Generalization of Quantum Machine Learning Models Using Quantum Fisher Information Metric

Tobias Haug[1,2,*] and M. S. Kim[2]
[1]*Quantum Research Center, Technology Innovation Institute, Abu Dhabi, United Arab Emirates*
[2]*Blackett Laboratory, Imperial College London, London SW7 2AZ, United Kingdom*

(Received 30 March 2023; revised 27 June 2024; accepted 3 July 2024; published 31 July 2024)

Generalization is the ability of machine learning models to make accurate predictions on new data by learning from training data. However, understanding generalization of quantum machine learning models has been a major challenge. Here, we introduce the data quantum Fisher information metric (DQFIM). It describes the capacity of variational quantum algorithms depending on variational ansatz, training data, and their symmetries. We apply the DQFIM to quantify circuit parameters and training data needed to successfully train and generalize. Using the dynamical Lie algebra, we explain how to generalize using a low number of training states. Counterintuitively, breaking symmetries of the training data can help to improve generalization. Finally, we find that out-of-distribution generalization, where training and testing data are drawn from different data distributions, can be better than using the same distribution. Our work provides a useful framework to explore the power of quantum machine learning models.

DOI: 10.1103/PhysRevLett.133.050603

The key challenge in quantum machine learning is to design models that can learn from data and apply their acquired knowledge to perform well on new data [1]. This latter ability is called generalization and has been intensely studied recently [2–19]. Constructing models that generalize well is essential for quantum machine learning tasks such as learning unitaries [20–27], classification [28,29], compiling [11,30,31], generative modeling [32,33], quantum simulation [10,34,35], quantum autoencoders [36,37], and black-hole recovery protocols [38]. However, the conditions for generalization are not well understood. Recently proposed uniform generalization bounds [7,39] have been shown to be loose [40], do not account for symmetries, and are unable to explain numerical observations of generalization with few training data [10,39,40].

Thus, there is an urgent need for a framework to understand the conditions for successful training and generalization [8,41–51] to potentially gain advantage over classical models [52–55]. In classical machine learning, generalization has been evaluated using the classical Fisher information [4,5,56–58]. Recent works proposed the quantum Fisher information metric (QFIM) to characterize capacity and overparametrization of parametrized quantum states [49,59–61], however a connection with generalization has not been established.

Here, we introduce the data quantum Fisher information metric (DQFIM) to study generalization and overparametrization. In contrast to the QFIM, the DQFIM correctly captures the effect of data and circuit symmetries on the capacity of quantum machine learning models. The rank of the DQFIM quantifies the circuit depth and amount of data needed for generalization and convergence to a global minimum of the cost function. We apply our methods to learning unitaries, quantum control, generative models, finding excited states, and classification tasks. Using the connection between DQFIM and dynamical Lie algebra (DLA), we explain why quantum machine learning can generalize with few training data. While symmetries have been known to benefit quantum machine learning, we surprisingly find that symmetries in data can also hinder generalization. Finally, we show that out-of-distribution generalization, i.e., the training data are drawn from a different distribution than the test data, can exhibit better performance compared to in-distribution generalization. Our methods provide a quantum geometric picture to understand generalization which guides the design of better quantum machine learning models.

Model—We consider a unitary $U(\boldsymbol{\theta})$ parametrized by M-dimensional parameter vector $\boldsymbol{\theta}$ and training set $S_L = \{|\psi_\ell\rangle, O_\ell\}_{\ell=1}^L$ of size L. S_L consists of input states $|\psi_\ell\rangle$ drawn from a distribution $|\psi_\ell\rangle \in W$, as well as Hermitian operator O_ℓ which represents the label [10,35,39]. We now learn by minimizing the cost function

$$C_{\mathrm{train}}(\boldsymbol{\theta}, S_L) = 1 - \frac{1}{L}\sum_{\ell=1}^{L}\langle\psi_\ell|U(\boldsymbol{\theta})^\dagger O_\ell U(\boldsymbol{\theta})|\psi_\ell\rangle. \quad (1)$$

[*] Contact author: tobias.haug@u.nus.edu

Here, we assume without loosing generality that the eigenvalues of O_ℓ are (tightly) upper bounded by 1 such that $C_{\text{train}} \geq 0$. The trained model generalizes when the test error in respect to unseen test data $|\psi\rangle \in W$ and corresponding label $O_{|\psi\rangle}$ is small

$$C_{\text{test}}(\boldsymbol{\theta}, W) = 1 - \mathop{\mathbb{E}}_{|\psi\rangle \in W}[\langle\psi|U(\boldsymbol{\theta})^\dagger O_{|\psi\rangle} U(\boldsymbol{\theta})|\psi\rangle]. \quad (2)$$

Let us now give two important examples of our model. First, unitary learning or quantum compiling aims to represent a target unitary V with a parametrized unitary $U(\boldsymbol{\theta})$ such that $V|\psi_\ell\rangle = U(\boldsymbol{\theta})|\psi_\ell\rangle$ [26,30]. Here, $|\psi_\ell\rangle$ are initial states with corresponding label operator $O_\ell = V|\psi_\ell\rangle\langle\psi_\ell|V^\dagger$ being the target state to be learned. This learning model also describes quantum control problems [62]. Further, unsupervised generative models to learn a probability distribution $p(x)$ can be converted into unitary learning tasks [63] by encoding empirical distribution $q(x)$ into a state $|\Phi\rangle$ with $q(x) \sim |\langle x|\Phi\rangle|^2$, and choosing $O_\ell = |\Phi\rangle\langle\Phi|$.

Another important task is classification [28]. Here, the goal is to identify two classes, e.g., images of cats and dogs. One encodes the feature vector $\boldsymbol{x}_\ell$ into $|\psi_\ell(\boldsymbol{x}_\ell)\rangle$ with corresponding label $y_\ell = \pm 1$ and label operator $O_\ell = y_\ell \sigma^z$, where cats have $y = 1$ and dogs $y = -1$. The trained model infers the class y by measuring $y \sim \langle\psi|U(\boldsymbol{\theta})^\dagger \sigma^z U(\boldsymbol{\theta})|\psi\rangle$. We note that data reuploading [64], where data and parametrized unitary are interlayered, can also be mapped onto this model [65].

The parametrized unitary $U(\boldsymbol{\theta}) = \prod_{k=1}^G U^{(k)}(\boldsymbol{\theta}_k)$ commonly consists of G repeating layers of unitaries $U^{(k)}(\boldsymbol{\theta}_k) = \prod_{n=1}^K \exp(-i\theta_{kn} H_n)$, where H_n are Hermitian operators, $\boldsymbol{\theta}_k$ a K-dimensional vector, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$ the $M = GK$ dimensional parameter vector [22,62]. The optimization program starts with a randomly chosen $\boldsymbol{\theta}$ and iteratively minimizes Eq. (1) with the gradient $\nabla C_{\text{train}}(\boldsymbol{\theta})$, which can be efficiently estimated by a quantum computer [66]. Gradient descent iteratively updates $\boldsymbol{\theta} \to \boldsymbol{\theta} - \alpha \nabla C_{\text{train}}$ with some α until reaching a minimum after E training steps, where $\nabla C_{\text{train}}(\boldsymbol{\theta}^*) = 0$ with converged parameter $\boldsymbol{\theta}^*$. We assume that ansatz $U(\boldsymbol{\theta})$ can solve the learning task, i.e., we ensure that there is a parameter $\boldsymbol{\theta}_{\text{g}}$ such that $C_{\text{test}}(\boldsymbol{\theta}_{\text{g}}, W) = C_{\text{train}}(\boldsymbol{\theta}_{\text{g}}, S_L) = 0$.

After training we have three possible outcomes: (i) become stuck in local minimum $C_{\text{test}} \geq C_{\text{train}} \gg 0$; (ii) reach global minimum $C_{\text{train}} \approx 0$, however, no generalization with $C_{\text{test}} \gg 0$; (iii) generalization with $C_{\text{train}} \approx C_{\text{test}} \approx 0$. In the following, we show that the DQFIM determines the critical number of circuit parameters $M_{\text{c}}(L)$ for overparametrization as a function of L and training states L_{c} for generalization.

DQFIM—First, we define what can be learned about ansatz unitary $U(\boldsymbol{\theta})$ via training set S_L:

Definition 1 (unitary mapped onto data)—The data state for training set $S_L = \{|\psi_\ell\rangle, O_\ell\}_{\ell=1}^L$ of L states is

$$\rho_L = \frac{1}{L}\sum_{\ell=1}^{L} |\psi_\ell\rangle\langle\psi_\ell| \quad (3)$$

Training with cost function Eq. (1) and S_L gives only information about the unitary mapped onto the subspace of the training data $U_L(\boldsymbol{\theta}) \sim U(\boldsymbol{\theta})\rho_L$.

To understand Definition 1, consider the d-dimensional unitary $U \equiv U(\boldsymbol{u}) = \sum_{n,k=1}^d u_{nk}|n\rangle\langle k|$ with complex parameters $\boldsymbol{u} = \{u_{11}, u_{12}, \ldots, u_{dd}\}$ and training data $\{|\ell\rangle\}_{\ell=1}^L$, where $|\ell\rangle \in W$ are computational basis states and our goal is to learn some unitary $V = U(\boldsymbol{u}^*)$. For $L = 1$, training with Eq. (1) optimizes $U|1\rangle = \sum_{n=1}^d u_{n1}|n\rangle$. Thus, only the column vector $u_1 = (u_{11}, u_{21}, \ldots, u_{d1})$ of U can be trained, while $u_{n>1}$ are not learnable. For arbitrary L, applying U on the training states gives us $\{U|\ell\rangle = \sum_{n=1}^d u_{n\ell}|n\rangle\}_{\ell=1}^L$. The learnable parameters of U correspond to the $d \times L$-dimensional (unnormalized) isometry $U_L = (u_1, \ldots, u_L) \equiv U\rho_L$ with (unnormalized) projector $\rho_L = L^{-1}\sum_{\ell=1}^L |\ell\rangle\langle\ell|$ [see Fig. 1(a)]. Even if we find a global minima with $C_{\text{train}} = 0$, for $L < d$ we gain no information about the column vectors $(u_{L+1}, \ldots, u_d)$. The trained model $U(\boldsymbol{u}^*)$ randomly guesses these column vectors, resulting in a large generalization error C_{test}. Only for $L = d$, we have a complete training set that can achieve generalization with $C_{\text{test}} = 0$.

To understand generalization, we count the independent parameters of U_L, which we call the effective dimension D_L. For $L = 1$, $U|1\rangle = \sum_{n=1}^d u_{n1}|n\rangle = \sum_{n=1}^d (a_{n1} + ib_{n1})|n\rangle$ has $2d$ real parameters a_{n1}, b_{n1}. However, due

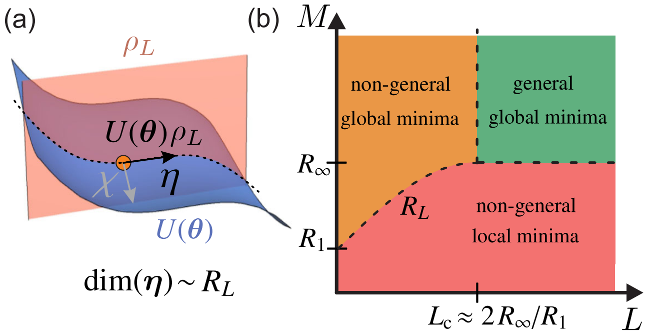

FIG. 1. (a) Ansatz unitary $U(\boldsymbol{\theta})$ and M-dimensional parameter vector $\boldsymbol{\theta}$ is optimized in respect to cost function Eq. (1) using L training data described by data state ρ_L [Eq. (3)]. Only the subspace of the unitary that acts on the training data $U_L \equiv U(\boldsymbol{\theta})\rho_L$ can be learned. Its learnable degrees of freedom are given by the maximal rank of the data quantum Fisher information metric (DQFIM) R_L. (b) Phase diagram of generalization with M and L. Convergence to global minimum ($C_{\text{train}} \approx 0$) is likely for overparametrization $M \geq R_L$. Generalization to unseen test data ($C_{\text{test}} \approx 0$) for overcomplete training data when $L \geq L_{\text{c}} \approx 2R_\infty/R_1$ and $M \geq R_\infty$.

to global phase and norm, there are only $D_1 = 2d - 2$ independent parameters. For $L = d$, parametrizing a complete unitary U requires $D_d = d^2 - 1$ parameters. For example, a single qubit has $D_1 = 2$ (Bloch sphere) and $D_2 = 3$ (arbitrary unitary) free parameters [67], and thus we require $L \geq 2$ states to generalize. However, depending on ansatz and data structure D_L can decrease. Let us consider $U(\boldsymbol{\theta}) = e^{-i\sigma_z\theta_k} \cdots e^{-i\sigma_z\theta_M}$ and distribution $W = \{|+\rangle, |-\rangle\}$ with $|\pm\rangle \sim |0\rangle \pm |1\rangle$ and z-Pauli σ_z. While we have M parameters, the generators commute and $L = 1$ is sufficient to generalize as $D_1 = D_d = 1$. In contrast, for $W = \{|0\rangle, |1\rangle\}$ we have $D_L = 0$ as only the trivial global phase is rotated.

We now propose the DQFIM to quantify the effective dimension [see Supplemental Material (SM) A [68]].

Definition 2 (DQFIM)—For unitary $U(\boldsymbol{\theta}) \equiv U$ and training set S_L, the DQFIM is defined as

$$\mathcal{Q}_{nm}(\rho_L, U) = 4\mathrm{Re}[\mathrm{tr}(\partial_n U \rho_L \partial_m U^\dagger) - \mathrm{tr}(\partial_n U \rho_L U^\dagger)\mathrm{tr}(U\rho_L \partial_m U^\dagger)], \tag{4}$$

where ∂_n is the derivative in respect to the nth entry of the M-dimensional vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$.

In SM A, we derive $\mathcal{Q}[\rho_L, U(\boldsymbol{\theta})]$ as the metric that describes how a variation $\boldsymbol{\theta} \to \boldsymbol{\theta} + \mathrm{d}\boldsymbol{\theta}$ changes the mapping of $U(\boldsymbol{\theta})$ onto the span of ρ_L, and relate the DQFIM to the QFIM of the purification of ρ_L. For $L = 1$, we recover the QFIM $\mathcal{F}_{nm} = 4\mathrm{Re}(\langle\partial_n\psi|\partial_m\psi\rangle - \langle\partial_n\psi|\psi\rangle\langle\psi|\partial_m\psi\rangle)$ [60,76].

The rank of $\mathcal{Q}$ gives the effective dimension

$$D_L[\rho_L, U(\boldsymbol{\theta})] = \mathrm{rank}[\mathcal{Q}(\rho_L, U(\boldsymbol{\theta})] \leq M. \tag{5}$$

The case $L = 1$ has been studied previously [49]: the effective dimension D_1 increases with M, until reaching a maximal value R_1 [see Fig. 2(a)]. Once maximal, the parametrized state $U(\boldsymbol{\theta})|\psi_1\rangle$ is overparametrized as it can explore all its degrees of freedom [59]. While $D_L(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, it turns out that due to self-averaging, a randomly chosen $\boldsymbol{\theta}_{\mathrm{rand}}$ nearly always assumes its maximal rank $\max_{\boldsymbol{\theta}} D_L(\boldsymbol{\theta}) \approx D_L(\boldsymbol{\theta}_{\mathrm{rand}})$ [49]. Just as D_1, our D_L increases with M until a maximal R_L, which describes the maximal number of degrees of freedom that U_L can explore and heralds overparametrization for arbitrary L:

Definition 3 (overparametrization)—Ansatz $U(\boldsymbol{\theta})$ with training data ρ_L is overparametrized when effective dimension D_L does not increase further upon increasing the number of parameters M. The maximal rank R_L reached at critical number of parameters $M \geq M_{\mathrm{c}}(L)$:

$$R_L \equiv \max_{M \geq M_{\mathrm{c}}(L), \boldsymbol{\theta}} D_L[\rho_L, U(\boldsymbol{\theta})]. \tag{6}$$

For overparametrization with $M \geq M_{\mathrm{c}}(L)$, a variation of $\boldsymbol{\theta}$ can explore all degrees of freedom of U_L and thus likely find the global minimum [59,77–80]:

Observation 1 (convergence to global minimum)—Global minimum $C_{\mathrm{train}}(\boldsymbol{\theta}^*) \approx 0$ with training set S_L is reached with high probability when $M \geq M_{\mathrm{c}}(L) \geq R_L$.

As seen in Fig. 2(b), R_L increases with L, where the growth slows down due to unitary constraints. We find the tight upper bound (SM, Sec. B or [81])

$$R_L \leq 2dL - L^2 - 1 \ \text{for}\, L \leq d; \quad R_L \leq d^2 - 1 \ \text{for}\, L > d. \tag{7}$$

R_L increases with L until its maximal possible value $R_{L_{\mathrm{c}}} \equiv R_\infty$. Here, the training data are overcomplete and sufficient to learn all degrees of freedom of $U(\boldsymbol{\theta})$:

Definition 4 (overcomplete data)—A given model $U(\boldsymbol{\theta})$ and ρ_L is overcomplete when R_L does not increase further upon increasing L. Its maximal rank $R_\infty = R_{L_{\mathrm{c}}}$ is reached for a critical number of training data $L \geq L_{\mathrm{c}}$

$$R_\infty = R_{L_{\mathrm{c}}} \equiv \max_{L \geq L_{\mathrm{c}}} R_L(\rho_L, U). \tag{8}$$

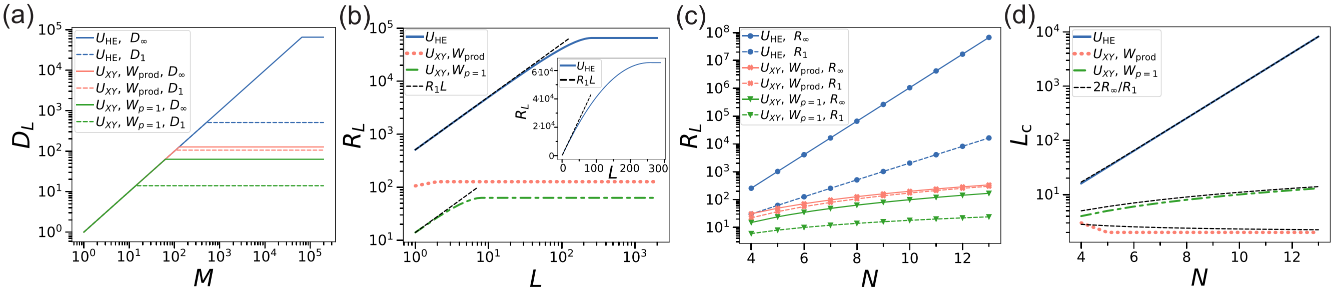

FIG. 2. DQFIM for different unitaries U with M parameters and L training states. As defined in SM Sec. D, we show hardware efficient circuit U_{HE} with no symmetries and Haar random training states (blue curves), as well as U_{XY} with particle number symmetry using as training data either product states W_{prod} (orange) or symmetry-conserving states $W_{p=1}$ (green). (a) Effective dimension D_L increases linearly with M, until it reaches a maximal value R_L for $M \geq M_{\mathrm{c}}(L)$. We have $N = 8$ qubits. (b) R_L increases with L until converging to R_∞ for $L \geq L_{\mathrm{c}}$. Black dashed line shows approximation $R_L \sim R_1 L$. The inset shows generic ansatz without log-plot, highlighting the nonlinear behavior of R_L. (c) Scaling of R_1 and R_∞ with qubit number N. (d) Number L_{c} of training states needed for generalization. Black dashed line shows $L_{\mathrm{c}} \approx 2R_\infty/R_1$.

We bound R_L similar to R_1 for Ref. [59] (see SM, Sec. C [68]):

Theorem 1—The maximal rank R_L is bounded by the dimension of the DLA $R_L \leq \dim(\mathfrak{g})$ where $\mathfrak{g} = \mathrm{span}\langle iH_1, \ldots, iH_K \rangle_{\mathrm{Lie}}$ is generated by the repeated nested commutators of the generators H_k of $U(\boldsymbol{\theta})$.

Thus, using an ansatz with restricted Lie algebra [43,45,84] with $\dim(\mathfrak{g}) \sim \mathrm{poly}(N)$ generalizes with $L_{\mathrm{c}}, M_{\mathrm{c}} \sim \mathrm{poly}(N)$ where N is the number of qubits.

We can estimate L_{c} with the following consideration: to generalize we have to learn all R_∞ degrees of freedom of the unitary. The first training state allows us to learn R_1 degrees of freedom, while each additional state provides a bit less as seen in Eq. (7). For the upper bound of Eq. (7) we have $L_{\mathrm{c}} \approx 2R_\infty/R_1$, which we numerically find to be a good estimator also for other models:

Observation 2 (generalization for learning unitaries)—A trained model generalizes $C_{\mathrm{test}}(\boldsymbol{\theta}^*) \approx 0$ with high probability when the model is overparametrized (i.e., $M \geq M_{\mathrm{c}} \geq R_L$ for Definition 3) and overcomplete (i.e., $L \geq L_{\mathrm{c}}$ for Definition 4). The critical number of training states L_{c} needed to generalize can be approximated by

$$L_{\mathrm{c}} \approx 2R_\infty/R_1. \tag{9}$$

Applications—We want to learn the unitary evolution $V_{XY} = \exp(-iH_{XY}t)$ at time t of the XY Hamiltonian $H_{XY} = \sum_{k=1}^{N}(\sigma_k^x\sigma_{k+1}^x + \sigma_k^y\sigma_{k+1}^y + h_k\sigma_k^z)$, where σ_k^α, $\alpha \in \{x, y, z\}$ is the Pauli operator acting on qubit k and $h_k \in \mathbb{R}$. We learn V_{XY} with $U_{XY}(\boldsymbol{\theta})$ ansatz (see SM Sec. D for definition [68]), which can represent any V_{XY} with polynomial number of parameters [85,86]. H_{XY} and U_{XY} conserve the particle number operator $P = \sum_{k=1}^{N}\frac{1}{2}(1 - \sigma_k^z)$ with $[U_{XY}, P] = [H_{XY}, P] = 0$, where $[.]$ is the commutator. As training states, we use $|\psi_\ell\rangle \in W_{p=1}$ which are states symmetric in regard to P, i.e., $P|\psi_\ell\rangle = p|\psi_\ell\rangle$ with the same eigenvalue $p = 1$ for all $|\psi_\ell\rangle \in W_{p=1}$. Further, we have the single-qubit product states W_{prod} with $|\psi_\ell\rangle = \otimes_{k=1}^{N} |\phi_\ell^k\rangle$, $|\phi_\ell^k\rangle \in \mathcal{H}(\mathbb{C}^2)$ which are not symmetric in respect to P.

Observation 3 (nonsymmetric data improve generalization)—We train $U_{XY}(\boldsymbol{\theta})$ with (i) particle-number conserving states $|\psi_\ell\rangle \in W_{p=1}$ and (ii) single-qubit product states $|\psi_\ell\rangle \in W_{\mathrm{prod}}$. For $W_{p=1}$ we find exactly $R_1 = 2N - 2$, $R_\infty = N^2 - 1$, while for W_{prod} we find via numerical extrapolation $R_1 = 2N^2 - 3N + 2$ and $R_\infty = 2N^2 - 1$ [Figs. 2(c) and 2(d)]. Generalization requires less $L_{\mathrm{c}} \approx 2R_\infty/R_1$ training states for nonsymmetric data:

$$\text{(i) Symmetric: } L_{\mathrm{c}} = N \text{ for } |\psi_\ell\rangle \in W_{p=1}$$

$$\text{(ii) Nonsymmetric: } L_{\mathrm{c}} = 2 \text{ for } |\psi_\ell\rangle \in W_{\mathrm{prod}}, \quad N > 4.$$

Intuitively, nonsymmetric data require less L as it can use information from other symmetry sectors.

Next, we consider out-of-distribution generalization where the training data is drawn from a different distribution than the test data [8]:

Observation 4 (out-of-distribution generalization requires less data)—Training $U_{XY}(\boldsymbol{\theta})$ with product states $|\psi_\ell\rangle \in W_{\mathrm{prod}}$, but testing with number-conserving data $W_{p=1}$ achieves out-of-distribution generalization with only $L \geq 2$ training data. In contrast, in-distribution training and testing with number-conserving data $|\psi_\ell\rangle \in W_{p=1}$ requires $L \geq N$ states to generalize.

This result follows from Observation 3 and product states being sufficient to learn arbitrary unitaries [8]. We confirm our result numerically in Fig. 3(d).

Numerical results—In Figs. 3(a)–3(c) we study learning of unitaries with hardware-efficient ansatz $U_{\mathrm{HE}}(\boldsymbol{\theta})$ (see SM Sec. D [68]). In Fig. 3(a), we converge to local minima with $C_{\mathrm{train}} \gg 0$ for $M \leq R_L$, while we find global minimum $C_{\mathrm{train}} \approx 0$ for $M \geq R_L$, which is indicated as black dashed line. In Fig. 3(b), generalization $C_{\mathrm{test}} \approx 0$ is achieved only for $M \geq R_\infty$ and $L \geq L_{\mathrm{c}} \approx 2R_\infty/R_1$ indicated by the vertical black line. In Fig. 3(c), the number of training steps E to converge show characteristic peaks close to M_{c} and L_{c} indicated by black dashed lines. In Fig. 3(d) we show the test error against L for the U_{XY} ansatz which

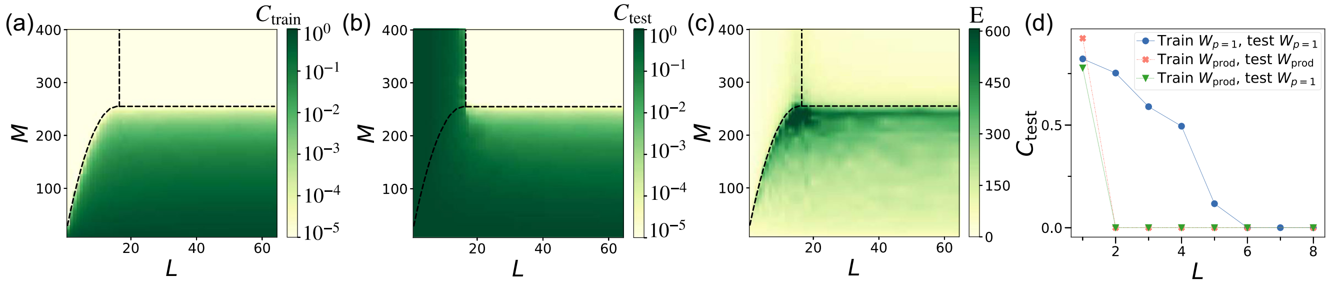

FIG. 3. (a) Median C_{train} against M and L for learning unitaries. Dashed black lines indicate $M_{\mathrm{c}}(L) = R_L$ and $L_{\mathrm{c}} = 2R_\infty/R_1$. We have a $N = 4$ qubit hardware-efficient ansatz trained with random training states and gradient descent [82] simulated with [83]. Target unitary is $V = U(\boldsymbol{\theta}_{\mathrm{g}})$ with random parameter $\boldsymbol{\theta}_{\mathrm{g}}$, where we take median over 10 random instances. (b) Average C_{test} against M and L. (c) Number of training steps E until reaching $C_{\mathrm{train}} < 10^{-4}$. (d) C_{test} against L with particle-number conserving U_{XY} ansatz for $N = 6$ qubits and $M = 90$. We train and test with product states W_{prod} and particle-number conserving states $W_{p=1}$.

conserves particle number P. We find that training with symmetric data W_{prod} generalizes for $L \geq 2$, while training with nonsymmetric $W_{p=1}$ generalizes for $L \geq N$ which numerically confirms Observation 3. Further, the green curve shows out-of-distribution generalization where training with W_{prod} generalizes with test data from $W_{p=1}$ using only $L \geq 2$, while in-distribution learning (blue curve) requires $L \geq N$, confirming Observation 4. We study U_{XY} in more detail in SM, Sec. E and other models which generalize for constant L in SM, Sec. F [68].

Conclusion—Our newly introduced DQFIM $\mathcal{Q}$ and its maximal rank R_L quantify the learnable degrees of freedom of ansatz $U(\boldsymbol{\theta})$ using L training states. R_L increases with L until the training data become overcomplete at $R_{L_\mathrm{c}} = R_\infty$ and $L_\mathrm{c} \approx 2R_\infty/R_1$ where one is able to generalize.

Overparametrized models converge to global minima with high probability [59,77,79,80,87–89]. We show that overparametrization depends on L and occurs for $M \geq M_\mathrm{c}(L) \geq R_L$ circuit parameters. Overparametrization and generalization appear in three distinct regimes, where training time increases substantially at the transitions, potentially indicating computational phase transitions [26,79].

While symmetries have been shown to improve generalization [45,46], we show that symmetries in data can also increase L needed to generalize due to higher R_∞/R_1 ratio compared to nonsymmetric data. This also implies that out-of-distribution generalization can outperform in-distribution generalization when training on nonsymmetric data, but testing on symmetric data. Note that nonsymmetric data have larger M_c, which implies an interesting trade-off between L_c and M_c.

The DQFIM accurately characterizes overparametrization and generalization depending on the specific structure and symmetries of ansatz $U(\boldsymbol{\theta})$ and training data ρ_L. In contrast, previously considered uniform generalization bounds provide only a loose bound on generalization error $\sim\sqrt{1/L}$ without accounting for symmetries [7,39]. We demonstrate the relationship between DLA and generalization, showing that polynomial DLA implies overparametrization and generalization with polynomial circuit depth and dataset size. Generalization with few data is possible whenever $R_\infty/R_1 = \mathrm{const}$, explaining the numerical observations of Refs. [10,39] (see also SM Sec. F). Thus, problem classes with polynomial DLA [41,84,90] can be trained with low data cost and avoid barren plateaus [91,92].

Our results apply to various quantum machine learning algorithms. We study unitary learning problems, which includes quantum compiling [30], quantum control (SM, Sec. G), and quantum generative models (SM, Sec. H). In SM Sec. I the DQFIM determines convergence of the subspace-search variational quantum eigensolver for finding eigenstates of Hamiltonians [93]. In SM Sec. J we apply the DQFIM for classification tasks. Here, the label operator $O_\ell = y_\ell \sigma^z$ is not a projector and thus has not only one, but 2^{N-1} degenerate solutions. This reduces $M_\mathrm{c}(L) \approx R_L\gamma$ by a constant factor $\gamma \leq 1$ where for a generic ansatz we find $\gamma = \frac{1}{2}$. γ can be smaller when the ansatz has symmetries which opens an interesting approach to reduce circuit depth in classification tasks.

Numerical evaluation of the DQFIM is straightforward via differentiation (with code available online [94]) and is scalable for matrix product states. Quantum computers can efficiently measure the DQFIM using the Hadamard test with a single ancilla and two control operations, or alternatively the shift rule and purification [95] in SM, Sec. A.

While the complexity of unitaries grows linearly with M [49,96], we find that the learnable degrees of freedom R_L of unitaries grows only sublinearly with L. Generalization error for overparametrized models scales as $C_{\mathrm{test}} \sim 1 - (L/L_\mathrm{c})^2$ (see SM Sec. E) which saturates the lower bound derived in Ref. [6]. We note that for underparametrized models the empirical generalization error $C_{\mathrm{test}} - C_{\mathrm{train}}$ [40] is not a good indicator of generalization due to convergence to bad local minima (see SM, Sec. K).

Finally, future work can apply the DQFIM for kernel models [97,98], noisy systems [61], and quantum natural gradients [99].

Acknowledgments—We thank Adithya Sireesh and Raghavendra Peddinti for discussions. This work is supported by a Samsung GRC project and the U.K. Hub in Quantum Computing and Simulation, part of the U.K. National Quantum Technologies Programme with funding from UKRI EPSRC Grant No. EP/T001062/1.

[1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[2] K. Poland, K. Beer, and T. J. Osborne, No free lunch for quantum machine learning, arXiv:2003.14103.

[3] M. C. Caro and I. Datta, Pseudo-dimension of quantum circuits, Quantum Mach. Intell. **2**, 14 (2020).

[4] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, Nat. Comput. Sci. **1**, 403 (2021).

[5] A. Abbas, D. Sutter, A. Figalli, and S. Woerner, Effective dimension of machine learning models, arXiv:2112.04807.

[6] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, Reformulation of the no-free-lunch theorem for entangled datasets, Phys. Rev. Lett. **128**, 070501 (2022).

[7] L. Banchi, J. Pereira, and S. Pirandola, Generalization in quantum machine learning: A quantum information standpoint, PRX Quantum **2**, 040321 (2021).

[8] 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, Nat. Commun. **14**, 3751 (2023).

[9] E. Peters and M. Schuld, Generalization despite overfitting in quantum machine learning models, Quantum **7**, 1210 (2023).

[10] J. Gibbs, Z. Holmes, M. C. Caro, N. Ezzell, H.-Y. Huang, L. Cincio, A. T. Sornborger, and P. J. Coles, Dynamical simulation via quantum machine learning with provable generalization, Phys. Rev. Res. **6**, 013241 (2024).

[11] T. Volkoff, Z. Holmes, and A. Sornborger, Universal compiling and (no-) free-lunch theorems for continuous-variable quantum learning, PRX Quantum **2**, 040327 (2021).

[12] H. Cai, Q. Ye, and D.-L. Deng, Sample complexity of learning parametric quantum circuits, Quantum Sci. Technol. **7**, 025014 (2022).

[13] C. M. Popescu, Learning bounds for quantum circuits in the agnostic setting, Quantum Inf. Process. **20**, 286 (2021).

[14] M. C. Caro, E. Gil-Fuster, J. J. Meyer, J. Eisert, and R. Sweke, Encoding-dependent generalization bounds for parametrized quantum circuits, Quantum **5**, 582 (2021).

[15] K. Bu, D. E. Koh, L. Li, Q. Luo, and Y. Zhang, Statistical complexity of quantum circuits, Phys. Rev. A **105**, 062431 (2022).

[16] J. Bowles, V. J. Wright, M. Farkas, N. Killoran, and M. Schuld, Contextuality and inductive bias in quantum machine learning, arXiv:2302.01365.

[17] Y. Du, Y. Yang, D. Tao, and M.-H. Hsieh, Problem-dependent power of quantum neural networks on multi-class classification, Phys. Rev. Lett. **131**, 140601 (2023).

[18] K. Bu, D. E. Koh, L. Li, Q. Luo, and Y. Zhang, Effects of quantum resources and noise on the statistical complexity of quantum circuits, Quantum Sci. Technol. **8**, 025013 (2023).

[19] K. Gili, M. Mauri, and A. Perdomo-Ortiz, Evaluating generalization in classical and quantum generative models, arXiv:2201.08770.

[20] A. Bisio, G. Chiribella, G. M. D'Ariano, S. Facchini, and P. Perinotti, Optimal quantum learning of a unitary transformation, Phys. Rev. A **81**, 032324 (2010).

[21] I. Marvian and S. Lloyd, Universal quantum emulator, arXiv:1606.02734.

[22] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum algorithms, Rev. Mod. Phys. **94**, 015004 (2022).

[23] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[24] S. Xue, Y. Liu, Y. Wang, P. Zhu, C. Guo, and J. Wu, Variational quantum process tomography of unitaries, Phys. Rev. A **105**, 032427 (2022).

[25] V. Gebhart, R. Santagati, A. A. Gentile, E. M. Gauger, D. Craig, N. Ares, L. Banchi, F. Marquardt, L. Pezzè, and C. Bonato, Learning quantum systems, Nat. Rev. Phys. **5**, 141 (2023).

[26] B. T. Kiani, S. Lloyd, and R. Maity, Learning unitaries by gradient descent, arXiv:2001.11897.

[27] Z. Yu, X. Zhao, B. Zhao, and X. Wang, Optimal quantum dataset for learning a unitary transformation, Phys. Rev. Appl. **19**, 034017 (2023).

[28] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, arXiv:1802.06002.

[29] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).

[30] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, Quantum-assisted quantum compiling, Quantum **3**, 140 (2019).

[31] N. Ezzell, E. M. Ball, A. U. Siddiqui, M. M. Wilde, A. T. Sornborger, P. J. Coles, and Z. Holmes, Quantum mixed state compiling, Quantum Sci. Technol. **8**, 035001 (2023).

[32] B. Coyle, D. Mills, V. Danos, and E. Kashefi, The Born supremacy: Quantum advantage and training of an Ising born machine, npj Quantum Inf. **6**, 60 (2020).

[33] K. Gili, M. Hibat-Allah, M. Mauri, C. Ballance, and A. Perdomo-Ortiz, Do quantum circuit born machines generalize?, Quantum Sci. Technol. **8**, 035021 (2023).

[34] C. Cirstoiu, Z. Holmes, J. Iosue, L. Cincio, P. J. Coles, and A. Sornborger, Variational fast forwarding for quantum simulation beyond the coherence time, npj Quantum Inf. **6**, 82 (2020).

[35] J. Gibbs, K. Gili, Z. Holmes, B. Commeau, A. Arrasmith, L. Cincio, P. J. Coles, and A. Sornborger, Long-time simulations for fixed input states on quantum hardware, npj Quantum Inf. **8**, 135 (2022).

[36] J. Romero, J. P. Olson, and A. Aspuru-Guzik, Quantum autoencoders for efficient compression of quantum data, Quantum Sci. Technol. **2**, 045001 (2017).

[37] H. Zhang, L. Wan, T. Haug, W.-K. Mok, S. Paesani, Y. Shi, H. Cai, L. K. Chin, M. F. Karim, L. Xiao *et al.*, Resource-efficient high-dimensional subspace teleportation with a quantum autoencoder, Sci. Adv. **8**, eabn9783 (2022).

[38] L. Leone, S. F. E. Oliviero, S. Piemontese, S. True, and A. Hamma, Retrieving information from a black hole using quantum machine learning, Phys. Rev. A **106**, 062434 (2022).

[39] 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, Nat. Commun. **13**, 4919 (2022).

[40] E. Gil-Fuster, J. Eisert, and C. Bravo-Prieto, Understanding quantum machine learning also requires rethinking generalization, Nat. Commun. **15**, 2277 (2024).

[41] L. Schatzki, M. Larocca, Q. T. Nguyen, F. Sauvage, and M. Cerezo, Theoretical guarantees for permutation-equivariant quantum neural networks, npj Quantum Inf. **10**, 12 (2024).

[42] M. Ragone, P. Braccia, Q. T. Nguyen, L. Schatzki, P. J. Coles, F. Sauvage, M. Larocca, and M. Cerezo, Representation theory for geometric quantum machine learning, arXiv:2210.07980.

[43] Q. T. Nguyen, L. Schatzki, P. Braccia, M. Ragone, P. J. Coles, F. Sauvage, M. Larocca, and M. Cerezo, Theory for equivariant quantum neural networks, PRX Quantum **5**, 020328 (2024).

[44] H. Zheng, Z. Li, J. Liu, S. Strelchuk, and R. Kondor, Speeding up learning quantum states through group equivariant convolutional quantum ansätze, PRX Quantum **4**, 020327 (2023).

[45] J. J. Meyer, M. Mularski, E. Gil-Fuster, A. A. Mele, F. Arzani, A. Wilms, and J. Eisert, Exploiting symmetry in variational quantum machine learning, PRX Quantum **4**, 010328 (2023).

[46] M. Larocca, F. Sauvage, F. M. Sbahi, G. Verdon, P. J. Coles, and M. Cerezo, Group-invariant quantum machine learning, PRX Quantum **3**, 030341 (2022).

[47] F. Sauvage, M. Larocca, P. J. Coles, and M. Cerezo, Building spatial symmetries into parametrized quantum circuits for faster training, Quantum Sci. Technol. **9** (2024).

[48] A. Skolik, M. Cattelan, S. Yarkoni, T. Bäck, and V. Dunjko, Equivariant quantum circuits for learning on weighted graphs, npj Quantum Inf. **9**, 47 (2023).

[49] T. Haug, K. Bharti, and M. S. Kim, Capacity and quantum geometry of parametrized quantum circuits, PRX Quantum **2**, 040309 (2021).

[50] K. Bu, D. E. Koh, L. Li, Q. Luo, and Y. Zhang, Rademacher complexity of noisy quantum circuits, arXiv:2103.03139.

[51] E. R. Anschuetz, A. Bauer, B. T. Kiani, and S. Lloyd, Efficient classical algorithms for simulating symmetric quantum systems, Quantum **7**, 1189 (2023).

[52] 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**, 1 (2021).

[53] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).

[54] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Provably efficient machine learning for quantum many-body problems, Science **377**, eabk3333 (2022).

[55] J. Liu, M. Liu, J.-P. Liu, Z. Ye, Y. Wang, Y. Alexeev, J. Eisert, and L. Jiang, Towards provably efficient quantum algorithms for large-scale machine-learning models, Nat. Commun. **15**, 434 (2024).

[56] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, Spectrally-normalized margin bounds for neural networks, Adv. Neural Inf. Process. Syst. **30**, 6240 (2017), http://papers.nips.cc/paper/7204-spectrally-normalized.

[57] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, Fantastic generalization measures and where to find them, arXiv:1912.02178.

[58] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, Fisher-Rao metric, geometry, and complexity of neural networks, in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, Barcelona, 2019), pp. 888–896.

[59] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, Theory of overparametrization in quantum neural networks, NATO ASI series Series F, Computer and system sciences **3**, 542 (2023).

[60] J. J. Meyer, Fisher information in noisy intermediate-scale quantum applications, Quantum **5**, 539 (2021).

[61] D. García-Martín, M. Larocca, and M. Cerezo, Effects of noise on the overparametrization of quantum neural networks, Phys. Rev. Res. **6**, 013295 (2024).

[62] R. Chakrabarti and H. Rabitz, Quantum control landscapes, Int. Rev. Phys. Chem. **26**, 671 (2007).

[63] M. S. Rudolph, S. Lerch, S. Thanasilp, O. Kiss, S. Vallecorsa, M. Grossi, and Z. Holmes, Trainability barriers and opportunities in quantum generative modeling, arXiv:2305.02881.

[64] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, Quantum **4**, 226 (2020).

[65] S. Jerbi, L. J. Fiderer, H. Poulsen Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Quantum machine learning beyond kernel methods, Nat. Commun. **14**, 517 (2023).

[66] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[67] M. A. Nielsen and I. L Chuang, *Quantum Computation and Quantum Information*, 10th Anniversary Edition (Cambridge Univ. Press, 2011).

[68] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.133.050603, which includes Refs. [69–75], for detailed derivations, as well as additional theoretical and numerical results.

[69] R. Cheng, Quantum geometric tensor (Fubini-Study metric) in simple quantum system: A pedagogical introduction, arXiv:1012.1337.

[70] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, Phys. Rev. X **7**, 021050 (2017).

[71] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. C. Benjamin, Theory of variational quantum simulation, Quantum **3**, 191 (2019).

[72] D. d'Alessandro, *Introduction to Quantum Control and Dynamics* (Chapman and Hall/CRC, London, 2021).

[73] J. Werschnik and E. Gross, Quantum optimal control theory, J. Phys. B **40**, R175 (2007).

[74] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, Optimal control of coupled spin dynamics: Design of nmr pulse sequences by gradient ascent algorithms, J. Magn. Reson. **172**, 296 (2005).

[75] E. Recio-Armengol, J. Eisert, and J. J. Meyer, Single-shot quantum machine learning, arXiv:2406.13812.

[76] J. Liu, H. Yuan, X.-M. Lu, and X. Wang, Quantum Fisher information matrix and multiparameter estimation, J. Phys. A **53**, 023001 (2020).

[77] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, Quantum optimally controlled transition landscapes, Science **303**, 1998 (2004).

[78] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Reinforcement learning in different phases of quantum control, Phys. Rev. X **8**, 031086 (2018).

[79] E. R. Anschuetz, Critical points in quantum generative models, arXiv:2109.06957.

[80] X. You, S. Chakrabarti, and X. Wu, A convergence theory for over-parametrized variational quantum eigensolvers, arXiv:2205.12481.

[81] J. Polcari, Representing unitary matrices by independent parameters, Technical Report (Working Paper, Rev 0, 2016).

[82] R. Fletcher, *Practical Methods of Optimization* (John Wiley & Sons, New York, 2013).

[83] J. R. Johansson, P. D. Nation, and F. Nori, Qutip: An open-source python framework for the dynamics of open quantum systems, Comput. Phys. Commun. **183**, 1760 (2012).

[84] R. Wiersema, E. Kökcü, A. F. Kemper, and B. N. Bakalov, Classification of dynamical lie algebras for translation-invariant 2-local spin systems in one dimension, arXiv:2309.05690.

[85] E. Kökcü, D. Camps, L. B. Oftelie, J. K. Freericks, W. A. de Jong, R. Van Beeumen, and A. F. Kemper, Algebraic compression of quantum circuits for Hamiltonian evolution, Phys. Rev. A **105**, 032420 (2022).

[86] E. Kökcü, T. Steckmann, Y. Wang, J. K. Freericks, E. F. Dumitrescu, and A. F. Kemper, Fixed depth Hamiltonian simulation via cartan decomposition, Phys. Rev. Lett. **129**, 070501 (2022).

[87] J. Kim, J. Kim, and D. Rosa, Universal effectiveness of high-depth circuits in variational eigenproblems, Phys. Rev. Res. **3**, 023203 (2021).

[88] E. Campos, A. Nasrallah, and J. Biamonte, Abrupt transitions in variational quantum circuit training, Phys. Rev. A **103**, 032607 (2021).

[89] J. Kim and Y. Oz, Quantum energy landscape and circuit optimization, Phys. Rev. A **106**, 052424 (2022).

[90] F. Sauvage, M. Larocca, P. J. Coles, and M. Cerezo, Building spatial symmetries into parametrized quantum circuits for faster training, Quantum Sci. Technol. **9**, 015029 (2024).

[91] E. Fontana, D. Herman, S. Chakrabarti, N. Kumar, R. Yalovetzky, J. Heredge, S. H. Sureshbabu, and M. Pistoia, The adjoint is all you need: Characterizing barren plateaus in quantum ansätze, arXiv:2309.07902.

[92] M. Ragone, B. N. Bakalov, F. Sauvage, A. F. Kemper, C. O. Marrero, M. Larocca, and M. Cerezo, A unified theory of barren plateaus for deep parametrized quantum circuits, arXiv:2309.09342.

[93] K. M. Nakanishi, K. Mitarai, and K. Fujii, Subspace-search variational quantum eigensolver for excited states, Phys. Rev. Res. **1**, 033062 (2019).

[94] T. Haug, Generalization with quantum geometry, https://github.com/txhaug/geometric_generalization.

[95] A. Mari, T. R. Bromley, and N. Killoran, Estimating the gradient and higher-order derivatives on quantum hardware, Phys. Rev. A **103**, 012405 (2021).

[96] J. Haferkamp, P. Faist, N. B. Kothakonda, J. Eisert, and N. Yunger Halpern, Linear growth of quantum circuit complexity, Nat. Phys. **18**, 528 (2022).

[97] M. Schuld and N. Killoran, Quantum machine learning in feature Hilbert spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[98] T. Haug, C. N. Self, and M. S. Kim, Quantum machine learning of large datasets using randomized measurements, Mach. Learn. **4**, 015005 (2023).

[99] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum natural gradient, Quantum **4**, 269 (2020).