


Effect of alternating layered *Ansätze* on trainability of projected quantum kernelsYudai Suzuki¹ and Muyuan Li² ¹*Department of Mechanical Engineering, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama 223-8522, Japan*²*IBM Quantum, IBM-MIT Watson AI Lab, Cambridge, Massachusetts 02142, USA*

(Received 9 April 2024; accepted 7 June 2024; published 1 July 2024)

Quantum kernel methods have been actively examined from both theoretical and practical perspectives due to the potential of quantum advantage in machine learning tasks. Despite a provable advantage of fine-tuned quantum kernels for specific problems, widespread practical usage of quantum kernel methods requires resolving the so-called vanishing similarity issue, where exponentially vanishing variance of the quantum kernels causes implementation infeasibility and trainability problems. In this work, we analytically and numerically investigate the vanishing similarity issue in projected quantum kernels with alternating layered *Ansätze*. We find that variance depends on circuit depth, the size of local unitary blocks, and the initial state, indicating the issue is avoidable if shallow alternating layered *Ansätze* are used and the initial state is not highly entangled. Our work provides some insight into design principles of projected quantum kernels and implies the need for caution when using highly entangled states as input to quantum kernel-based learning models.

DOI: [10.1103/PhysRevA.110.012409](https://doi.org/10.1103/PhysRevA.110.012409)**I. INTRODUCTION**

Recent advances in quantum devices and their public accessibility have led a number of researchers to explore the applicability of quantum computing in various fields. Machine learning is one such field where quantum computers can possibly enhance the capability of conventional methods. Remarkably, it has been shown that some quantum machine learning (QML) methods are theoretically guaranteed to outperform their existing classical counterparts for certain tasks [1–8]. Motivated by these works, QML approaches have also been heuristically examined with the hope to discover practical advantages over classical ones.

Quantum kernel methods are promising QML methods where the Hilbert space accessed by quantum computers is utilized as a feature space for machine learning tasks [9,10]. More specifically, quantum computers are used to map data into quantum feature space (i.e., the Hilbert space) via quantum circuits; then a quantum kernel, an inner product of a pair of data-dependent quantum features, is computed. The core idea is that the quantum kernel can measure the similarity between data points in the quantum feature space without explicitly determining the corresponding feature vectors that are exponentially large in the number of qubits. Much attention has been paid to quantum kernel methods because the provable advantage for a specific learning task has been shown [4] and supervised QML models can be recast in terms of kernel methods [11].

Despite the hope of quantum advantages for real-world machine learning tasks, it has been suggested that quantum kernel methods suffer from the so-called vanishing similarity issue or exponential concentration issue [12,13], which undermines implementation feasibility and trainability of quantum kernel-based learning models. Analogous to the well-known barren plateau problems in variational quantum algorithms [14–18], vanishing similarity is a phenomenon where the

expectation value and variance of the quantum kernel decay exponentially quickly in the number of qubits. As a result, output values of quantum kernels for any pairs of data points result in the same value, i.e., concentrated around the expectation value. This implies that an exponential number of measurement shots is needed to estimate each quantum kernel on quantum hardware. It also implies that models constructed from quantum kernels fail to distinguish between data points, leading to overfitting and poor performance for new unseen data [12,13].

Recent works have attempted to analytically clarify the phenomenon and seek a remedy to this issue. In particular, Ref. [13] analyzed the phenomenon for two types of fidelity-based quantum kernels, the commonly used fidelity-based quantum kernel [9] and projected quantum kernels [5]. In addition, four causes of the problem were elucidated in the literature: expressivity of quantum circuits, global measurement, how entangled the data-embedded quantum states are, and quantum noise. The analysis gives insight into design principles for quantum kernels. Scaling the rotation angles for data encoding gates could help avoid the issue at the cost of expressivity of quantum circuits [19–21]. Moreover, it has been shown that a new type of quantum kernel, called the quantum Fisher kernel, can mitigate the vanishing similarity issue because local similarities are measured via the information geometric quantity of quantum circuits [12].

In this work we further examine projected quantum kernels from the perspective of the vanishing similarity issue. As mentioned above, Ref. [13] analyzed projected quantum kernels for globally random quantum circuits and reached the conclusion that one cannot mitigate the exponential concentration for the quantum circuits. On the other hand, according to Ref. [18] on how to remedy the barren plateau problem, using local cost functions and the so-called alternating layered *Ansätze* (ALAs) possibly resolves vanishing gradients. This suggests a possibility that projected quantum kernels can alleviate the

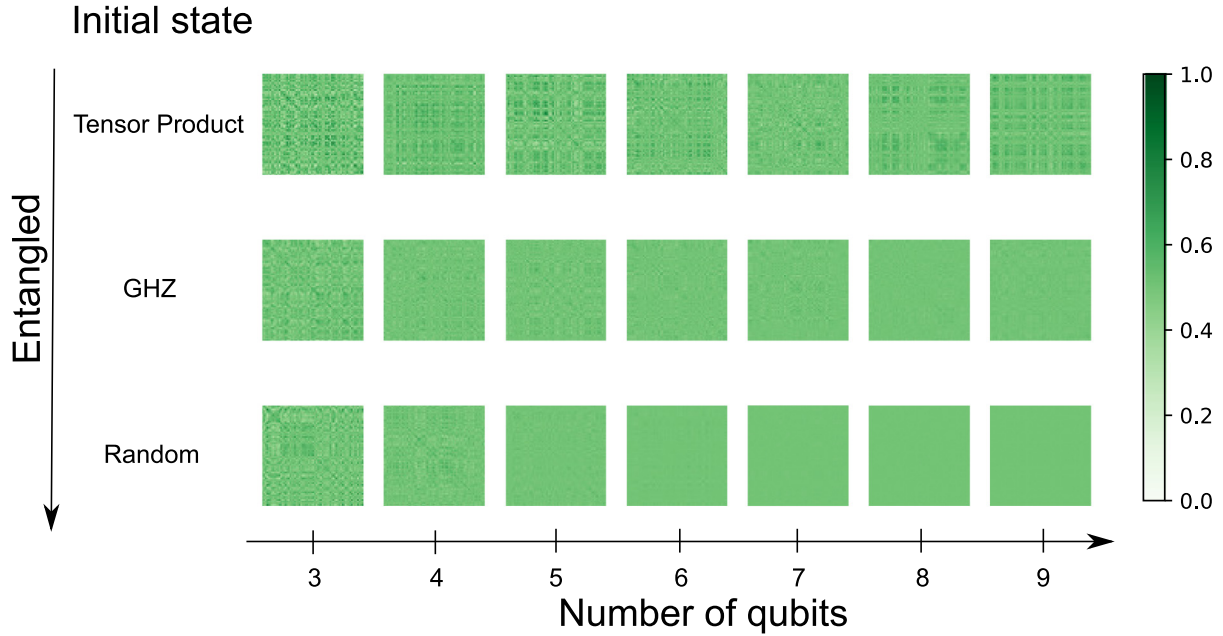


FIG. 1. Gram matrices of projected quantum kernels for different numbers of qubits and initial states. The Gram matrices, where the entry (i, j) contains the value of the quantum kernel for a data pair (x_i, x_j) , are computed using alternating layered *Ansätze* with depth $L = 6$ for 100 randomly generated data points. Here a tensor product state (top row), the GHZ state (middle row), and a quantum state randomly sampled from the Haar measure (bottom row) are prepared as the initial state. The more entangled the initial state is, the more identical every element of the Gram matrix is as the number of qubits increases; in other words, vanishing similarity arises.

issue because the difference of data is measured via a local quantity, i.e., reduced density matrices of the data-dependent quantum states. Therefore, this work analytically and numerically investigates the presence of the vanishing similarity issue in projected quantum kernels for different types of quantum circuits.

To be more specific, we provide analytical expressions for the expectation value and variance of projected quantum kernels using (1) n -qubit random quantum circuits and (2) the ALA with m -qubit local unitary blocks. We assume here that globally random quantum circuits and local unitary blocks in the ALAs form 2-designs [22–26]. With this assumption, the globally random quantum circuits fail to avoid the issue, as demonstrated in Ref. [13]. As for the ALAs, we find that the variance of projected quantum kernels depends on not only the circuit depth and size of the local unitary blocks, but also the initial state. This result indicates that the variance of the projected quantum kernel with shallow ALAs can avoid the vanishing similarity issue if the initial state is not highly entangled, such as a tensor product state. Figure 1 illustrates this result. Moreover, we observe a dependence on position of the reduced density matrices (accordingly, the light cone of the reduced subsystem) used to calculate projected quantum kernels. This suggests that the contribution of the term in the summed projected quantum kernels differs depending on the position of the subsystems. We then validate these analytical results by performing numerical simulation.

The rest of this paper is organized as follows. We provide an overview of quantum kernel methods and details of projected quantum kernels in Sec. II A. Then we elaborate the setting of our analysis in Sec. II B. Our main analytical results

on the vanishing similarity issue in projected quantum kernels is detailed in Sec. III A, which is followed by numerical simulation to demonstrate examples supporting the analytical results in Sec. III B. Section IV summarizes the paper and discusses the implication of our results. In Appendix A we provide the preliminaries of the integration over the Haar random unitary used in our analysis. In Appendixes B 1 and B 2 we provide the proof of our analytical results on the expectation value and variance of projected quantum kernels for the case of (1) n -qubit random quantum circuits and (2) alternating layered *Ansätze* with m -qubit local unitary blocks, respectively. In addition, we explain the difference between Eqs. (5) and (8) in the variance in Appendix B 3.

II. PRELIMINARIES

In this section we first review quantum kernel methods and provide the details of projected quantum kernels. We also introduce the settings in our analysis.

A. Quantum kernel methods

Quantum kernel methods measure the similarity between all possible pairs of data using a function called quantum kernel. Originally, a fidelity quantum kernel defined as

$$k_Q(x_i, x_j) = \text{Tr}[\rho(x_i, \theta)\rho(x_j, \theta)], \quad (1)$$

was proposed, where $\rho(x, \theta) = U(x, \theta)\rho_0U^\dagger(x, \theta)$ is the density matrix representation of the quantum state generated by applying a unitary operator $U(x, \theta)$ to the initial state ρ_0 . The unitary operator is realized by a quantum circuit dependent on

data \mathbf{x} and tunable parameters θ and plays the role of feature mapping, in which classical or quantum data are mapped to certain quantum states that have rich information on the data set. Note that we also introduce parameters θ , because such a quantum feature map can be engineered by optimizing θ in practical situations [27].

Then a Gram matrix G whose (i, j) element corresponds to a kernel function with an input pair $(\mathbf{x}_i, \mathbf{x}_j)$, i.e.,

$$G_{i,j} = k_Q(\mathbf{x}_i, \mathbf{x}_j),$$

is used to perform machine learning tasks. Typically, kernel methods are used for classification tasks in combination with support vector machines. The classification problem is reduced to minimizing the cost function $L(\alpha)$ with respect to the parameter α ,

$$L(\alpha) = -\sum_i^N \alpha_i + \frac{1}{2} \sum_{i,j}^N \alpha_i \alpha_j y_i y_j G_{ij}, \quad (2)$$

where N is the number of data points and $y_i \in \{+1, -1\}$ is the label of data \mathbf{x}_i . With optimal parameter α^{opt} obtained by solving Eq. (2), the prediction $y(\mathbf{x}_{\text{new}})$ of unseen data \mathbf{x}_{new} can be written as

$$y(\mathbf{x}_{\text{new}}) = \text{sgn} \left(\sum_i \alpha_i^{\text{opt}} y_i k_Q(\mathbf{x}_{\text{new}}, \mathbf{x}_i) \right). \quad (3)$$

While it has been proven that there exists a data set that is efficiently learnable not by classical models but by quantum kernels [4], the fidelity-based quantum kernel in Eq. (1) suffers from the vanishing similarity issue: The expectation and variance of the quantum kernel decline exponentially as the number of qubits increases. More concretely, the vanishing similarity issue is mathematically defined as

$$\text{Var}_{\{\mathbf{x}, \mathbf{x}'\}} [k_Q(\mathbf{x}, \mathbf{x}')] \leq B, \quad B \in O(c^{-n}), \quad (4)$$

with $c > 1$ and the number of qubits n . Here the variance is taken over all possible input data pairs $\{\mathbf{x}, \mathbf{x}'\}$. We remark that, as the quantum kernel depends on the data via a quantum feature map $U(\mathbf{x}, \theta)$, the variance can be equivalently taken over $\{U(\mathbf{x}, \theta), U(\mathbf{x}', \theta)\}$ sampled from a data- (and parameter-) dependent unitary ensemble, i.e., $\text{Var}_{\{U(\mathbf{x}, \theta), U(\mathbf{x}', \theta)\}} [k_Q(\mathbf{x}, \mathbf{x}')]$. The reason why this is detrimental is twofold [12,13]. One is that an exponential number of measurements must be done to precisely estimate the quantum kernel. The other is a trainability issue. The Gram matrix will be close to the identity matrix for a large number of qubits and thus the model of Eq. (3) obtained by minimizing the cost function in Eq. (2) would cause overfitting.

A possible remedy to this problem is projected quantum kernels proposed in Ref. [5], where a few variations were introduced. A simple one is a linear projected quantum kernel defined as

$$k_{\text{PQ}}^L(\mathbf{x}, \mathbf{x}') = \sum_{\kappa} \text{Tr} \{ \text{Tr}_{\bar{S}_{\kappa}} [\rho(\mathbf{x}, \theta)] \text{Tr}_{\bar{S}_{\kappa}} [\rho(\mathbf{x}', \theta)] \}, \quad (5)$$

where S_{κ} denotes the subspace for the κ th qubit and $\text{Tr}_{\bar{S}_{\kappa}}[\cdot]$ is the partial trace operation over the subspace \bar{S}_{κ} . Note that \bar{S} is the complement of the subspace S . Also, the Gaussian

projected quantum kernel is proposed,

$$k_{\text{PQ}}^G(\mathbf{x}, \mathbf{x}') = \exp \left(-\gamma \sum_{\kappa} \|\text{Tr}_{\bar{S}_{\kappa}} [\rho(\mathbf{x}, \theta)] - \text{Tr}_{\bar{S}_{\kappa}} [\rho(\mathbf{x}', \theta)]\|_2^2 \right), \quad (6)$$

with a hyperparameter $\gamma \in \mathbb{R}^+$ and the Hilbert–Schmidt norm $\|\cdot\|_2$. A key point of projected quantum kernels is that the similarity of data is measured using the reduced density matrix $\text{Tr}_{\bar{S}_{\kappa}}[\rho(\mathbf{x}, \theta)]$ instead of the density matrix $\rho(\mathbf{x}, \theta)$, namely, the local difference between data is compared in projected quantum kernels. According to Ref. [18], the barren plateau problem in variational quantum algorithms can be circumvented using local cost functions and the ALA. Similarly, projected quantum kernels also possess a local property that can help mitigate the vanishing similarity issue, which makes it favorable over traditional quantum kernels for practical applications.

B. Setting in our analysis

Although Ref. [13] demonstrates that projected quantum kernels with globally random quantum circuits cannot avoid the issue, it is a seemingly promising approach because of their locality. Thus, this work further analyzes projected quantum kernels from the vanishing similarity perspective, considering two types of quantum circuits. One is the n -qubit random quantum circuit and the other is the ALA with m -qubit local unitary blocks [18], as depicted in Figs. 2(a) and 2(b), respectively. We note that the former quantum circuit is the same setting in Ref. [13], but the latter has not been examined for use with projected quantum kernels. We perform analytical calculation for the globally random quantum circuits as well to make sure of the validity of our analysis and show an exact expression of the variance. For ease of analytical investigation, we then assume that the globally random quantum circuits and local unitary blocks in the ALAs are independent and 2-designs [22–26], meaning that the quantum circuits (unitary blocks) have the same statistical property with a Haar random unitary up to the second moment. In a broad sense, this assumption indicates that the quantum circuits or unitary blocks are expressive enough to uniformly explore the ensemble of Haar random states. We remark that, while quantum circuits might not be 2-designs in practice, some previous works have made similar assumptions to check the problems such as a barren plateau [14,18,28–30] and vanishing similarity [12,13,19]. Specifically, we express the ALA as

$$U(\mathbf{x}, \theta) = \prod_{d=1}^L V_d(\mathbf{x}, \theta) = \prod_{d=1}^L \left(\prod_{l=1}^{\zeta} W_{l,d}(\mathbf{x}, \theta_{l,d}) \right), \quad (7)$$

where L is the circuit depth and ζ is the number of unitary blocks in each layer. Here we assume that the total number of qubits n satisfies $n = m\zeta$. We note that the number of qubits on which both a unitary block in a layer and the one in the adjacent layer act is $m/2$; for example, $S_{(2,1)}$ and $S_{(1,2)}$ have an

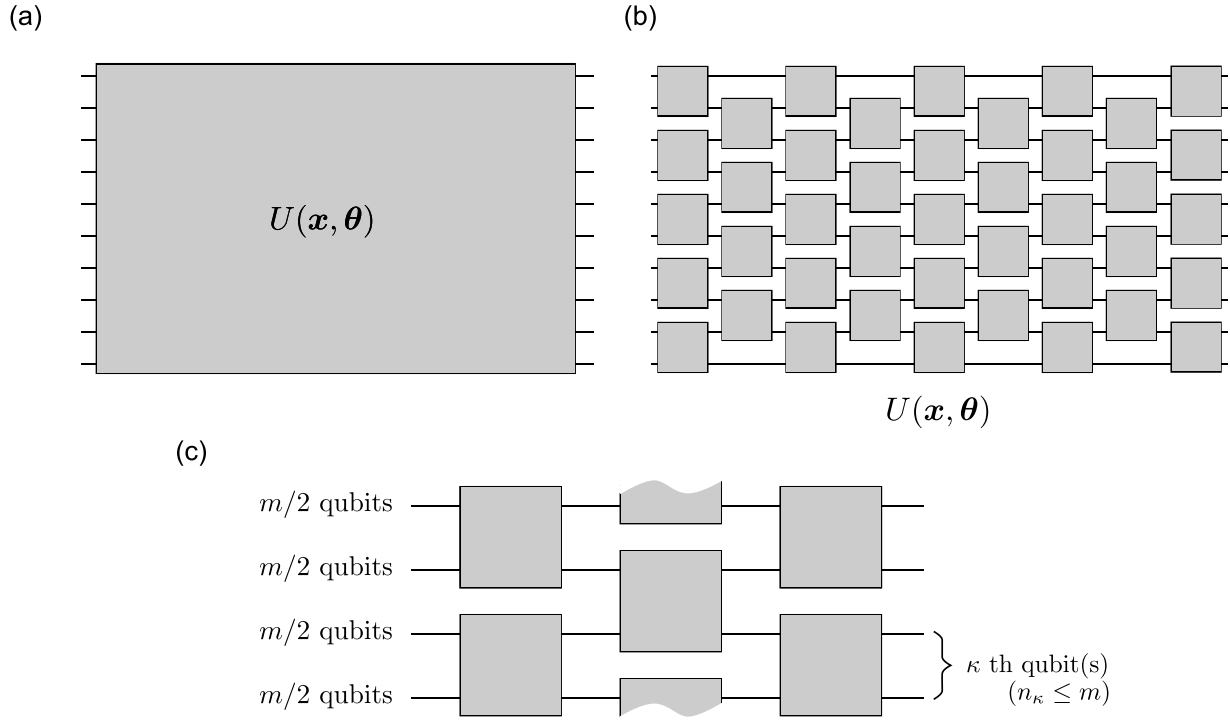


FIG. 2. Quantum circuits used in our analysis. The globally random quantum circuit is shown acting on (a) all qubits and (b) the ALA. (c) Details of the ALA and the setting of the projected quantum kernel in our analysis.

$m/2$ -qubit subspace in common, where $S_{(l,d)}$ is the subspace of qubits which the unitary block $W_{l,d}$ acts on. Details are illustrated in Fig. 2(c).

Throughout this paper, in lieu of Eqs. (5) and (6), we consider the quantity

$$k_{\text{PQ}}^{(\kappa)}(\mathbf{x}, \mathbf{x}') = \text{Tr}\{\text{Tr}_{S_{\kappa}}[\rho(\mathbf{x}, \boldsymbol{\theta})]\text{Tr}_{S_{\kappa}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\}. \quad (8)$$

We focus on this quantity because exploring it is sufficient to confirm the tendency of projected quantum kernels. Of course, the variance of Eq. (5) depends on the covariance terms and thus is not necessarily equal to that of the summation of Eq. (8) over possible κ . However, in this case, every covariance term is equal to or more than zero and the difference between them does not matter in terms of scaling (see Appendix B 3 for more details). Moreover, without loss of generality, we assume that the subspace for the κ th reduced density matrix (composed of n_{κ} qubits) appearing in Eq. (8), S_{κ} , is completely included in the subspace on which one of the unitary blocks in the last layer acts, as is shown in Fig. 2(c). We also assume that the initial state ρ_0 is an arbitrary pure state.

III. RESULTS

In what follows, we provide analytical results on the vanishing similarity issue in projected quantum kernels. Then we show numerical results to check the reliability of our analysis.

A. Main results

We analytically calculate the expectation value and variance of projected quantum kernels to check the existence of the vanishing similarity issue. Here we focus on two types of quantum circuits, that is, globally random quantum circuits and ALAs. Although the case for globally random quantum circuits has been analyzed in Ref. [13], here we check to confirm our analytical procedure and give an exact expression of the variance.

We first show analytical results for the globally random quantum circuits, with the full proof included in Appendix B 1.

Proposition 1. Let us denote the expectation value and variance of the projected quantum kernel defined in Eq. (8) with n -qubit random quantum circuits by $\langle k_{\text{PQ,RQC}}^{(\kappa)} \rangle$ and $\text{Var}(k_{\text{PQ,RQC}}^{(\kappa)})$, respectively. If the n -qubit random quantum circuits form t -designs with $t \geq 2$ and independent, then we have

$$\langle k_{\text{PQ,RQC}}^{(\kappa)} \rangle = \frac{1}{2^{n_{\kappa}}}, \quad (9)$$

$$\text{Var}(k_{\text{PQ,RQC}}^{(\kappa)}) = \frac{2^{2n_{\kappa}} - 1}{2^{2n_{\kappa}}(2^n + 1)^2} \approx \frac{1}{2^{2n}}. \quad (10)$$

We remind the reader that n_{κ} is the number of κ th qubits and n is the total number of qubits. Proposition 1 implies that the similarity between a pair of different data will be hard to distinguish regardless of the size of reduced density matrix for a large number of qubits. Therefore, projected quantum kernels with globally random quantum circuits cannot avoid the vanishing similarity issue. Note that the result is different from the result in Ref. [13] in that we calculate the exact

expectation rather than its upper bound, but the implication is consistent.

Next we provide the result obtained for the case of ALAs. We obtain here the lower bound of the variance to see the absence of the vanishing similarity issue. (See Appendix B 2 for the proof).

Theorem 1. For the projected quantum kernel defined in Eq. (8) and the ALA defined in Eq. (7), we denote the expectation value and variance by $\langle k_{\text{PQ,ALA}}^{(\kappa)} \rangle$ and $\text{Var}(k_{\text{PQ,ALA}}^{(\kappa)})$, respectively. Also, we assume that every unitary block in the ALA, $U(\mathbf{x}, \boldsymbol{\theta})$ and $U(\mathbf{x}', \boldsymbol{\theta})$, is a t -design with $t \geq 2$ and independent. Then the expectation value is

$$\langle k_{\text{PQ,ALA}}^{(\kappa)} \rangle = \frac{1}{2^{n_\kappa}}. \quad (11)$$

As for the variance, its lower bound is

$$\text{Var}(k_{\text{PQ,ALA}}^{(\kappa)}) \geq \frac{2^{2m(L-1)}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2(2^m + 1)^{4(L-1)}2^{2n_\kappa}} F(\rho_0, L), \quad (12)$$

with a function $F(\rho_0, L)$ of the initial state ρ_0 and the depth L . More specifically, we define the function as

$$F(\rho_0, L) = \left(2^m \sum_{h \in P(S_{(k_u,1)} : S_{(k_l,1)})} t_h \text{Tr}(\rho_{0,\bar{h}}^2) - \sum_{\tau=0}^{L-1} \frac{c_\tau}{2^{m\tau}} \right)^2, \quad (13)$$

where $t_h, c_\tau \in \mathbb{R}^+$ and $\rho_{0,\bar{h}} = \text{Tr}_{\bar{h}}(\rho_0)$ is the partial trace of the initial state over the subspace \bar{h} . Also, $P(S_{(k_u,1)} : S_{(k_l,1)})$ is the set containing all the possible neighboring subspaces in $\bigcup_{i=0}^{k_l-k_u} S_{(k_u+i,1)}$. Here $W_{k_u,1}$ ($W_{k_l,1}$) denotes the unitary block located at the upper (lower) edge of the light cone in the first layer. We note that $F(\rho_0, L) = 0$ if $\rho_{0,\bar{h}}$ is the completely mixed state for all subspaces h .

Like the case for a globally random quantum circuit, the expectation value is dependent not on the total number of qubits but on the system size of the reduced density matrix. However, Eq. (12) shows that the lower bound of the variance depends not only on the depth L and the size of the local unitary blocks m , but also on the initial state via the function $F(\rho_0, L)$. As shown in Eq. (13), the function contains the purity of some subspace of the initial state. Thus, depending on the choice of initial state, the vanishing similarity issue can be avoided. For example, if the initial state can be represented as a tensor product of arbitrary single-qubit states, i.e., $\rho_0 = \sigma_1 \otimes \sigma_2 \otimes \dots \otimes \sigma_n$ with arbitrary single-qubit states $\{\sigma_i\}$, then the function has a maximum value and the variance scales as $\Omega(2^{-2mL})$. On the other hand, if the initial state is so entangled that $\text{Tr}(\rho_{0,\bar{h}}^2)$ is the completely mixed state for almost all h , then the variance could decrease exponentially fast with respect to the number of qubits regardless of circuit depth. Note that it has been reported that the initial state matters for the vanishing gradient problem in variational quantum algorithms [31,32]. Thus, our result suggests that the initial state should also be taken into account for the usage of projected quantum kernels. Moreover, we check the dependence of the variance on the position of the κ th qubit. To be more specific, we consider the situations in which the κ th qubit(s) is (are) located (i)

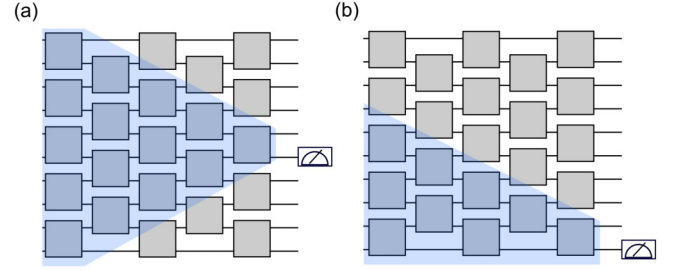


FIG. 3. Light cone depending on the position of the κ th qubit(s). The blue regions represent the light cone in the quantum circuits. (a) Case (i), where the number of local unitary blocks in the light cone is the largest. (b) Case (ii), with the smallest number of unitary blocks.

in the middle of the last layer and (ii) in the unitary block at the edge i.e., $W_{1,L}$ or $W_{\zeta,L}$, illustrated in Figs. 3(a) and 3(b), respectively. In addition, we assume that the initial state is a tensor product state to check the relationship between the depth and the position of the κ th qubit(s). In the first case, this is exactly the same as the result shown in Eq. (12), i.e., $\Omega(2^{-2mL})$. For the second case, as demonstrated in Appendix B 2, the variance is $\Omega(2^{-mL})$. The difference comes down to the number of unitary blocks in the light cone. This implies that reduced density matrices at the edge of the layer contribute to the linear projected quantum kernel in Eq. (5) more than the ones in the middle due to the quadratic difference. We remark that the dependence of the variance on the observables' position was argued in the context of variational quantum algorithms in Ref. [18], and the result we newly obtained here from the viewpoint of quantum kernel methods is similar to the statement shown in the literature (see Fig. 2 in the Supplementary Information of Ref. [18]). Moreover, we note that this result could suggest that the expressivity of models depends on the qubit positions, according to the connection between trainability issues and expressivity [29]. Although we do not believe that taking into account the position of observables helps to resolve such trainability problems, it would be interesting to explore the link of the expressivity and qubit positions of observables, which we leave for future work.

B. Numerical results

We perform numerical simulations to demonstrate examples that support our analytical results. In particular, we focus on the behavior of the variance for the ALA, because the one for the globally random quantum circuits has been analyzed in Ref. [13]. In the numerical experiments, ALAs with two-qubit local unitary blocks shown in Fig. 4 are considered, where we employ data reuploading techniques [33]. More specifically, each local unitary block consists of an embedding layer and the parametrized quantum circuit layers. Here we use rotation Y and Z gates as single-qubit rotation gates acting on the i th qubit, i.e., $R_{\sigma_i}(\beta) = \exp(-\beta\sigma_i/2)$, $\sigma_i \in \{Y_i, Z_i\}$, and the controlled- Z gate as an entangler. As for the input data, we set the number of qubits equal to the dimension of the data and each component is randomly chosen from the uniform distribution in the range $[-\pi, \pi)$. Analogously, each parameter in

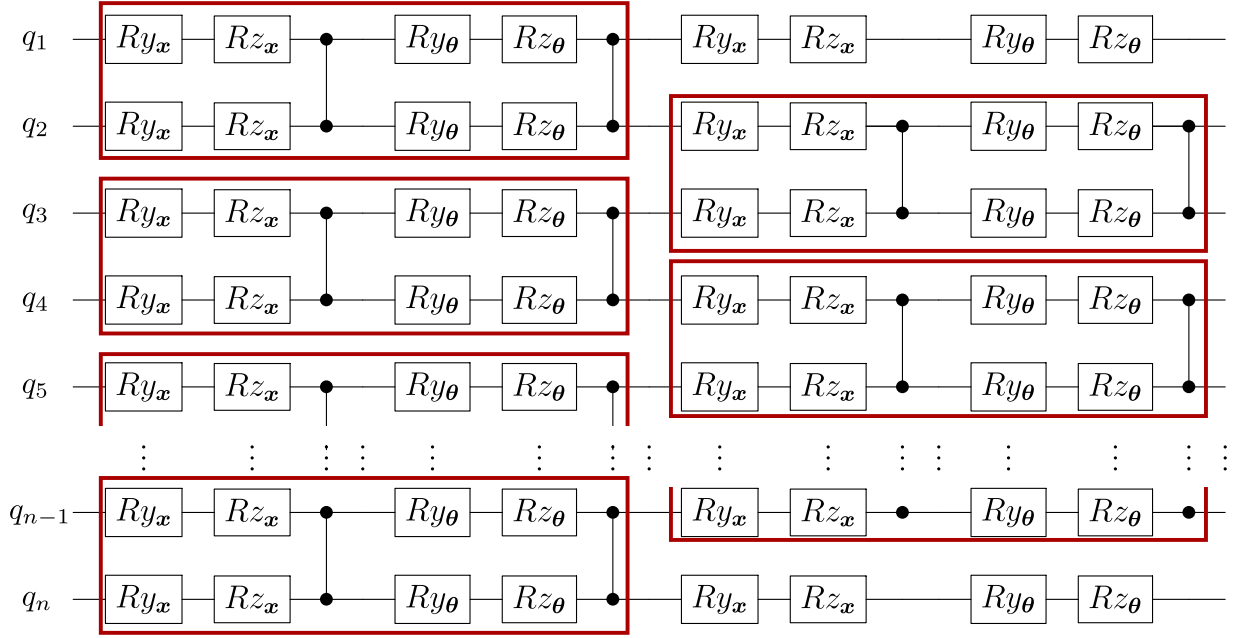


FIG. 4. Alternating layered *Ansatz* used in our simulation. As an example, here we show an even n -qubit alternating layered *Ansatz* with depth $L = 2$. The quantum circuit consists of two-qubit local unitary blocks denoted by red boxes, each of which has a data embedding layer and a parametrized quantum circuit layer. We note that Ry_a (Rz_a) represents a single-qubit rotation gate on the Y (Z) axis, whose angle is determined by a function of \mathbf{x} or $\boldsymbol{\theta}$ shown in the subscript. In the numerical experiments, the i th element of data x_i is encoded into single-qubit rotation gates (Ry and Rz) acting on the i th qubit in every embedding layer. Also, each parameter is assigned to different single-qubit rotation gates in the parametrized quantum circuit layers.

the parametrized quantum circuit layers is selected uniformly at random from the range $[-\pi, \pi)$. Then we prepare five sets of parameters and five data sets containing 50 data points to compute $k_{\text{pQ}}^{(\kappa)}(\mathbf{x}, \mathbf{x}')$ in Eq. (8) with $\mathbf{x} \neq \mathbf{x}'$. We note that $n_\kappa = 1$ for our numerical simulations. Afterward, the variance is calculated using the projected quantum kernels computed for different 25 settings of the input data set and the parameter set. The computation is performed for all possible κ . When we encode the data into the quantum circuit, the i th component of the input data, x_i , is injected into the angle of the single-qubit rotation gates acting on the i th qubit in every embedding layer, that is, $Ry_i(x_i)$ [$Rz_i(x_i)$]. We also assign each parameter to a single-qubit rotation gate in parametrized quantum circuit layers, namely, no parameters are shared with different rotation gates. Figure 4 depicts the details of the quantum circuit. The numerical simulation is performed using Qiskit [34].

We summarize here the numerical results from the following perspectives: (i) the dependence of the variance on circuit depth for different initial states, (ii) the dependence on the position of the κ th qubit, and (iii) the relation between the variance and the number of qubits n .

1. Dependence on circuit depth

Figure 5 shows the variance of projected quantum kernels as a function of the depth L for different initial states, where the number of qubits $n = 9$ and the reduced density matrix with respect to the fifth qubit are considered for three initial states: a tensor product state $\rho_0 = |0^{\otimes n}\rangle\langle 0^{\otimes n}|$, the Greenberger-Horne-Zeilinger (GHZ) state $\rho_0 = |\psi_{\text{GHZ}}\rangle\langle\psi_{\text{GHZ}}|$ with $|\psi_{\text{GHZ}}\rangle = 2^{-1/2}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$, and initial states randomly sampled from the Haar measure. We

choose these initial states with different degrees of entanglement to examine how entanglement of initial states affects the variance. As for the random initial states, we prepare five different states and the variance is averaged over the trials. It turns out that the variance decreases exponentially in circuit depth L for the case of the tensor product state and the GHZ

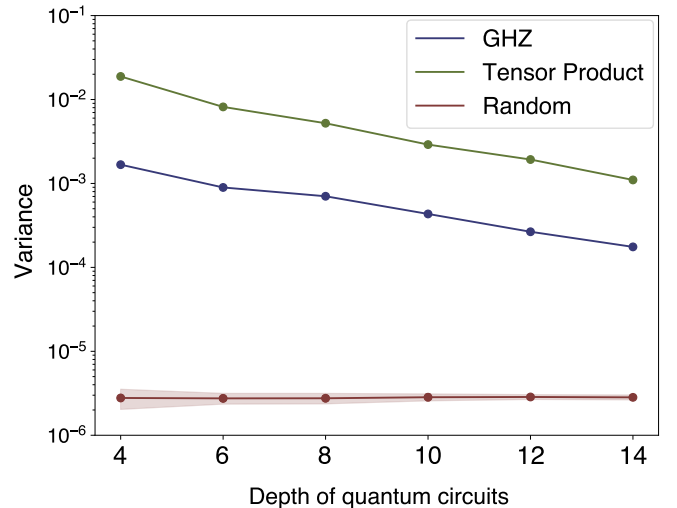


FIG. 5. Variance of the projected quantum kernel as a function of the depth of quantum circuits. Here we used the nine-qubit ALAs with depth $L \in \{4, 6, 8, 10, 12, 14\}$ and the reduced density matrix for the fifth qubit to compute the projected quantum kernel. We consider three initial states: a tensor product state (green), the GHZ state (blue), and random quantum states (red). The shaded region illustrates the standard deviation over five different random states.

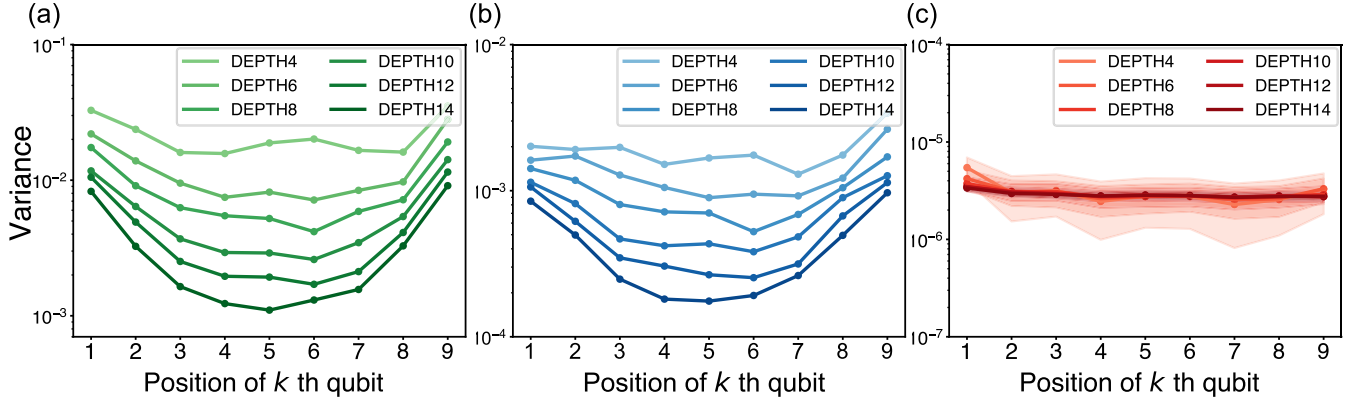


FIG. 6. Variance of the projected quantum kernel as a function of the position of the κ th qubit. We used the ALAs with nine qubits. The variances of the projected quantum kernel with depth $L \in \{4, 6, 8, 10, 12, 14\}$ are shown for the cases of (a) a tensor product state, (b) the GHZ state, and (c) random quantum states. In (c) the standard deviation is represented by the shaded region.

state. On the other hand, if a random quantum state is prepared as the initial state, the variance is independent of the depth and much smaller than the ones for other cases. This is consistent with the analytical result shown in Theorem 1. As demonstrated in Eq. (12), the variance is determined by the depth and the function of the initial state $F(\rho, L)$. For the first two cases, $F(\rho, L)$ does not contribute to the variance so much because the reduced systems of the initial states are far from the completely mixed states; the purity is 1 for the tensor product state over any subspace h and the purity is $1/2$ for the GHZ state if $\bar{h} \neq \emptyset$ or $h \neq \emptyset$ and otherwise 1. Thus, a term other than $F(\rho, L)$ comes into play; the variance vanishes exponentially with respect to the depth. However, the partial trace of a random quantum state can be close to the completely mixed state and thus $F(\rho, L)$ plays a significant role in the variance rather than the remaining term. Hence, the variance is consistently small regardless of the depth.

2. Dependence on positions of reduced subsystems

The variance as a function of the positions of the κ th qubit for the nine-qubit system is shown in Fig. 6. We notice that the variance of the reduced system at the edge of the layer is

smaller than that of the systems in the middle for the tensor product state and the GHZ state, shown in Figs. 6(a) and 6(b), respectively. Also, the gap of the variance between the systems at the edge and in the middle gets larger as the depth increases. This numerical result agrees with the statement in the previous section that the scaling of the variance differs depending on the number of local unitary blocks in the light cone and accordingly the position of the κ th qubit. As for the random quantum state case in Fig. 6(c), the depth and the position are less significant in the variance because the term $F(\rho, L)$ contributes dominantly.

3. Dependence on the total number of qubits

Figures 7(a)–7(c) show the variance for different numbers of qubits using a tensor product state, the GHZ state, and random quantum states, respectively. For the tensor product and the GHZ state, the variance levels off for all cases of circuit depth when the number of qubits is larger than a certain number. This is because the purity is constant for these cases and thus $F(\rho, L)$ is saturated. Thus, we can confirm that the variance of these cases is irrelevant to the number of qubits. However, Fig. 7(c) shows that the variance vanishes exponen-

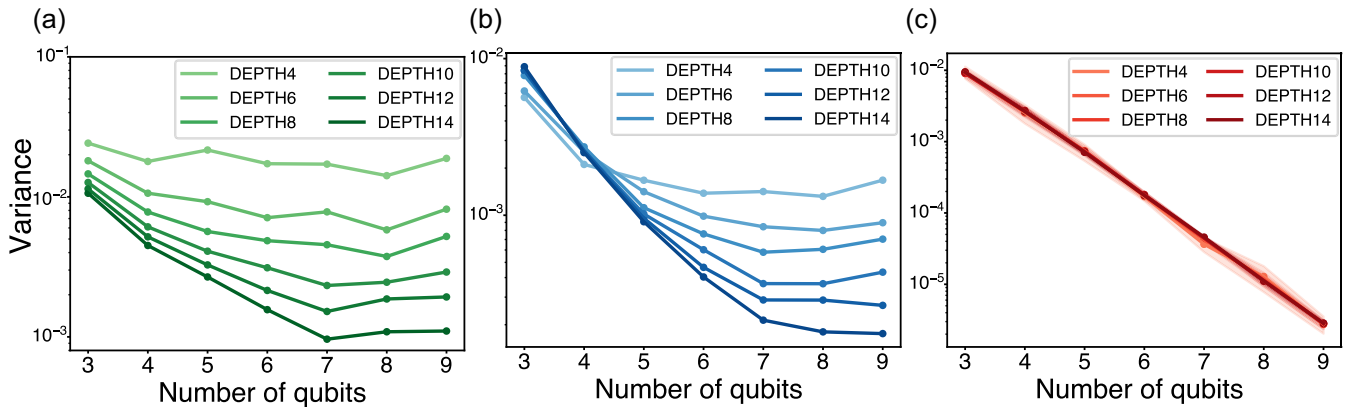


FIG. 7. Variance of the projected quantum kernel as a function of the number of qubits. We used different numbers of qubits $n \in \{3, 4, 5, 6, 7, 8, 9\}$ and the ALAs with different depths $L \in \{4, 6, 8, 10, 12, 14\}$. Here we consider the reduced system of the $\lceil n/2 \rceil$ th qubit, i.e., the qubit in the middle of the width. Variances are shown for (a) a tensor product state, (b) the GHZ state, and (c) random quantum states.

tially fast in the number of qubits. This would be attributed to the fact that there is an exponential decay in $F(\rho, L)$ with respect to the number of qubits. Hence, this indicates that initial states really matter in the variance of projected quantum kernels.

IV. DISCUSSION AND CONCLUSION

In this paper we investigated the vanishing similarity issue in projected quantum kernels from both analytical and numerical perspectives. We analytically showed that this issue is not avoidable for the case of globally random quantum circuits, which is consistent with results in Ref. [13]. In contrast, we found that projected quantum kernels with ALAs can avoid exponential decay of the variance if the quantum circuits are shallow and the initial state is not highly entangled. This implies the potential of projected quantum kernels for practical usability. In addition, we showed that the initial state plays a significant role in the variance scaling and thus caution is needed in preparing input states. We discuss the implication of our results in QML tasks below.

First, our results suggest that there is a caveat when quantum data are used as input states in QML tasks. Some QML tasks handle quantum states as an input state and then parametrized quantum circuits are applied to the state to seek features suitable for the tasks. In this situation, the initial state could be more entangled than a tensor product state. Hence, there is a possibility that the vanishing similarity issue for projected quantum kernels could be exacerbated for some tasks.

We also showed that the variance differs depending on the position of the reduced density matrix. Thus, the contribution to projected quantum kernels in Eqs. (5) and (6) of reduced systems at the edge of the layer is larger than that of systems in the middle; the tendency gets worse as we increase the circuit depth. This might result in a poor performance for some tasks because the relevant information could be undermined. Hence, in some situations, it would be better to consider the gap, for example, by modifying the weight of the projected quantum kernel for the κ th qubit(s).

Moreover, our results indicate a situation where classical shadows can reduce the quantum resources required to compute projected quantum kernels. Classical shadow is a technique to estimate properties of quantum states with a small number of measurement shots [35]. We note that the classical shadow technique can also be applied to estimate the quantum fidelity efficiently (see Ref. [35] for more details). Thus, some works have exploited the technique for efficient estimation of quantum kernel methods [5,36]. On the other hand, the classical shadow does not work when vanishing similarity arises. This is because the resolution needed to tell the difference in a pair of data through the quantum kernel is significantly high. Our Theorem 1 suggests that projected quantum kernels can utilize the power of classical shadows when shallow ALAs are used and the initial state is not highly entangled.

We finally remark that our analytical results are based on the assumption that quantum circuits and the local unitary blocks in the ALAs are 2-designs. This result is of significance in that we shed light on the trainability and limitations of

projected quantum kernels in general. On the other hand, as the no-free-lunch theorems [37–39] suggest, domain knowledge should be incorporated into the model. Actually, an emerging field called geometric quantum machine learning [40–44], where inductive bias such as symmetry is considered in constructing quantum models, has attracted much attention. Therefore, it would be worthwhile to explore the existence of the vanishing similarity issue by incorporating domain knowledge into the model for practical purposes. It would also be important to investigate advantages of projected quantum kernels for practical machine learning tasks handling quantum data as well as classical data.

ACKNOWLEDGMENTS

The authors thank Kunal Sharma, Ryan Sweke, Khadijeh Najafi, and Antonio Mezzacapo for stimulating discussions and comments on the manuscript. Part of this work was done when Y.S. was a research intern at IBM. Y.S. was supported by Grant-in-Aid for JSPS Fellows No. 22KJ2709.

APPENDIX A: PRELIMINARIES

We utilize formulas of integration over the Haar random unitary to calculate the expectation value and variance of projected quantum kernels (PQKs). Hence we here present lemmas related to the calculation.

Formulas of integrals over Haar random unitaries

Our analysis assumes that quantum circuits form t -designs [22–26]. When a quantum circuit W is a 1-design, i.e., the ensemble has the same statistical properties with the Haar random unitary up to the first moment, we can have the expression

$$\int d\mu(W) W_{i,j} W_{l,k}^* = \frac{\delta_{i,l} \delta_{j,k}}{d}, \quad (\text{A1})$$

where d is the dimension of the unitary W and $\delta_{i,j}$ represents the Kronecker delta. Similarly, we can exploit the following formula for the 2-design case:

$$\begin{aligned} \int d\mu(W) W_{i_1, j_1} W_{l_1, k_1}^* W_{i_2, j_2} W_{l_2, k_2}^* \\ = \frac{\delta_{i_1, l_1} \delta_{i_2, l_2} \delta_{j_1, k_1} \delta_{j_2, k_2} + \delta_{i_1, l_2} \delta_{i_2, l_1} \delta_{j_1, k_2} \delta_{j_2, k_1}}{d^2 - 1} \\ - \frac{\delta_{i_1, l_1} \delta_{i_2, l_2} \delta_{j_1, k_2} \delta_{j_2, k_1} + \delta_{i_1, l_2} \delta_{i_2, l_1} \delta_{j_1, k_1} \delta_{j_2, k_2}}{d(d^2 - 1)}. \end{aligned} \quad (\text{A2})$$

Furthermore, as we consider the alternating layered *Ansatz* in our analysis, we show below the five lemmas derived and shown in the Supplementary Information of Ref. [18]. In these lemmas, we define a unitary operator W acting on the Hilbert space \mathcal{H}_w and W' acting on the bipartite system $\mathcal{H}_{w_1} \otimes \mathcal{H}_{w_2}$ as

$$W = \sum_{i,j} W_{i,j} |i\rangle\langle j|, \quad W' = \sum_{i_1, j_1, i_2, j_2} W'_{i_1, j_1, i_2, j_2} |i_1 i_2\rangle\langle j_1 j_2|. \quad (\text{A3})$$

Lemma 1. Let a unitary W acting on the d -dimensional Hilbert space \mathcal{H}_w be a t -design with $t \geq 1$. Then, for arbitrary

operators $A, B : \mathcal{H}_w \rightarrow \mathcal{H}_w$, we have

$$\sum_i p_i \text{Tr}(W_i A W_i^\dagger B) = \int d\mu(W) \text{Tr}(W A W^\dagger B) = \frac{\text{Tr}(A) \text{Tr}(B)}{d}. \quad (\text{A4})$$

Lemma 2. Let a unitary W acting on the d -dimensional Hilbert space \mathcal{H}_w be a t -design with $t \geq 2$. Then, for arbitrary operators $A, B, C, D : \mathcal{H}_w \rightarrow \mathcal{H}_w$, we have

$$\begin{aligned} & \sum_i p_i \text{Tr}(W_i A W_i^\dagger B W_i C W_i^\dagger D) \\ &= \int d\mu(W) \text{Tr}(W A W^\dagger B W C W^\dagger D) \\ &= \frac{1}{d^2 - 1} [\text{Tr}(A) \text{Tr}(C) \text{Tr}(B D) + \text{Tr}(A C) \text{Tr}(B) \text{Tr}(D)] \\ &\quad - \frac{1}{d(d^2 - 1)} [\text{Tr}(A) \text{Tr}(B) \text{Tr}(C) \text{Tr}(D) + \text{Tr}(A C) \text{Tr}(B D)]. \end{aligned} \quad (\text{A5})$$

Lemma 3. Let a unitary W on the d -dimensional Hilbert space \mathcal{H}_w be a t -design with $t \geq 2$. Then, for arbitrary operators $A, B, C, D : \mathcal{H}_w \rightarrow \mathcal{H}_w$, we have

$$\begin{aligned} & \sum_i p_i \text{Tr}(W_i A W_i^\dagger B) \text{Tr}(W_i C W_i^\dagger D) \\ &= \int d\mu(W) \text{Tr}(W A W^\dagger B) \text{Tr}(W C W^\dagger D) \\ &= \frac{1}{d^2 - 1} [\text{Tr}(A) \text{Tr}(B) \text{Tr}(C) \text{Tr}(D) + \text{Tr}(A C) \text{Tr}(B D)] \\ &\quad - \frac{1}{d(d^2 - 1)} [\text{Tr}(A) \text{Tr}(C) \text{Tr}(B D) + \text{Tr}(A C) \text{Tr}(B) \text{Tr}(D)]. \end{aligned} \quad (\text{A6})$$

Lemma 4. Let a unitary W acting on the d_w -dimensional Hilbert space \mathcal{H}_w be a t -design with $t \geq 2$. In addition, suppose $\mathcal{H} = \mathcal{H}_{\bar{w}} \otimes \mathcal{H}_w$ is $d_w d_{\bar{w}}$ dimensional. Then, for arbitrary operators $A, B : \mathcal{H} \rightarrow \mathcal{H}$, we have

$$\begin{aligned} & \sum_i p_i (\mathbb{I}_{\bar{w}} \otimes W_i) A (\mathbb{I}_{\bar{w}} \otimes W_i^\dagger) B \\ &= \int d\mu(W) (\mathbb{I}_{\bar{w}} \otimes W) A (\mathbb{I}_{\bar{w}} \otimes W^\dagger) B \\ &= \frac{\text{Tr}_w(A) \otimes \mathbb{I}_w B}{d_w} \end{aligned} \quad (\text{A7})$$

and

$$\begin{aligned} & \sum_i p_i \text{Tr}[(\mathbb{I}_{\bar{w}} \otimes W_i) A (\mathbb{I}_{\bar{w}} \otimes W_i^\dagger) B] \\ &= \int d\mu(W) \text{Tr}[(\mathbb{I}_{\bar{w}} \otimes W) A (\mathbb{I}_{\bar{w}} \otimes W^\dagger) B] \\ &= \frac{1}{d_w} \text{Tr}[\text{Tr}_w(A) \text{Tr}_w(B)]. \end{aligned} \quad (\text{A8})$$

Here \mathbb{I}_w ($\mathbb{I}_{\bar{w}}$) represents the identity matrix acting on the Hilbert space \mathcal{H}_w ($\mathcal{H}_{\bar{w}}$) and the partial trace over \mathcal{H}_w ($\mathcal{H}_{\bar{w}}$) is denoted by Tr_w ($\text{Tr}_{\bar{w}}$). Also, \bar{A} denotes the complement of A .

Lemma 5. Let W be a unitary operator acting on the d_w -dimensional Hilbert space \mathcal{H}_w . In addition, suppose $\mathcal{H} = \mathcal{H}_{\bar{w}} \otimes \mathcal{H}_w$ is $d_w d_{\bar{w}}$ dimensional with $d_w = 2^m$ and $d_{\bar{w}} = 2^{n-m}$. Then, for arbitrary operators $A, B : \mathcal{H} \rightarrow \mathcal{H}$, we have

$$\text{Tr}[(\mathbb{I}_{\bar{w}} \otimes W) A (\mathbb{I}_{\bar{w}} \otimes W^\dagger) B] = \sum_{p,q} \text{Tr}(W A_{qp}, W^\dagger B_{pq}) \quad (\text{A9})$$

where

$$A_{qp} = \text{Tr}_{\bar{w}}[(|p\rangle\langle q| \otimes \mathbb{I}_w) A], \quad B_{pq} = \text{Tr}_{\bar{w}}[(|q\rangle\langle p| \otimes \mathbb{I}_w) B]. \quad (\text{A10})$$

Here q and p represent bit strings of length $n - m$.

APPENDIX B: VANISHING SIMILARITY ISSUE IN PROJECTED QUANTUM KERNELS

In this Appendix we analytically derive the expectation value and variance of PQKs for two types of quantum circuits, i.e., globally random quantum circuits acting on all n qubits and the ALA. The PQK we consider in our analysis is [5]

$$k_{\text{PQ}}^{(\kappa)}(\mathbf{x}, \mathbf{x}') = \text{Tr}\{\text{Tr}_{\bar{S}_\kappa}[\rho(\mathbf{x}, \boldsymbol{\theta})] \text{Tr}_{\bar{S}_\kappa}[\rho(\mathbf{x}', \boldsymbol{\theta})]\}, \quad (\text{B1})$$

where $\rho(\mathbf{x}, \boldsymbol{\theta}) = U(\mathbf{x}, \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}, \boldsymbol{\theta})$, with initial state ρ_0 and the input- and parameter-dependent unitary operator $U(\mathbf{x}, \boldsymbol{\theta})$; $\text{Tr}_{\bar{S}_\kappa}[\cdot]$ is the partial trace over the subspace \bar{S}_κ . Also, the number of κ th qubits is denoted by n_κ . In our analysis, we assume that S_κ is completely included in the subspace on which one of the unitary blocks in the last layer of the ALA acts, as is shown in Fig. 2(c). We also assume that the initial state ρ_0 is an arbitrary pure state. Finally, we will state the difference of the variance between Eq. (B1) and the linear PQK in Appendix B 3.

1. Case 1: Globally random quantum circuits

Here we calculate the expectation value and variance of the PQK in Eq. (B1), considering the n -qubit random quantum circuits.

a. Expectation value

We derive the expectation value of the PQK. We assume that either $U(\mathbf{x}, \boldsymbol{\theta})$ or $U(\mathbf{x}', \boldsymbol{\theta})$ is a t -design with $t \geq 1$ without loss of generality. We utilize here the symmetry of the PQK in Eq. (B1). In particular, we assume that $U(\mathbf{x}, \boldsymbol{\theta})$ is a t -design with $t \geq 1$. Then the expectation value of the PQK over the Haar random unitary, $\langle k_{\text{PQ}}^{(\kappa)} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})}$, is calculated as

$$\begin{aligned} \langle k_{\text{PQ}}^{(\kappa)} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})} &= \langle \text{Tr}\{\text{Tr}_{\bar{S}_\kappa}[U(\mathbf{x}, \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}, \boldsymbol{\theta})] \text{Tr}_{\bar{S}_\kappa}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})} \\ &= \frac{1}{2^n} \text{Tr}\{\text{Tr}_{\bar{S}_\kappa}(\mathbb{I}) \text{Tr}_{\bar{S}_\kappa}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \\ &= \frac{2^{n-n_\kappa}}{2^n} \text{Tr}[\rho(\mathbf{x}', \boldsymbol{\theta})] = \frac{1}{2^{n_\kappa}}, \end{aligned} \quad (\text{B2})$$

where Lemma 1 is used for the second equality and the property of the density matrix, i.e., $\text{Tr}(\rho) = 1$, is utilized for the last equality.

b. Variance

Next we calculate the variance. The variance $\text{Var}(k_{\text{PQ}}^{(\kappa)})$ is expressed as $\text{Var}(k_{\text{PQ}}^{(\kappa)}) = \langle k_{\text{PQ}}^{(\kappa)2} \rangle - \langle k_{\text{PQ}}^{(\kappa)} \rangle^2$. As we already have $\langle k_{\text{PQ}}^{(\kappa)} \rangle = 1/2^{2n_\kappa}$, we focus on $\langle k_{\text{PQ}}^{(\kappa)2} \rangle$. Here we assume that $U(\mathbf{x}, \boldsymbol{\theta})$ and $U(\mathbf{x}', \boldsymbol{\theta})$ are t -designs with $t \geq 2$. Due to

the independence of $U(\mathbf{x}, \boldsymbol{\theta})$ and $U(\mathbf{x}', \boldsymbol{\theta})$ from our assumptions, the expectation value can be obtained by integrating the square of the PQK over these unitaries, that is, $\langle k_{\text{PQ}}^{(\kappa)2} \rangle = \langle k_{\text{PQ}}^{(\kappa)2} \rangle_{U(\mathbf{x}, \boldsymbol{\theta}), U(\mathbf{x}', \boldsymbol{\theta})}$. Thus, we first calculate the expectation value over $U(\mathbf{x}, \boldsymbol{\theta})$, i.e., $\langle k_{\text{PQ}}^{(\kappa)2} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})}$. Then we obtain

$$\begin{aligned}
\langle k_{\text{PQ}}^{(\kappa)2} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})} &= \langle \text{Tr} \{ \text{Tr}_{S_\kappa} [U(\mathbf{x}, \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}, \boldsymbol{\theta})] \text{Tr}_{S_\kappa} [\rho(\mathbf{x}', \boldsymbol{\theta})] \} \text{Tr} \{ \text{Tr}_{S_\kappa} [U(\mathbf{x}, \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}, \boldsymbol{\theta})] \text{Tr}_{S_\kappa} [\rho(\mathbf{x}', \boldsymbol{\theta})] \} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})} \\
&= \langle \text{Tr} \{ [U(\mathbf{x}, \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}, \boldsymbol{\theta}) \otimes \rho(\mathbf{x}', \boldsymbol{\theta})] \text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}} \} \\
&\quad \times \text{Tr} \{ [U(\mathbf{x}, \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}, \boldsymbol{\theta}) \otimes \rho(\mathbf{x}', \boldsymbol{\theta})] \text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}} \} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})} \\
&= \frac{1}{2^{2n} - 1} \{ 2^{2(n-n_\kappa)} \text{Tr}(\rho_0) \text{Tr}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) \text{Tr}(\rho_0) \text{Tr}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) + 2^{n-n_\kappa} \text{Tr}(\rho_0^2) \text{Tr}[\text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) \text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}})] \} \\
&\quad - \frac{1}{2^n(2^{2n} - 1)} \{ 2^{n-n_\kappa} \text{Tr}(\rho_0) \text{Tr}(\rho_0) \text{Tr}[\text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) \text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}})] + 2^{2(n-n_\kappa)} \text{Tr}(\rho_0^2) \text{Tr}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) \text{Tr}[\rho_{\mathbf{x}', \boldsymbol{\theta}}] \} \\
&= \frac{2^{n-n_\kappa}}{2^n(2^n + 1)} \{ 2^{n-n_\kappa} + \text{Tr}[\text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) \text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}})] \}. \tag{B3}
\end{aligned}$$

In the second equality, we utilize the fact that

$$\text{Tr}[\text{Tr}_S(A) \text{Tr}_S(B)] = \text{Tr}[(A \otimes B) \text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{S_1 \otimes S_2}], \tag{B4}$$

where A and B are arbitrary matrices and $\mathbb{I}_{S_1 \otimes S_2}$ and SWAP_{S_1, S_2} denote the identity operator and the SWAP operator acting on systems S_1 and S_2 , respectively. Note that the subspace labeled with the number in the subscript, i.e., $i \in \{1, 2\}$ in S_i , dictates one of the duplicated subsystems. In the third equality, we use the result

$$\begin{aligned}
&\langle \text{Tr}[\text{Tr}_S(wAw^\dagger) \text{Tr}_S(B)] \text{Tr}[\text{Tr}_S(wAw^\dagger) \text{Tr}_S(B)] \rangle_w \\
&= \langle \text{Tr}[(wAw^\dagger \otimes B) \text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{S_1 \otimes S_2}] \text{Tr}[(wAw^\dagger \otimes B) \text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{S_1 \otimes S_2}] \rangle_w \\
&= \frac{1}{d^2 - 1} \left[\left(\frac{d}{\dim(S)} \right)^2 \text{Tr}(A) \text{Tr}(B) \text{Tr}(A) \text{Tr}(B) + \frac{d}{\dim(S)} \text{Tr}(A^2) \text{Tr}[\text{Tr}_S(B) \text{Tr}_S(B)] \right] \\
&\quad - \frac{1}{d(d^2 - 1)} \left[\frac{d}{\dim(S)} \text{Tr}(A) \text{Tr}(A) \text{Tr}[\text{Tr}_S(B) \text{Tr}_S(B)] + \left(\frac{d}{\dim(S)} \right)^2 \text{Tr}(A^2) \text{Tr}(B) \text{Tr}(B) \right], \tag{B5}
\end{aligned}$$

with the arbitrary matrices A and B of size d and a $d \times d$ unitary matrix w . Here $\dim(S)$ represents the dimension of the space S . We note that Eq. (B5) can be obtained using Eq. (A2) and the property of SWAP operators regarding the trace operation, i.e., $\text{Tr}(\text{SWAP}_{S_1, S_2}) = \dim(S)$, where $\dim(S_1) = \dim(S_2) = \dim(S)$. Also, in the last equality, the property of the pure state, i.e., $\text{Tr}(\rho) = \text{Tr}(\rho^2) = 1$, is used.

In Eq. (B3), only the second term in the last equality depends on $U(\mathbf{x}', \boldsymbol{\theta})$. Then the expectation value of the term can be calculated as

$$\begin{aligned}
&\langle \text{Tr}[\text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}}) \text{Tr}_{S_\kappa}(\rho_{\mathbf{x}', \boldsymbol{\theta}})] \rangle_{U(\mathbf{x}', \boldsymbol{\theta})} \\
&= \langle \text{Tr} \{ [U(\mathbf{x}', \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}', \boldsymbol{\theta}) \otimes U(\mathbf{x}', \boldsymbol{\theta}) \rho_0 U^\dagger(\mathbf{x}', \boldsymbol{\theta})] \text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}} \} \rangle_{U(\mathbf{x}', \boldsymbol{\theta})} \\
&= \frac{1}{2^{2n} - 1} \{ \text{Tr}(\rho_0) \text{Tr}(\rho_0) \text{Tr}(\text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}}) + \text{Tr}(\rho_0^2) \text{Tr}[\text{SWAP}_{S_{\kappa_1} \cup S_{\kappa_1}, S_{\kappa_2} \cup S_{\kappa_2}} (\text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}})] \} \\
&\quad - \frac{1}{2^n(2^{2n} - 1)} \{ \text{Tr}(\rho_0) \text{Tr}(\rho_0) \text{Tr}[\text{SWAP}_{S_{\kappa_1} \cup S_{\kappa_1}, S_{\kappa_2} \cup S_{\kappa_2}} (\text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}})] + \text{Tr}(\rho_0^2) \text{Tr}(\text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{S_{\kappa_1} \otimes S_{\kappa_2}}) \} \\
&= \frac{1}{2^{2n} - 1} \left(1 - \frac{1}{2^n} \right) (2^{2n-n_\kappa} + 2^{n+n_\kappa}) = \frac{1}{2^n + 1} (2^{n-n_\kappa} + 2^{n_\kappa}), \tag{B6}
\end{aligned}$$

where Eqs. (B4) and (A2) are used for the first and second equalities and the property of the SWAP operator with respect to the trace operation is utilized for the third equality. Therefore, we can obtain

$$\langle k_{\text{PQ}}^{(\kappa)2} \rangle = \frac{2^{n-n_\kappa}}{2^n(2^n + 1)} \left(2^{n-n_\kappa} + \frac{1}{2^n + 1} (2^{n-n_\kappa} + 2^{n_\kappa}) \right). \tag{B7}$$

As a result, we have

$$\begin{aligned}\text{Var}(k_{\text{PQ}}^{(\kappa)}) &= \langle k_{\text{PQ}}^{(\kappa)2} \rangle - \langle k_{\text{PQ}}^{(\kappa)} \rangle^2 \\ &= \frac{2^{n-n_\kappa}}{2^n(2^n+1)} \left(2^{n-n_\kappa} + \frac{1}{2^n+1} (2^{n-n_\kappa} + 2^{n_\kappa}) \right) - \frac{1}{2^{2n_\kappa}} \\ &= \frac{2^{2n_\kappa} - 1}{2^{2n_\kappa} (2^n + 1)^2}.\end{aligned}\quad (\text{B8})$$

We note that the same result can be obtained for the case in which different initial states are prepared for $\rho(\mathbf{x}, \boldsymbol{\theta})$ and $\rho(\mathbf{x}', \boldsymbol{\theta})$.

2. Case 2: Alternating layered *Ansätze*

In what follows, we calculate the expectation value and variance of the PQQ in Eq. (B1) considering the ALA.

a. Expectation value

We note that expectation value $\langle k_{\text{PQ}}^{(\kappa)} \rangle_{U(\mathbf{x}, \boldsymbol{\theta})}$ can be obtained by integrating the quantity over every unitary block, that is, $\langle k_{\text{PQ}}^{(\kappa)} \rangle_{W_{1,1}(\mathbf{x}, \boldsymbol{\theta}), W_{2,1}(\mathbf{x}, \boldsymbol{\theta}), \dots, W_{\zeta,L}(\mathbf{x}, \boldsymbol{\theta}_{\zeta,L})}$. Thus, we start with the integration over the unitary block in the last layer that acts on the κ th qubit(s), which we denote by \tilde{W} . Then we obtain

$$\begin{aligned}\langle k_{\text{PQ}}^{(\kappa)} \rangle_{\tilde{W}} &= \langle \text{Tr} \{ \text{Tr}_{\tilde{S}_\kappa} (\tilde{W} \rho_{x,r} \tilde{W}^\dagger) \text{Tr}_{\tilde{S}_\kappa} [\rho(\mathbf{x}', \boldsymbol{\theta})] \} \rangle_{\tilde{W}} = \langle \text{Tr} \{ [\tilde{W} \rho_{x,r} \tilde{W}^\dagger \otimes \rho(\mathbf{x}', \boldsymbol{\theta})] \text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{\tilde{S}_{\kappa_1} \otimes \tilde{S}_{\kappa_2}} \} \rangle_{\tilde{W}} \\ &= \frac{1}{2^m} \text{Tr} \{ [\text{Tr}_{S_{\tilde{W}}} (\rho_{x,r}) \otimes \rho(\mathbf{x}', \boldsymbol{\theta})] \text{Tr}_{S_{\tilde{W}}} (\text{SWAP}_{S_{\kappa_1}, S_{\kappa_2}} \otimes \mathbb{I}_{\tilde{S}_{\kappa_1} \otimes \tilde{S}_{\kappa_2}}) \} \\ &= \frac{1}{2^m} \text{Tr} \left([\text{Tr}_{S_{\tilde{W}}} (\rho_{x,r}) \otimes \rho(\mathbf{x}', \boldsymbol{\theta})] \frac{2^m}{2^{n_\kappa}} \mathbb{I} \right) = \frac{1}{2^{n_\kappa}} \text{Tr}(\rho_{x,r}) \text{Tr}[\rho(\mathbf{x}', \boldsymbol{\theta})] = \frac{1}{2^{n_\kappa}},\end{aligned}\quad (\text{B9})$$

where Lemma 4 is used for the third equality, the property of the SWAP operator for the trace operation is used for the fourth equality, and the property of the density matrix, i.e., $\text{Tr}(\rho) = 1$, is utilized for the last equality. Here $\rho_{x,r}$ is the quantum state resulting from the initial state to which the unitary operator $U(\mathbf{x}, \boldsymbol{\theta})$, except for \tilde{W} , is applied, namely, the equality $\rho(\mathbf{x}, \boldsymbol{\theta}) = \tilde{W} \rho_{x,r} \tilde{W}^\dagger$ holds. As is dictated in Eq. (B9), the rest of the unitary blocks in the ALA do not contribute to the calculation of the expectation value, namely, the unitary blocks can be canceled out in terms of the calculation after the integration over \tilde{W} . Thus the expectation value reads

$$\langle k_{\text{PQ}}^{(\kappa)} \rangle = \langle k_{\text{PQ}}^{(\kappa)} \rangle_{\tilde{W}} = \frac{1}{2^{n_\kappa}}. \quad (\text{B10})$$

b. Variance

Finally, we compute the variance of the PQQ for ALAs. As the variance can be written as $\text{Var}(k_{\text{PQ}}^{(\kappa)}) = \langle k_{\text{PQ}}^{(\kappa)2} \rangle - \langle k_{\text{PQ}}^{(\kappa)} \rangle^2$, we again focus on the quantity $\langle k_{\text{PQ}}^{(\kappa)2} \rangle$. Also, analogously to the calculation for the expectation value, we integrate the quantity over all local unitary blocks in the ALAs, $U(\mathbf{x}, \boldsymbol{\theta})$ and $U(\mathbf{x}', \boldsymbol{\theta})$. In particular, we begin with the integration over the unitary blocks in the last layer. Without loss of generality, we assume that the unitary block in the last layer that acts on the κ th qubit(s) is the p th unitary block in the last layer, i.e., $W_{p,L}(\mathbf{x}, \boldsymbol{\theta}_{p,L})$. Moreover, for the sake of clarity, we define $W_{l,d}(\mathbf{x}, \boldsymbol{\theta}_{l,d}) \equiv W_{l,d}[\mathbf{x}, \boldsymbol{\theta}_{l,d}] \equiv W_{l,d}'$ hereafter.

To obtain the expectation value over the unitary blocks in the last layer, we have to calculate the integration of the following quantity repeatedly:

$$\langle \text{Tr}[\text{Tr}_{\tilde{S}}(wAw^\dagger) \text{Tr}_{\tilde{S}}(B)] \text{Tr}[\text{Tr}_{\tilde{S}}(wAw^\dagger) \text{Tr}_{\tilde{S}}(B)] \rangle_w. \quad (\text{B11})$$

Thus, based on the calculation to be performed below, we consider the following situations: a unitary block w of size d_w acting on (1) a subspace of \tilde{S} and (2) a subspace of both S and \tilde{S} . Then, for arbitrary operators $A, B : S \otimes \tilde{S} \rightarrow S \otimes \tilde{S}$, the expectation value of Eq. (B11) over $w : S_w \rightarrow S_w$ can be obtained as follows: For (1) $S_w \subseteq \tilde{S}$,

$$\begin{aligned}\langle \text{Tr}[\text{Tr}_{\tilde{S}}(wAw^\dagger) \text{Tr}_{\tilde{S}}(B)] \text{Tr}[\text{Tr}_{\tilde{S}}(wAw^\dagger) \text{Tr}_{\tilde{S}}(B)] \rangle_w &= \langle \text{Tr}[\text{Tr}_{\tilde{S}}(Aw^\dagger w) \text{Tr}_{\tilde{S}}(B)] \text{Tr}[\text{Tr}_{\tilde{S}}(Aw^\dagger w) \text{Tr}_{\tilde{S}}(B)] \rangle_w \\ &= \text{Tr}[\text{Tr}_{\tilde{S}}(A) \text{Tr}_{\tilde{S}}(B)] \text{Tr}[\text{Tr}_{\tilde{S}}(A) \text{Tr}_{\tilde{S}}(B)],\end{aligned}\quad (\text{B12})$$

and (2) $S_w = S \otimes S_{\tilde{h}}$ with $S_{\tilde{h}} \subset \tilde{S}$,

$$\begin{aligned}\langle \text{Tr}[\text{Tr}_{\tilde{S}}(wAw^\dagger) \text{Tr}_{\tilde{S}}(B)] \text{Tr}[\text{Tr}_{\tilde{S}}(wAw^\dagger) \text{Tr}_{\tilde{S}}(B)] \rangle_w \\ = \langle \text{Tr}[(wAw^\dagger \otimes B) \text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\tilde{S}_1 \otimes \tilde{S}_2}] \text{Tr}[(wAw^\dagger \otimes B) \text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\tilde{S}_1 \otimes \tilde{S}_2}] \rangle_w\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{d_w^2 - 1} \left[\left(\frac{d_w}{\dim(S)} \right)^2 \text{Tr}(A)\text{Tr}(B)\text{Tr}(A)\text{Tr}(B) + \frac{d_w}{\dim(S)} \text{Tr}[\text{Tr}_{S_w}(A)\text{Tr}_{S_w}(A)]\text{Tr}[\text{Tr}_{\bar{S}}(B)\text{Tr}_{\bar{S}}(B)] \right] \\
&\quad - \frac{1}{d_w(d_w^2 - 1)} \left[\frac{d_w}{\dim(S)} \text{Tr}(A)\text{Tr}(A)\text{Tr}[\text{Tr}_{\bar{S}}(B)\text{Tr}_{\bar{S}}(B)] + \left(\frac{d_w}{\dim(S)} \right)^2 \text{Tr}[\text{Tr}_{S_w}(A)\text{Tr}_{S_w}(A)]\text{Tr}(B)\text{Tr}(B) \right]. \quad (\text{B13})
\end{aligned}$$

Then we can obtain

$$\begin{aligned}
\langle k_{\text{PQ}}^{(\kappa)^2} \rangle_{W_{1,L}, \dots, W_{\zeta,L}} &= \langle \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}, \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}, \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \rangle_{W_{1,L}, \dots, W_{\zeta,L}} \\
&= \langle \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}(W_{\kappa,L} \rho_{\mathbf{x},L-1} W_{\kappa,L}^\dagger) \text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}(W_{\kappa,L} \rho_{\mathbf{x},L-1} W_{\kappa,L}^\dagger) \text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \rangle_{W_{\rho,L}} \\
&= \frac{1}{2^{2m} - 1} \left[\left(\frac{2^m}{2^{n_\kappa}} \right)^2 \text{Tr}(\rho_{\mathbf{x},L-1})\text{Tr}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}(\rho_{\mathbf{x},L-1})\text{Tr}[\rho(\mathbf{x}', \boldsymbol{\theta})] \right. \\
&\quad \left. + \frac{2^m}{2^{n_\kappa}} \text{Tr}\{\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\} \text{Tr}[\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \right] \\
&\quad - \frac{1}{2^m(2^{2m} - 1)} \left[\frac{2^m}{2^{n_\kappa}} \text{Tr}(\rho_{\mathbf{x},L-1})\text{Tr}(\rho_{\mathbf{x},L-1})\text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \right. \\
&\quad \left. + \left(\frac{2^m}{2^{n_\kappa}} \right)^2 \text{Tr}[\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})] \text{Tr}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}[\rho(\mathbf{x}', \boldsymbol{\theta})] \right] \\
&= \frac{1}{2^{2m} - 1} \left[\left(\frac{2^m}{2^{n_\kappa}} \right)^2 + \frac{2^m}{2^{n_\kappa}} \text{Tr}[\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})] \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \right] \\
&\quad - \frac{1}{2^m(2^{2m} - 1)} \left[\frac{2^m}{2^{n_\kappa}} \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} + \left(\frac{2^m}{2^{n_\kappa}} \right)^2 \text{Tr}[\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})] \right] \\
&= \frac{1}{(2^{2m} - 1)2^{2n_\kappa}} \left\{ \left[2^{2m} \text{Tr}[\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})] - 1 \right] 2^{n_\kappa} \text{Tr}\{\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}_{S_{\bar{\kappa}}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\} \right. \\
&\quad \left. + 2^m \text{Tr}[\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})\text{Tr}_{S_{(\rho,L)}}(\rho_{\mathbf{x},L-1})] - 2^{2m} \right\}, \quad (\text{B14})
\end{aligned}$$

where Eqs. (B12) and (B13) are used in the second and third equalities, respectively, and the trace property of the density matrix is used in the last equality. Also, $\rho_{\mathbf{x},d}$ denotes the quantum state resulting from the initial state to which the unitary operator $U(\mathbf{x}, \boldsymbol{\theta})$, except for the unitary blocks from the $(d+1)$ th layer through the last layer, is applied, i.e., $\rho(\mathbf{x}, \boldsymbol{\theta}) = [\prod_{l=d}^L V_l(\mathbf{x}, \boldsymbol{\theta})] \rho_{\mathbf{x},d} [\prod_{l=d}^L V_l(\mathbf{x}, \boldsymbol{\theta})]^\dagger$. We remind the reader that $S_{(l,d)}$ denotes the subspace of qubits on which the unitary block $W_{l,d}$ ($W'_{l,d}$) acts.

Next we compute the integration of $\langle k_{\text{PQ}}^{(\kappa)^2} \rangle_{W_{1,L}, \dots, W_{\zeta,L}}$ in Eq. (B14) over the unitary blocks in the last layer of $U(\mathbf{x}', \boldsymbol{\theta})$. In this case, only $\text{Tr}\{\text{Tr}_{\bar{S}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\text{Tr}_{\bar{S}}[\rho(\mathbf{x}', \boldsymbol{\theta})]\}$ in Eq. (B14) matters. To compute the integral of the quantity over the unitary blocks in the last layer, the following situations can be expected: a unitary block w acting on (a) a subspace of S , (b) a subspace of \bar{S} , (c) a subspace of both S and \bar{S} , and (d) S and a subspace of \bar{S} . Then, for arbitrary operator $A : S' \otimes \bar{S}' \rightarrow S' \otimes \bar{S}'$, the expectation value of $\text{Tr}[\text{Tr}_{\bar{S}}(wAw^\dagger)\text{Tr}_{\bar{S}}(wAw^\dagger)]$ over $w : S_w \rightarrow S_w$ can be obtained as follows: For

(a) $S_w \subseteq S$,

$$\langle \text{Tr}[\text{Tr}_{\bar{S}}(wAw^\dagger)\text{Tr}_{\bar{S}}(wAw^\dagger)] \rangle_w = \langle \text{Tr}[w\text{Tr}_{\bar{S}}(A)w^\dagger w\text{Tr}_{\bar{S}}(A)w^\dagger] \rangle_w = \text{Tr}[\text{Tr}_{\bar{S}}(A)\text{Tr}_{\bar{S}}(A)]; \quad (\text{B15})$$

(b) $S_w \subset \bar{S}$,

$$\langle \text{Tr}[\text{Tr}_{\bar{S}}(wAw^\dagger)\text{Tr}_{\bar{S}}(wAw^\dagger)] \rangle_w = \langle \text{Tr}[\text{Tr}_{\bar{S}}(Aw^\dagger w)\text{Tr}_{\bar{S}}(Aw^\dagger w)] \rangle_w = \text{Tr}[\text{Tr}_{\bar{S}}(A)\text{Tr}_{\bar{S}}(A)]; \quad (\text{B16})$$

(c) $S_w = S_h \otimes S_{\bar{h}}$ with $d^{1/2}$ -dimensional spaces $S_h \subset S$ and $S_{\bar{h}} \subset \bar{S}$,

$$\begin{aligned}
&\langle \text{Tr}[\text{Tr}_{\bar{S}}(wAw^\dagger)\text{Tr}_{\bar{S}}(wAw^\dagger)] \rangle_w \\
&= \langle \text{Tr}[(wAw^\dagger \otimes wAw^\dagger)(\text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\bar{S}_1 \otimes \bar{S}_2})] \rangle_w = \frac{1}{d^2 - 1} \left\{ \text{Tr}[(\mathbb{I}_{S_{w,1} \otimes S_{w,2}} \otimes \text{Tr}_{S_{w,1}}(A) \otimes \text{Tr}_{S_{w,2}}(A))(\text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\bar{S}_1 \otimes \bar{S}_2})] \right. \\
&\quad \left. + \text{Tr}[(\text{SWAP}_{S_{w,1}, S_{w,2}} \otimes \text{Tr}_{S_w} \otimes \text{Tr}_{S_{w,1} \cup S_{w,2}}[A \otimes A(\text{SWAP}_{S_{w,1}, S_{w,2}} \otimes \mathbb{I}_{S_{w,1} \otimes S_{w,2}})])](\text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\bar{S}_1 \otimes \bar{S}_2}) \right\}
\end{aligned}$$

$$\begin{aligned}
& - \frac{1}{d(d^2-1)} \left[\text{Tr} \left(\left\{ \mathbb{I}_{S_{w,1} \otimes S_{w,1}} \otimes \text{Tr}_{S_{w,1} \cup S_{w,2}} \left[A \otimes A \left(\text{SWAP}_{S_{w,1}, S_{w,2}} \otimes \mathbb{I}_{S_{w,1}^- \otimes S_{w,2}^-} \right) \right] \right\} \left(\text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\bar{S}_1 \otimes \bar{S}_2} \right) \right) + \text{Tr} \right. \\
& \quad \times \left. \left\{ \left[\text{SWAP}_{S_{w,1}, S_{w,2}} \otimes \text{Tr}_{S_{w,1}}(A) \otimes \text{Tr}_{S_{w,2}}(A) \right] \left(\text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\bar{S}_1 \otimes \bar{S}_2} \right) \right\} \right] \\
& = \frac{d^{1/2}}{d+1} \left\{ \text{Tr} \left[\text{Tr}_{S \cup S_{\bar{h}}}(A) \text{Tr}_{S \cup S_{\bar{h}}}(A) \right] + \text{Tr} \left[\text{Tr}_{S \setminus S_{\bar{h}}}(A) \text{Tr}_{S \setminus S_{\bar{h}}}(A) \right] \right\}; \tag{B17}
\end{aligned}$$

and (d) $S_w = S \otimes S_{\bar{h}}$ with $d^{1/2}$ -dimensional spaces S and $S_{\bar{h}} \subset \bar{S}$,

$$\begin{aligned}
\langle \text{Tr} \left[\text{Tr}_{\bar{S}}(wAw^\dagger) \text{Tr}_{\bar{S}}(wAw^\dagger) \right] \rangle_w & = \langle \text{Tr} \left[(wAw^\dagger \otimes wAw^\dagger) \left(\text{SWAP}_{S_1, S_2} \otimes \mathbb{I}_{\bar{S}_1 \otimes \bar{S}_2} \right) \right] \rangle_w \\
& = \frac{d^{1/2}}{d+1} \left\{ \text{Tr}(A) \text{Tr}(A) + \text{Tr} \left[\text{Tr}_{S \setminus S_{\bar{h}}}(A) \text{Tr}_{S \setminus S_{\bar{h}}}(A) \right] \right\}. \tag{B18}
\end{aligned}$$

Hence, using Eqs. (B15) to (B18), we obtain

$$\begin{aligned}
& \langle \text{Tr} \left\{ \text{Tr}_{\bar{S}_\kappa}[\rho(\mathbf{x}', \boldsymbol{\theta})] \text{Tr}_{\bar{S}_\kappa}[\rho(\mathbf{x}', \boldsymbol{\theta})] \right\} \rangle_{W'_{1,L}, \dots, W'_{\zeta,L}} \\
& = \langle \text{Tr} \left[\text{Tr}_{\bar{S}_\kappa} \left(W'_{p,L} \rho_{\mathbf{x}', L-1} W_{p,L}^\dagger \right) \text{Tr}_{\bar{S}_\kappa} \left(W'_{p,L} \rho_{\mathbf{x}', L-1} W_{p,L}^\dagger \right) \right] \rangle_{W'_{p,L}} \\
& = \frac{1}{2^{2m}-1} \left[\left(\frac{2^m}{2^{n_\kappa}} \right)^2 2^{n_\kappa} + \left(\frac{2^m}{2^{n_\kappa}} \right) 2^{2n_\kappa} \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \right] \right] \\
& \quad - \frac{1}{2^m(2^{2m}-1)} \left[\left(\frac{2^m}{2^{n_\kappa}} \right) 2^{2n_\kappa} + \left(\frac{2^m}{2^{n_\kappa}} \right)^2 2^{n_\kappa} \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \right] \right], \\
& = \frac{1}{2^{2m}-1} \left[\left(2^{m+n_\kappa} - \frac{2^m}{2^{n_\kappa}} \right) \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \right] + \frac{2^{2m}}{2^{n_\kappa}} - 2^{2n_\kappa} \right]. \tag{B19}
\end{aligned}$$

Thus, from Eqs. (B14) and (B19), the expectation value $\langle k_{\text{PQ}}^{(\kappa)^2} \rangle_{W_{1,L}, \dots, W_{\zeta,L}, W'_{1,L}, \dots, W'_{\zeta,L}}$ can read

$$\begin{aligned}
& \langle k_{\text{PQ}}^{(\kappa)^2} \rangle_{W_{1,L}, \dots, W_{\zeta,L}, W'_{1,L}, \dots, W'_{\zeta,L}} \\
& = \frac{1}{(2^{2m}-1)2^{2n_\kappa}} \left\{ 2^m \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \right] - 1 \right\} \\
& \quad \times 2^{n_\kappa} \frac{1}{2^{2m}-1} \left[\left(2^{m+n_\kappa} - \frac{2^m}{2^{n_\kappa}} \right) \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \right] + \frac{2^{2m}}{2^{n_\kappa}} - 2^{2n_\kappa} \right] \\
& \quad + 2^m \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \right] - 2^{2m} \left\} \\
& = \frac{1}{2^{2n_\kappa}} + \frac{2^{2m}(2^{2n_\kappa}-1)}{(2^{2m}-1)2^{2n_\kappa}} \left(\text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \right] - \frac{1}{2^m} \right) \left(\text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}', L-1}) \right] - \frac{1}{2^m} \right) \\
& = \frac{1}{2^{2n_\kappa}} + \frac{2^{2m}(2^{2n_\kappa}-1)}{(2^{2m}-1)2^{2n_\kappa}} \left(\text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \right] - \frac{1}{2^m} \right)^2. \tag{B20}
\end{aligned}$$

In the last equality, we utilize our assumption that any unitary blocks in $U(\mathbf{x}, \boldsymbol{\theta})$ and $U(\mathbf{x}', \boldsymbol{\theta})$ are 2-designs and an additional assumption that the same initial state is prepared for $\rho(\mathbf{x}, \boldsymbol{\theta})$ and $\rho(\mathbf{x}', \boldsymbol{\theta})$. Therefore, the variance of the PQK for ALAs can be written as

$$\begin{aligned}
\text{Var}(k_{\text{PQ}}^{(\kappa)}) & = \langle k_{\text{PQ}}^{(\kappa)^2} \rangle - \langle k_{\text{PQ}}^{(\kappa)} \rangle^2 \\
& = \langle \langle k_{\text{PQ}}^{(\kappa)^2} \rangle_{W_{1,L}, \dots, W_{\zeta,L}, W'_{1,L}, \dots, W'_{\zeta,L}} \rangle_{W_{1,1}, \dots, W_{\zeta, L-1}, W'_{1,1}, \dots, W'_{\zeta, L-1}} - \langle k_{\text{PQ}}^{(\kappa)} \rangle^2 \\
& = \frac{1}{2^{2n_\kappa}} + \frac{2^{2m}(2^{2n_\kappa}-1)}{(2^{2m}-1)2^{2n_\kappa}} \left(\langle \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \right] \rangle_{W_{1,1}, \dots, W_{\zeta, L-1}} - \frac{1}{2^m} \right)^2 - \frac{1}{2^{2n_\kappa}} \\
& = \frac{2^{2m}(2^{2n_\kappa}-1)}{(2^{2m}-1)2^{2n_\kappa}} \left(\langle \text{Tr} \left[\text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1}) \right] \rangle_{W_{1,1}, \dots, W_{\zeta, L-1}} - \frac{1}{2^m} \right)^2. \tag{B21}
\end{aligned}$$

The implication of Eq. (B21) is that the variance depends on the purity of the quantum state, i.e., $\text{Tr}[(\rho_{\mathbf{x}, L-1}^{S_{(p,L)}})^2]$ with $\rho_{\mathbf{x}, L-1}^{S_{(p,L)}} \equiv \text{Tr}_{\bar{S}_{(p,L)}}(\rho_{\mathbf{x}, L-1})$. We remind the reader that $\rho_{\mathbf{x}, L-1}$ is the quantum state resulting from the initial state to which the unitary operator

$U(\mathbf{x}, \theta)$, except for the unitary blocks in the last layer, is applied and $\text{Tr}_{\bar{S}_{(p,L)}}(\cdot)$ is the partial trace over the subspace $\bar{S}_{(p,L)}$, with $S_{(p,L)}$ the subspace on which the unitary block $W_{p,L}$ acts. This means that the variance is zero if $\rho_{\mathbf{x},L-1}^{S_{(p,L)}}$ is the completely mixed state, i.e., $\mathbb{I}/2^m$. Thus, Eq. (B21) indicates that how fast the quantum state converges to the mixed state is crucial to avoid the trainability issue, while the situation where $\rho_{\mathbf{x},L-1}^{S_{(p,L)}}$ is close to a pure state means the unitary operation is efficiently simulatable by classical computers.

Finally, we check the relationship between the variance and initial state as well as circuit depth of the ALA. We consider here two cases regarding the position of the κ th qubit(s): The κ th qubit(s) is (are) located (1) in the middle so that the number of qubits on which the unitary blocks in the first layer inside the light cone act is smaller than n and (2) in the unitary block at the edge i.e., $W_{1,L}$ or $W_{\zeta,L}$. Case 1 corresponds to the situation where the number of unitary blocks inside the light cone is the smallest and case 2 is the one where the number of blocks is the largest.

For ease of understanding, we first consider one-layer ALAs. In this case, the variance for both cases 1 and 2 can be written as

$$\text{Var}(k_{\text{PQ}}^{(\kappa)}) = \frac{2^{2m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 2^{2n_\kappa}} \left(\text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] - \frac{1}{2^m} \right)^2. \quad (\text{B22})$$

Next we deal with two-layer ALAs. Then, using Eqs. (B15)–(B18), we obtain

$$\begin{aligned} \text{Var}(k_{\text{PQ}}^{(\kappa)}) &= \frac{2^{2m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 2^{2n_\kappa}} \left(\langle \text{Tr}[\text{Tr}_{\bar{S}_{(p,2)}}(\rho_{\mathbf{x},1}) \text{Tr}_{\bar{S}_{(p,2)}}(\rho_{\mathbf{x},1})] \rangle_{W_{1,1}, \dots, W_{\zeta,1}} - \frac{1}{2^m} \right)^2 \\ &= \frac{2^{2m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 2^{2n_\kappa}} \left(\frac{2^m}{(2^m + 1)^2} \{ 1 + \text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] + \text{Tr}[\text{Tr}_{\bar{S}_{(p+1,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p+1,1)}}(\rho_0)] \right. \\ &\quad \left. + \text{Tr}[\text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0)] \right) - \frac{1}{2^m} \Big)^2 \\ &= \frac{2^{2m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 (2^m + 1)^4 2^{2n_\kappa}} \left(-2 - \frac{1}{2^m} + 2^m \text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] \right. \\ &\quad \left. + 2^m \text{Tr}[\text{Tr}_{\bar{S}_{(p+1,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p+1,1)}}(\rho_0)] + 2^m \text{Tr}[\text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0)] \right)^2 \quad (\text{B23}) \end{aligned}$$

for cases 1 and 2.

Subsequently, for case 1 with a three-layer ALA, we have

$$\begin{aligned} \text{Var}(k_{\text{PQ}}^{(\kappa)}) &= \frac{2^{2m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 2^{2n_\kappa}} \left(\langle \text{Tr}[\text{Tr}_{\bar{S}_{(p,3)}}(\rho_{\mathbf{x},2}) \text{Tr}_{\bar{S}_{(p,3)}}(\rho_{\mathbf{x},2})] \rangle_{W_{1,1}, \dots, W_{\zeta,2}} - \frac{1}{2^m} \right)^2 \\ &= \frac{2^{4m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 (2^m + 1)^8 2^{2n_\kappa}} \left(-5 - \frac{4}{2^m} - \frac{1}{2^{2m}} + 2^m \{ \text{Tr}[\text{Tr}_{\bar{S}_{(p-1,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p-1,1)}}(\rho_0)] + \text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] \right. \\ &\quad + \text{Tr}[\text{Tr}_{\bar{S}_{(p-1,1) \cup \bar{S}_{(p,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p-1,1) \cup \bar{S}_{(p,1)}}}(\rho_0)] \} + 2^m \{ \text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] + \text{Tr}[\text{Tr}_{\bar{S}_{(p+1,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p+1,1)}}(\rho_0)] \\ &\quad + \text{Tr}[\text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0)] \} + 2^m \{ \text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] + \text{Tr}[\text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0)] \\ &\quad \left. + \text{Tr}[\text{Tr}_{\bar{S}_{(p-1,1) \cup \bar{S}_{(p,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p-1,1) \cup \bar{S}_{(p,1)}}}(\rho_0)] + \text{Tr}[\text{Tr}_{\bar{S}_{(p-1,1) \cup \bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p-1,1) \cup \bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0)] \} \right). \quad (\text{B24}) \end{aligned}$$

For case 2 we get

$$\begin{aligned} \text{Var}(k_{\text{PQ}}^{(\kappa)}) &= \frac{2^{2m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 (2^m + 1)^4 2^{2n_\kappa}} \left(-1 - \frac{1}{2^m} + 2^{m/2} \{ 2 \text{Tr}[\text{Tr}_{\bar{S}_{(p,3) \setminus \bar{S}_{(p,2)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,3) \setminus \bar{S}_{(p,2)}}}(\rho_0)] \right. \\ &\quad \left. + \text{Tr}[\text{Tr}_{\bar{S}_{(p,3) \cup \bar{S}_{(p,2)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,3) \cup \bar{S}_{(p,2)}}}(\rho_0)] \} \right)^2 \\ &= \frac{2^{3m}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2 (2^m + 1)^4 2^{2n_\kappa}} \left(-\frac{2}{2^{m/2}} - \frac{1}{2^{3m/2}} + 2^{m/2} \{ 2 \text{Tr}[\text{Tr}_{\bar{S}_{(p,1)}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1)}}(\rho_0)] \right. \\ &\quad \left. + \text{Tr}[\text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0) \text{Tr}_{\bar{S}_{(p,1) \cup \bar{S}_{(p+1,1)}}}(\rho_0)] \} \right)^2. \quad (\text{B25}) \end{aligned}$$

We note that here we consider $p = 1$ without loss of generality.

Therefore, the variance in case 1 reads

$$\text{Var}(k_{\text{PQ}}^{(\kappa)}) = \frac{2^{2m(L-1)}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2(2^m + 1)^{4(L-1)}2^{2n_\kappa}} \left(2^m \sum_{h \in P(S_{(k_u,1)} : S_{(k_l,1)})} t_h \text{Tr}[\text{Tr}_{\bar{h}}(\rho_0)\text{Tr}_{\bar{h}}(\rho_0)] - \sum_{\tau=0}^{L-1} \frac{c_\tau}{2^{m\tau}} \right)^2, \quad (\text{B26})$$

where $c_\tau, t_h \in \mathbb{R}^+$ and $P(S_{(k_u,1)} : S_{(k_l,1)}) = \{\bigcup_{i=\xi}^{\xi+L} S_{(k_u+i,1)} | l \in \{1, \dots, L\}, \xi \in \{0, \dots, (k_l - k_u) - i + 1\}\}$ is the set containing all the possible neighboring subspaces in $\bigcup_{i=0}^{k_l-k_u} S_{(k_u+i,1)}$. We define here $W_{k_u,1}$ and $W_{k_l,1}$ as the the unitary blocks located at the edge of the light cone in the first layer. We also remind the reader that $S_{(l,d)}$ denotes the subspace of the qubits that the unitary operator $W_{l,d}$ acts on.

As for case 2 with odd L layers ($L \geq 3$), we get

$$\text{Var}(k_{\text{PQ}}^{(\kappa)}) = \frac{2^{mL}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2(2^m + 1)^{2(L-1)}2^{2n_\kappa}} \left(2^{m/2} \sum_{h \in P'(S_{(k_u,1)} : S_{(k_l,1)})} t'_h \text{Tr}[\text{Tr}_{\bar{h}}(\rho_0)\text{Tr}_{\bar{h}}(\rho_0)] - \sum_{\tau=1}^{(L+1)/2} \frac{c'_\tau}{2^{m(2\tau-1)/2}} \right)^2, \quad (\text{B27})$$

where $c'_\tau, t'_h \in \mathbb{R}^+$ and $P'(S_{(k_u,1)} : S_{(k_l,1)}) = \{S_{(p,1)} \cup S_{(p\pm 1,1)} \cup S_i | S_i \in P(S_{(k_u,1)} : S_{(k_l,1)})\}$. If L is even, the variance is written as

$$\text{Var}(k_{\text{PQ}}^{(\kappa)}) = \frac{2^{mL}(2^{2n_\kappa} - 1)}{(2^{2m} - 1)^2(2^m + 1)^{2(L-1)}2^{2n_\kappa}} \left(2^m \sum_{h \in P(S_{(k_u,1)} : S_{(k_l,1)})} t''_h \text{Tr}[\text{Tr}_{\bar{h}}(\rho_0)\text{Tr}_{\bar{h}}(\rho_0)] - \sum_{\tau=0}^{L/2} \frac{c''_\tau}{2^{m\tau}} \right)^2, \quad (\text{B28})$$

with $c''_\tau, t''_h \in \mathbb{R}^+$. We note that $W_{k_u,1}$ and $W_{k_l,1}$ denote the unitary blocks located at the edge of the light cone in the first layer. Also, Eqs. (B26)–(B28) go to zero if $\text{Tr}[\text{Tr}_{\bar{h}}(\rho_0)\text{Tr}_{\bar{h}}(\rho_0)]$ is the completely mixed state for any h and reaches the maximum when $\text{Tr}[\text{Tr}_{\bar{h}}(\rho_0)\text{Tr}_{\bar{h}}(\rho_0)] = 1$ for all h . We comment that the result for different initial states is also easily obtainable.

Here we summarize key implications obtained from Eqs. (B26)–(B28). These results indicate that variance of the PQK depends on not only the depth L but also the initial state ρ_0 . If the initial state is a tensor product of arbitrary single-qubit states, i.e., $\rho_0 = \sigma_1 \otimes \sigma_2 \otimes \dots \otimes \sigma_n$ with an arbitrary single-qubit states $\{\sigma_i\}$, then the variances are $\Omega(2^{-2mL})$ and $\Omega(2^{-mL})$ for cases 1 and 2, respectively; this means that it might be possible to preserve the trainability up to $\text{poly}[\log(n)]$ depth. On the other hand, if we prepare an entangled quantum state such as the GHZ state $|\psi_{\text{GHZ}}\rangle = 2^{-1/2}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$, then the variance would be smaller than the case for a tensor product of arbitrary single-qubit states. Note that $\text{Tr}[\text{Tr}_{\bar{h}}(|\psi_{\text{GHZ}}\rangle\langle\psi_{\text{GHZ}}|)\text{Tr}_{\bar{h}}(|\psi_{\text{GHZ}}\rangle\langle\psi_{\text{GHZ}}|)] = 1/2$ for $\bar{h} \neq \emptyset$ or $h \neq \emptyset$; otherwise $\text{Tr}[\text{Tr}_{\bar{h}}(|\psi_{\text{GHZ}}\rangle\langle\psi_{\text{GHZ}}|)\text{Tr}_{\bar{h}}(|\psi_{\text{GHZ}}\rangle\langle\psi_{\text{GHZ}}|)] = 1$. In the worst-case scenario where the initial state is random enough so that $\text{Tr}_{\bar{h}}(\rho_0)$ is the completely mixed state for almost all h , then the variance would be closer to zero.

3. Variance of the linear projected quantum kernel

In this section we further check the difference of the variance between PQKs in Eq. (8) and the linear PQK defined as

$$k_{\text{PQ}}^L(\mathbf{x}, \mathbf{x}') = \sum_{\kappa} \text{Tr}\{\text{Tr}_{S_\kappa}[\rho(\mathbf{x}, \boldsymbol{\theta})]\text{Tr}_{S_\kappa}[\rho(\mathbf{x}', \boldsymbol{\theta})]\}. \quad (\text{B29})$$

The variance of Eq. (B29) can be written as

$$\begin{aligned} \text{Var}(k_{\text{PQ}}^L) &= \text{Var}\left(\sum_{\kappa} k_{\text{PQ}}^{(\kappa)}(\mathbf{x}, \mathbf{x}')\right) \\ &= \sum_{\kappa} \text{Var}(k_{\text{PQ}}^{(\kappa)}) + 2 \sum_{\kappa > \kappa'} \text{Cov}[k_{\text{PQ}}^{(\kappa)}, k_{\text{PQ}}^{(\kappa')}], \end{aligned} \quad (\text{B30})$$

where $\text{Cov}[A, B]$ represents the covariance of A and B . Then Eq. (B30) means that the variance of the linear PQK is different from the simple summation of the variances of Eq. (B1) because of the covariance terms. Actually, the covariance of $k_{\text{PQ}}^{(\kappa)}$ and $k_{\text{PQ}}^{(\kappa')}$ differs depending on whether or not the κ th qubit(s) and κ' th qubit(s) are located in the same subspace of a local unitary block in the last layer. If $S_\kappa, S_{\kappa'} \subseteq S_W$ with a unitary block in the last layer W , then we obtain

$$\text{Cov}[k_{\text{PQ}}^{(\kappa)}, k_{\text{PQ}}^{(\kappa')}] = \text{Var}(k_{\text{PQ}}^{(\kappa)}). \quad (\text{B31})$$

Moreover, if $S_\kappa \subseteq S_W, S_{\kappa'} \subseteq S_{W'}$, and $S_W \cap S_{W'} = \emptyset$, then the covariance reads

$$\text{Cov}[k_{\text{PQ}}^{(\kappa)}, k_{\text{PQ}}^{(\kappa')}] = 0. \quad (\text{B32})$$

We note that the result can be obtained by following exactly what we did for the variance calculation in Appendixes B 1 and B 2, that is, the integration of the covariance over the unitary blocks in the last layer. These results indicate that the variance would not be recovered exponentially by the covariance terms in Eq. (B30), whereas the variance could possibly increase.

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

[2] P. Rebentrost, M. Mohseni, and S. Lloyd, Quantum support vector machine for big data classification, *Phys. Rev. Lett.* **113**, 130503 (2014).

- [3] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem, *Science* **292**, 472 (2001).
- [4] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, *Nat. Phys.* **17**, 1013 (2021).
- [5] 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).
- [6] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, On the quantum versus classical learnability of discrete distributions, *Quantum* **5**, 417 (2021).
- [7] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, *Phys. Rev. Lett.* **126**, 190505 (2021).
- [8] 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).
- [9] 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).
- [10] M. Schuld and N. Killoran, Quantum machine learning in feature Hilbert spaces, *Phys. Rev. Lett.* **122**, 040504 (2019).
- [11] M. Schuld, Supervised quantum machine learning models are kernel methods, [arXiv:2101.11020](https://arxiv.org/abs/2101.11020).
- [12] Y. Suzuki, H. Kawaguchi, and N. Yamamoto, Quantum Fisher kernel for mitigating the vanishing similarity issue, *Quantum Sci. Technol.* **9**, 035050 (2024).
- [13] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, Exponential concentration in quantum kernel methods, *Nat. Commun.* **15**, 5200 (2024).
- [14] 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).
- [15] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.* **12**, 6961 (2021).
- [16] M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, *Quantum Sci. Technol.* **6**, 035006 (2021).
- [17] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, *Quantum* **5**, 558 (2021).
- [18] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, *Nat. Commun.* **12**, 1791 (2021).
- [19] J. Kübler, S. Buchholz, and B. Schölkopf, The inductive bias of quantum kernels, *Adv. Neural Inf. Process. Syst.* **34**, 12661 (2021).
- [20] A. Canatar, E. Peters, C. Pehlevan, S. M. Wild, and R. Shaydulin, Bandwidth enables generalization in quantum kernel models, *Trans. Mach. Learn. Res.* (2023).
- [21] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, *Adv. Quantum Technol.* **2**, 1900070 (2019).
- [22] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Average-case complexity versus approximate simulation of commuting quantum computations, *Phys. Rev. Lett.* **117**, 080501 (2016).
- [23] L. A. Goldberg and H. Guo, The complexity of approximating complex-valued Ising and Tutte partition functions, *Comput. Complexity* **26**, 765 (2017).
- [24] A. W. Harrow and S. Mehraban, Approximate unitary t -designs by short random quantum circuits using nearest-neighbor and long-range gates, *Commun. Math. Phys.* **401**, 1531 (2023).
- [25] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, *Phys. Rev. A* **80**, 012304 (2009).
- [26] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, *J. Math. Phys.* **45**, 2171 (2004).
- [27] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, Quantum embeddings for machine learning, [arXiv:2001.03622](https://arxiv.org/abs/2001.03622).
- [28] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht, and A. T. Sornborger, Barren plateaus preclude learning scramblers, *Phys. Rev. Lett.* **126**, 190501 (2021).
- [29] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, *PRX Quantum* **3**, 010313 (2022).
- [30] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Equivalence of quantum barren plateaus to cost concentration and narrow gorges, *Quantum Sci. Technol.* **7**, 045015 (2022).
- [31] L. Leone, S. F. E. Oliviero, L. Cincio, and M. Cerezo, On the practical usefulness of the hardware efficient ansatz, [arXiv:2211.01477](https://arxiv.org/abs/2211.01477).
- [32] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, Diagnosing barren plateaus with tools from quantum optimal control, *Quantum* **6**, 824 (2022).
- [33] 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).
- [34] Qiskit contributors, Qiskit: An open-source framework for quantum computing (IBM, Armonk, 2023).
- [35] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nat. Phys.* **16**, 1050 (2020).
- [36] 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).
- [37] S. P. Adam, S.-A. N. Alexandropoulos, P. M. Pardalos, and M. N. Vrahatis, No free lunch theorem: A review, in *Approximation and Optimization*, edited by I. Demetriou and P. Pardalos, Springer Optimization and Its Applications Vol. 145 (Springer, Cham, 2019), pp. 57–82.
- [38] D. H. Wolpert and W. G. Macready, No free lunch theorems for optimization, *IEEE Trans. Evol. Computat.* **1**, 67 (1997).
- [39] D. H. Wolpert, in *Soft Computing and Industry*, edited by R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, and F. Hoffmann (Springer, London, 2002), pp. 25–42.
- [40] 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).
- [41] 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](https://arxiv.org/abs/2210.07980).

- [42] 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).
- [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] 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).