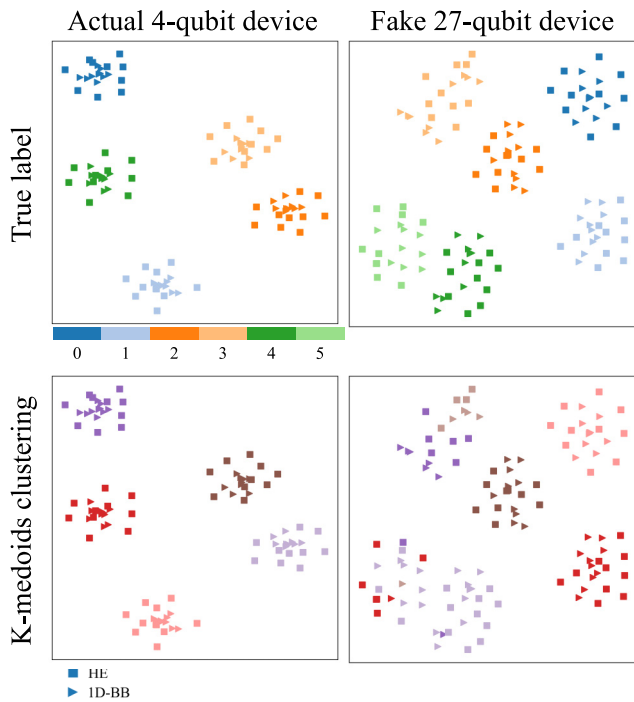


FIG. 3. Results of clustering quantum circuit in the dataset using IBM's four-qubit quantum computer and 27-qubit fake device. Left panel) The result of clustering on the four-qubit dataset by IBM's four-qubit quantum computer. (Right panel) The result of clustering on the 20-qubit dataset by noise model simulation using real noise data of IBM's 27-qubit quantum computer. (Top panel) Visualization of the dataset using t-SNE. The points are colored depending on their true labels. (Bottom panel) Clustering result by k -medoids algorithm.

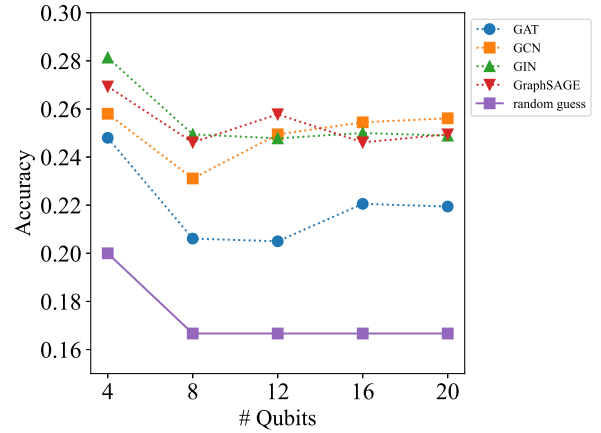


FIG. 4. Performance of different graph-based models for the classification of the dataset.

available. To this end, we perform the clustering by running quantum circuits $U_m^\dagger U_{m'}$ for all pairs of m and m' on the ibmq_manila device available at IBMQ to get fidelity between the two output states. The number of measurements is set to 2×10^4 for each $U_m^\dagger U_{m'}$. Only hardware-efficient ansatz (HE) and 1D brick-block (1D-BB) ansatz with $D = 3$ to 12 are used in the experiments, and thus the number of data is ten for each ansatz.

In Fig. 3, we visualize the dataset by t-SNE using the distance matrix obtained from the experiment. As we can observe, we are able to perfectly cluster the dataset. This is because fidelities between the quantum states belonging to the same labels are maintained to be much larger than those with different labels even in a noisy environment. This result shows the actual quantum computers are capable of learning the dataset we propose.

To investigate the possibility of learning our 20-qubit dataset on actual devices, we also perform clustering by running circuits on FakeAuckland backend available at IBMQ which is a simulator mimicking ibmq_auckland device. The result is visualized in Fig. 3 in the same manner as the above experiment. The ARI is 0.720. This indicates the

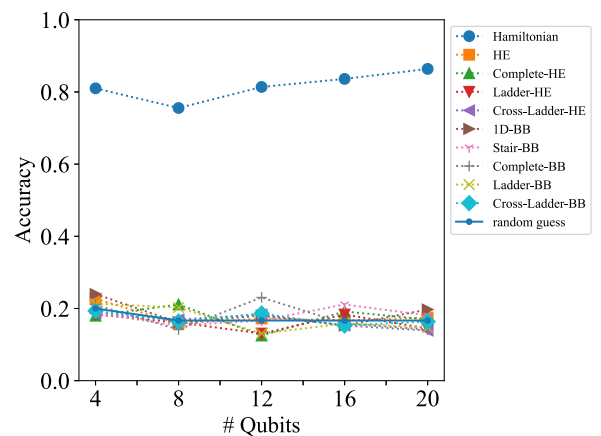


FIG. 5. Classification result using support vector machine on the circuit parameters.

possibility of learning our dataset using real devices even at the size of 20 qubits.

C. Classical machine learning applied to the dataset

We expect our dataset to be hard to be learned solely by classical computers. The most challenging aspect of applying classical machine learning algorithms is identifying an effective feature map that can transform the classical descriptions of a quantum circuit—such as gate types, gate positions, gate parameters, and the sequence of gates—into, for instance, a vector of real values, which can then be processed by various machine learning models. We adopted two approaches: First, we chose to represent quantum circuits as graphs, inspired by recent literature [45], and applied graph neural networks (GNNs) because they are capable of leveraging the full structural information of the circuits. Second, we directly use the circuit parameters as features and apply support vector machine (SVM) classifiers, to test how much information can be learned from a simplified representation.

In the former, we preprocessed the data transforming each circuit to graph. We selected a graph structure inspired by Ref. [45]. A circuit is translated to a graph whose nodes

represent gates and encode information about gates as real vectors, and edges are placed basically when there is a wire in a quantum circuit diagram. We test four graph-based models: graph convolutional networks (GCN) [46], graph attention networks (GAT) [47], graph isomorphism network (GIN) [48], and graph sample and aggregate (GraphSAGE) [49], and estimated the accuracy with threefold cross validation. The performance hardly reaches 30% multiclass accuracy on the training set, showing the hardness of the task. The resulting performance on the validation set can be seen in Fig. 4. For details about the models and their implementations, see [26].

In the latter, we consider the task of classifying a dataset comprised of circuits, all utilizing the same type of ansatz, by employing the circuits' parameter vector, θ , as features within a supervised learning framework. Here, the classical approach is restricted to use only a subset of the information available to the quantum circuits, which is the parameter vector θ . For the model, we employ kernel SVM [50]. To remove the difference in the lengths of the parameter vectors θ depending on the different depths of the circuits, we extend θ to match the deepest circuits within each ansatz type and fill the extended elements with zero. We split 80% of all data into training data and the rest into test data to ensure that the divided data has

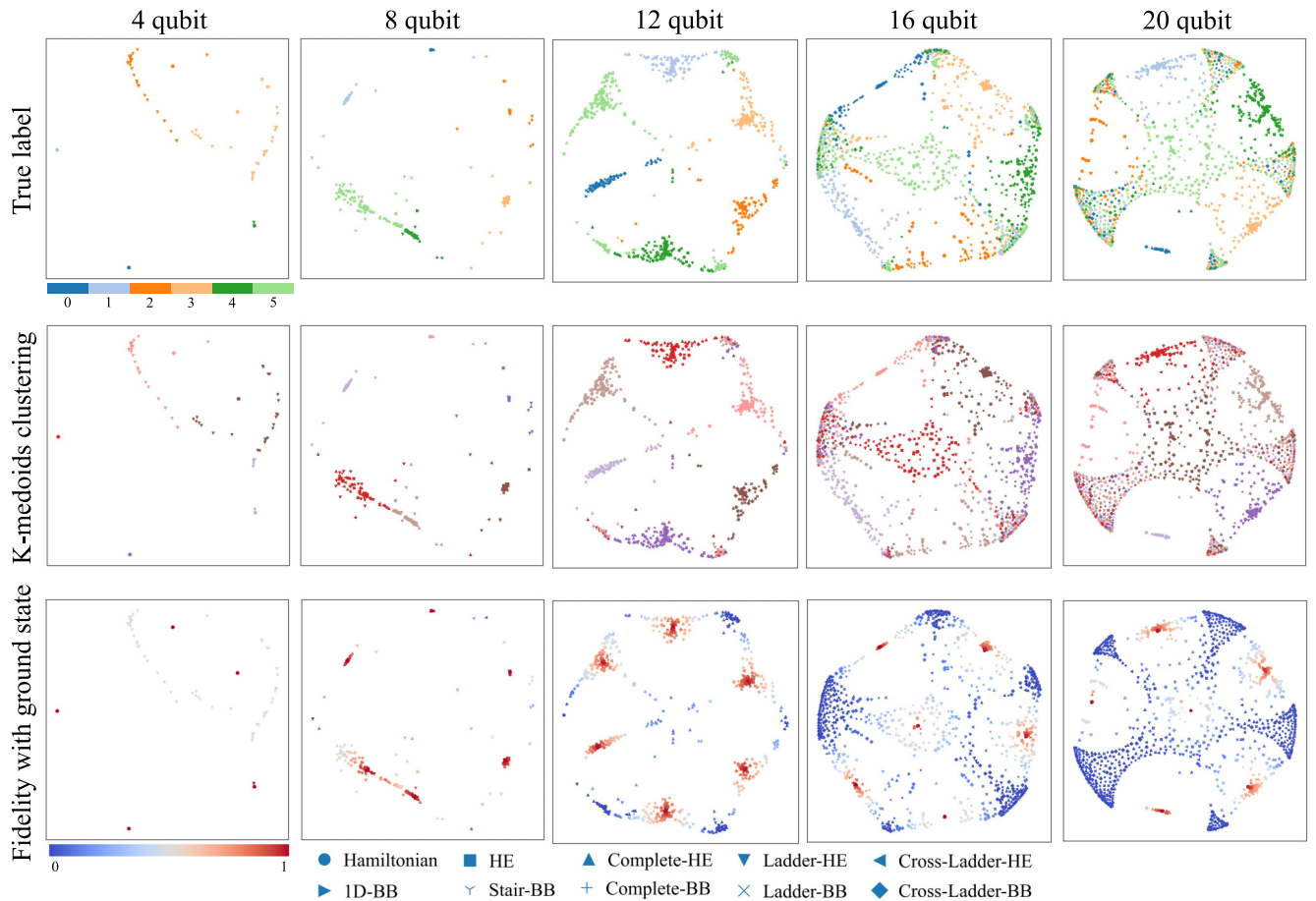


FIG. 6. Visualization and clustering result of the dataset. (Top panel) Visualization of the dataset using MDS. The points are colored depending on their true labels. (Middle panel) Clustering result by k -medoids algorithm. (Bottom panel) The same visualization but are colored depending on their fidelity with the ground state of each label's Hamiltonian.

roughly the same proportions of different labels. We train the model ten times with different splitting and report the mean accuracy score. The regularization strength, types of kernel, and the parameters in kernel are treated as hyperparameters and optimized through the grid search technique combined with stratified threefold cross validation. We note that other classification methods such as random forest have given us similar results.

Figure 5 shows the classification accuracy. For Hamiltonian ansatz, we reached a higher classification accuracy of up to about 80% than other ansätze for all n . Such a high accuracy may be caused by the difference of quantum circuit structures from one label to another, which only exist in the case of the Hamiltonian ansatz. Except for the Hamiltonian ansatz, we only reached a classification accuracy of up to about 20% accuracy. This means that labels are predicted almost randomly. Although there may be a possible improvement to be made to this result by using more sophisticated methods such as neural networks, we can at least say “standard” classification models such as SVM do not work well on the proposed dataset even for the very simplified task.

IV. CONCLUSION

In this paper, we proposed a quantum circuit classification problem as a more practical machine learning task. We introduced the dataset of $N = 4, 8, 12, 16, 20$ -qubit quantum circuits optimized by VQE for different Hamiltonians using different types of ansatz. We verified that the unsupervised clustering of the dataset is easy for ideal quantum computers. In 4-, 8-, 12-, and 16-qubit cases, we achieved the ARI score of over 0.88, and in the 20-qubit case, we achieved 0.69. In particular, we demonstrate a successful classification of our four-qubit subdataset using the actual four-qubit device.

Potential future directions include the following: First, the capability of generating datasets through VQE techniques suggests that actual quantum devices may be able to produce datasets encompassing an even greater number of qubits. This strategy is considered feasible with the careful selection of suitable ansatz and optimization methods. Pertinent research has explored various methods, including pretraining [51], layerwise training [52], and strategies employing tensor networks [53,54]. It is also interesting to explore whether other

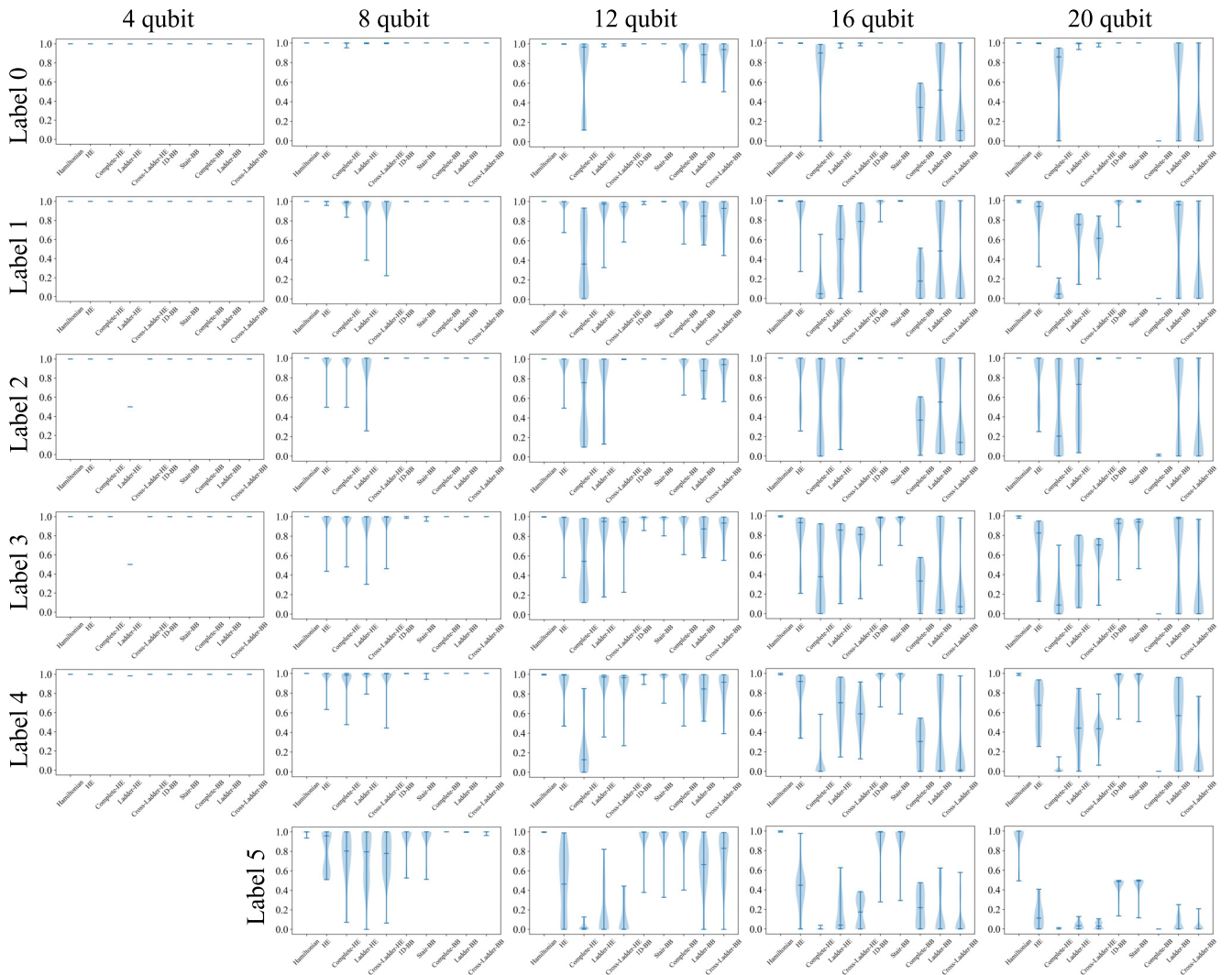


FIG. 7. Fidelity between each data and the ground state of each label's Hamiltonian.

variational quantum algorithms or fault tolerant quantum computer algorithms can also construct a dataset similar to the one provided in this work. Our dataset, which provides a set of practical quantum circuits optimized to output ground states, can also be useful as a benchmark for quantum circuit compilers/transpilers. The devices may be able to conduct similar experiments on our dataset of more numbers of qubits. We finally note that it is worth investigating whether state-of-the-art classical machine learning algorithms can solve the dataset. We published the dataset on GitHub [26].

ACKNOWLEDGMENTS

K.M. is supported by JST PRESTO Grants No. JPMJPR2019, JST FOREST Grant No. JPMJFR232Z, JSPS KAKENHI Grant No. 20K22330, 23H03819, 24K16980 and JST CREST Grant No. JPMJCR24I4. T.S. is supported by JST PRESTO Grant

No. JPMJPR24F4. K.F. is supported by JST ERATO Grant No. JPMJER1601 and JST CREST Grant No. JPMJCR1673. This work is supported by MEXT Quantum Leap Flagship Program (MEXT-QLEAP) Grant No. JPMXS0120319794 and JST COI-NEXT Grant No. JPMJPF2014.

DATA AVAILABILITY

The data that support the findings of this article are openly available [26].

APPENDIX: SUPPLEMENTARY DATA ANALYSIS

Here we present additional details for the main results in our manuscript. In Fig. 6, we present the visualizations of the dataset by MDS [42]. Figure 7 shows the fidelity between each data and ground state of the corresponding Hamiltonian by violin plots to understand how close the output of each circuit is to the ground states.

-
- [1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, *Nature (London)* **549**, 195 (2017).
 - [2] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, Challenges and opportunities in quantum machine learning, *Nat. Comput. Sci.* **2**, 567 (2022).
 - [3] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, [arXiv:1802.06002](https://arxiv.org/abs/1802.06002)
 - [4] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, *Phys. Rev. A* **98**, 032309 (2018).
 - [5] M. Schuld and N. Killoran, Quantum machine learning in feature hilbert spaces, *Phys. Rev. Lett.* **122**, 040504 (2019).
 - [6] J.-G. Liu and L. Wang, Differentiable learning of quantum circuit born machines, *Phys. Rev. A* **98**, 062324 (2018).
 - [7] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, A generative modeling approach for benchmarking and training shallow quantum circuits, *npj Quantum Inf.* **5**, 45 (2019).
 - [8] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, *Nature (London)* **567**, 209 (2019).
 - [9] T. Kusumoto, K. Mitarai, K. Fujii, M. Kitagawa, and M. Negoro, Experimental quantum kernel trick with nuclear spins in a solid, *npj Quantum Inf.* **7**, 94 (2021).
 - [10] K. Bartkiewicz, C. Gneiting, A. Černoč, K. Jiráková, K. Lemr, and F. Nori, Experimental kernel-based quantum machine learning in finite feature space, *Sci. Rep.* **10**, 1 (2020).
 - [11] M. S. Rudolph, N. B. Toussaint, A. Katabarwa, S. Johri, B. Peropadre, and A. Perdomo-Ortiz, Generation of high-resolution handwritten digits with an ion-trap quantum computer, *Phys. Rev. X* **12**, 031010 (2022).
 - [12] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, *Nat. Phys.* **17**, 1013 (2021).
 - [13] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, *Nat. Commun.* **12**, 2631 (2021).
 - [14] V. Dunjko, Y.-K. Liu, X. Wu, and J. M. Taylor, Exponential improvements for quantum-accessible reinforcement learning, [arXiv:1710.11160](https://arxiv.org/abs/1710.11160).
 - [15] S. Jerbi, C. Gyurik, S. Marshall, H. Briegel, and V. Dunjko, Parametrized quantum policies for reinforcement learning, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 28362–28375.
 - [16] N. Pirnay, R. Sweke, J. Eisert, and J.-P. Seifert, A super-polynomial quantum-classical separation for density modelling, *Phys. Rev. A* **107**, 042416 (2023).
 - [17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* **86**, 2278 (1998).
 - [18] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, [arXiv:1708.07747](https://arxiv.org/abs/1708.07747).
 - [19] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, and J. R. McClean, Quantum advantage in learning from experiments, *Science* **376**, 1182 (2022).
 - [20] L. Schatzki, A. Arrasmith, P. J. Coles, and M. Cerezo, Entangled datasets for quantum machine learning, [arXiv:2109.03400](https://arxiv.org/abs/2109.03400).
 - [21] E. Perrier, A. Youssry, and C. Ferrie, Qdataset, quantum datasets for machine learning, *Sci. Data* **9**, 582 (2022).
 - [22] L. Placidi, R. Hataya, T. Mori, K. Aoyama, H. Morisaki, K. Mitarai, and K. Fujii, Mnisq: A large-scale quantum circuit dataset for machine learning on/for quantum computers in the nisq era, [arXiv:2306.16627](https://arxiv.org/abs/2306.16627) (2023).
 - [23] E. Bernstein and U. Vazirani, Quantum complexity theory, *SIAM J. Comput.* **26**, 1411 (1997).
 - [24] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O’Brien, A variational eigenvalue solver on a photonic quantum processor, *Nat. Commun.* **5**, 4213 (2014).
 - [25] J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, E. Grant, L. Wossnig, I. Rungger, G. H. Booth, and J. Tennyson, The

- variational quantum eigensolver: A review of methods and best practices, *Phys. Rep.* **986**, 1 (2022).
- [26] N. Akimoto, M. Kosuke, P. Leonardo, S. Takanori, and F. Keisuke, Qulacs-Osaka, VQE-generated dataset, GitHub repository, <https://github.com/Qulacs-Osaka/VQE-generated-dataset>.
- [27] A. Cross, A. Javadi-Abhari, T. Alexander, N. De Beaudrap, L. S. Bishop, S. Heidel, C. A. Ryan, P. Sivarajah, J. Smolin, J. M. Gambetta, and B. R. Johnson, OpenQASM 3: A Broader and Deeper Quantum Assembly Language, *ACM Trans. Quantum Comput.* **3**, 1 (2022).
- [28] 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 (London)* **549**, 242 (2017).
- [29] K. M. Nakanishi, K. Mitarai, and K. Fujii, Subspace-search variational quantum eigensolver for excited states, *Phys. Rev. Res.* **1**, 033062 (2019).
- [30] J. M. Kübler, A. Arrasmith, L. Cincio, and P. J. Coles, An adaptive optimizer for measurement-frugal variational algorithms, *Quantum* **4**, 263 (2020).
- [31] K. M. Nakanishi, K. Fujii, and S. Todo, Sequential minimal optimization for quantum-classical hybrid algorithms, *Phys. Rev. Res.* **2**, 043158 (2020).
- [32] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nat. Commun.* **9**, 4812 (2018).
- [33] R. M. Parrish, E. G. Hohenstein, P. L. McMahon, and T. J. Martínez, Quantum computation of electronic transitions using a variational quantum eigensolver, *Phys. Rev. Lett.* **122**, 230401 (2019).
- [34] L. Slattery, B. Villalonga, and B. K. Clark, Unitary block optimization for variational quantum algorithms, *Phys. Rev. Res.* **4**, 023072 (2022).
- [35] J. Dborin, F. Barratt, V. Wimalaweera, L. Wright, and A. G. Green, Matrix product state pre-training for quantum machine learning, *Quantum Sci. Technol.* **7**, 035014 (2022).
- [36] M. S. Rudolph, J. Miller, J. Chen, A. Acharya, and A. Perdomo-Ortiz, Synergy between quantum circuits and tensor networks: Short-cutting the race to practical quantum advantage, [arXiv:2208.13673](https://arxiv.org/abs/2208.13673).
- [37] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, *Phys. Rev. A* **92**, 042303 (2015).
- [38] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the hamiltonian variational ansatz, *PRX Quantum* **1**, 020319 (2020).
- [39] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in python, *Nat. Methods* **17**, 261 (2020).
- [40] Y. Suzuki, Y. Kawase, Y. Masumura, Y. Hiraga, M. Nakadai, J. Chen, K. M. Nakanishi, K. Mitarai, R. Imai, S. Tamiya, T. Yamamoto, T. Yan, T. Kawakubo, Y. O. Nakagawa, Y. Ibe, Y. Zhang, H. Yamashita, H. Yoshimura, A. Hayashi, and K. Fujii, Qulacs: a fast and versatile quantum circuit simulator for research purpose, *Quantum* **5**, 559 (2021).
- [41] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* **9**, 2579 (2008).
- [42] J. Kruskal, Nonmetric multidimensional scaling: A numerical method, *Psychometrika* **29**, 115 (1964).
- [43] A. Novikov, PyClustering: Data Mining Library, *J. Open Source Softw.* **4**, 1230 (2019).
- [44] L. Hubert and P. Arabie, Comparing partitions, *J. Classif.* **2**, 193 (1985).
- [45] H. Wang, P. Liu, J. Cheng, Z. Liang, J. Gu, Z. Li, Y. Ding, W. Jiang, Y. Shi, X. Qian, D. Z. Pan, F. T. Chong, and S. Han, Quest: Graph transformer for quantum circuit reliability estimation, [arXiv:2210.16724](https://arxiv.org/abs/2210.16724) v1 [quant-ph].
- [46] T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, [arXiv:1609.02907](https://arxiv.org/abs/1609.02907) [cs.LG].
- [47] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks, [arXiv:1710.10903](https://arxiv.org/abs/1710.10903).
- [48] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, How powerful are graph neural networks? [arXiv:1810.00826](https://arxiv.org/abs/1810.00826).
- [49] W. L. Hamilton, R. Ying, and J. Leskovec, Inductive representation learning on large graphs, [arXiv:1706.02216](https://arxiv.org/abs/1706.02216).
- [50] J. Platt, Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, in *Advances in Large Margin Classifiers* (MIT Press, Cambridge, MA, 2000).
- [51] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, *Quantum* **3**, 214 (2019).
- [52] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, Layerwise learning for quantum neural networks, *Quantum Mach. Intell.* **3**, 5 (2021).
- [53] M. S. Rudolph, J. Miller, D. Motlagh, J. Chen, A. Acharya, and A. Perdomo-Ortiz, Synergy between quantum circuits and tensor networks: Short-cutting the race to practical quantum advantage, *Nat. Commun.* **14**, 8367 (2023).
- [54] R. Watanabe, K. Fujii, and H. Ueda, Variational quantum eigensolver with embedded entanglement using a tensor-network ansatz, *Phys. Rev. Res.* **6**, 023009 (2024).