On the practical usefulness of the Hardware Efficient Ansatz

Lorenzo Leone[1,2,3], Salvatore F.E. Oliviero[1,2,3], Lukasz Cincio[1], and M. Cerezo[4]

[1]Theoretical Division (T-4), Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
[2]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
[3]Physics Department, University of Massachusetts Boston, Boston, Massachusetts 02125, USA
[4]Information Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Variational Quantum Algorithms (VQAs) and Quantum Machine Learning (QML) models train a parametrized quantum circuit to solve a given learning task. The success of these algorithms greatly hinges on appropriately choosing an ansatz for the quantum circuit. Perhaps one of the most famous ansatzes is the one-dimensional layered Hardware Efficient Ansatz (HEA), which seeks to minimize the effect of hardware noise by using native gates and connectives. The use of this HEA has generated a certain ambivalence arising from the fact that while it suffers from barren plateaus at long depths, it can also avoid them at shallow ones. In this work, we attempt to determine whether one should, or should not, use a HEA. We rigorously identify scenarios where shallow HEAs should likely be avoided (e.g., VQA or QML tasks with data satisfying a volume law of entanglement). More importantly, we identify a Goldilocks scenario where shallow HEAs could achieve a quantum speedup: QML tasks with data satisfying an area law of entanglement. We provide examples for such scenario (such as Gaussian diagonal ensemble random Hamiltonian discrimination), and we show that in these cases a shallow HEA is always trainable and that there exists an anti-concentration of loss function values. Our work highlights the crucial role that input states play in the trainability of a parametrized quantum circuit, a phenomenon that is verified in our numerics.

1 Introduction

The advent of Noisy Intermediate-Scale Quantum (NISQ) [1] computers has generated a tremendous amount of excitement. Despite the presence of hardware noise and their limited qubit count, near-term quantum computers are already capable of outperforming the world's largest super-computers on certain contrived mathematical tasks [2–4]. This has started a veritable rat race to solve real-life tasks of interest in NISQ hardware.

One of the most promising strategies to make practical use of near-term quantum computers is to train parametrized hybrid quantum-classical models. Here, a quantum device is used to estimate a classically hard-to-compute quantity, while one also leverages classical optimizers to train the parameters in the model. When the algorithm is problem-driven, we usually refer to it as a Variational Quantum Algorithm (VQA) [5, 6]. VQAs can be used for a wide range of tasks such as finding the ground state of molecular Hamiltonians [7, 8], solving combinatorial optimization tasks [9, 10] and solving linear systems of equations [11–13], among others. On the other hand, when the algorithm is data-driven, we refer to it as a Quantum Machine Learning (QML) model [14, 15]. QML can be used in supervised [16, 17], unsupervised [18] and reinforced [19] learning problems, where the data processed in the quantum device can either be classical data embedded in quantum states [16, 20], or quantum data obtained from some physical process [21–23].

Both VQAs and QML models train parametrized quantum circuits $U(\boldsymbol{\theta})$ to solve their respective tasks. One of, if not *the*, most important aspect in determining the success of these near-term algorithms is the choice of ansatz for the parametrized quantum circuit [24]. By ansatz, we mean the specifications for the arrangement and type of quantum gates in $U(\boldsymbol{\theta})$, and how these depend on the set of trainable parameters $\boldsymbol{\theta}$. Recently, the field of ansatz design has seen a Cambrian explosion where researchers have proposed a plethora of ansatzes for VQAs and QML [5, 6]. These include variable structure ansatzes [25–29], problem-inspired ansatzes [30–34] and even the recently introduced field of

Lorenzo Leone: Lorenzo.Leone001@umb.edu
Salvatore F.E. Oliviero: s.oliviero001@umb.edu

geometric quantum machine learning where one embeds information about the data symmetries into $U(\boldsymbol{\theta})$ [35–41]. Choosing an appropriate ansatz is crucial as it has been shown that ill-defined ansatzes can be untrainable [42–51], and hence useless for large scale implementations.

Perhaps the most famous, and simultaneously infamous, ansatz is the so-called Hardware Efficient Ansatz (HEA). As its name implies, the main objective of HEA is to mitigate the effect of hardware noise by using gates native to the specific device being used. The previous avoids the gate overhead which arises when compiling [52] a non-native gate-set into a sequence of native gates. While the HEA was originally proposed within the framework of VQAs, it is now also widely used in QML tasks. The strengths of the HEA are that it can be as depth-frugal as possible and that it is problem-agnostic, meaning that one can use it in any scenario. However, its wide usability could also be its greater weakness, as it is believed that the HEA cannot have a good performance on all tasks [46] (this is similar to the famous no-free-lunch theorem in classical machine learning [53]). Moreover, it was shown that deep HEA circuits suffer from barren plateaus [42] due to their high expressibility [46]. Despite these difficulties, the HEA is not completely hopeless. In Ref. [43], the HEA saw a glimmer of hope as it was shown that shallow HEAs can be immune to barren plateaus, and thus have trainability guarantees.

From the previous, the HEA was left in a sort of gray-area of ansatzes, where its practical usefulness was unclear. On the one hand, there is a common practice in the field of using the HEA irrespective of the problem one is trying to solve. On the other hand, there is a significant push to move away from problem-agnostic HEA, and instead develop problem-specific ansatzes. However, the answer to questions such as "*Should we use (if at all) the HEA?*" or "*What problems are shallow HEAs good for?*" have not been rigorously tackled.

In this work, we attempt to determine what are the problems in VQAs and QML where HEAs should, or should not be used. As we will see, our results indicate that HEAs should likely be avoided in VQA tasks where the input state is a product state, as the ensuing algorithm can be efficiently simulated via classical methods. Similarly, we will rigorously prove that HEAs should not be used in QML tasks where the input data satisfies a volume law of entanglement. In these cases, we connect the entanglement in the input data to the phenomenon of cost concentration, and we show that high levels of entanglement lead to barren plateaus, and hence to untrainability. Finally, we identify a scenario where shallow HEAs can be useful and potentially capable of achieving a quantum advantage: QML tasks where the input data satisfies an area law of entanglement. In these cases, we can guarantee that the optimization landscape will not exhibit barren plateaus. Taken together our results highlight the critical importance that the input data plays in the trainability of a model.

2 Framework

2.1 Variational Quantum Algorithms and Quantum Machine Learning

Throughout this work, we will consider two related, but conceptually different, hybrid quantum-classical models. The first, which we will denote as a Variational Quantum Algorithm (VQA) model, can be used to solve the following tasks

Definition 1 (Variational Quantum Algorithms)**.** *Let O be a Hermitian operator, whose ground state encodes the solution to a problem of interest. In a VQA task, the goal is to minimize a cost function $C(\boldsymbol{\theta})$, parametrized through a quantum circuit $U(\boldsymbol{\theta})$, to prepare the ground state of O from a fiduciary state $|\psi_0\rangle$.*

In a VQA task one usually defines a cost function of the form

$$C(\boldsymbol{\theta}) = \mathrm{Tr}[U(\boldsymbol{\theta})\,|\psi_0\rangle\langle\psi_0|\,U^\dagger(\boldsymbol{\theta})O]\,, \tag{1}$$

and trains the parameters in $U(\boldsymbol{\theta})$ by solving the optimization task $\arg\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})$.

Then, while Quantum Machine Learning (QML) models can be used for a wide range of learning tasks, here we will focus on supervised problems

Definition 2 (Quantum Machine Learning)**.** *Let $\mathcal{S} = \{y_s, |\psi_s\rangle\}$ be a dataset of interest, where $|\psi_s\rangle$ are n-qubit states and y_s associated real-valued labels. In a QML task, the goal is to train a model, by minimizing a loss function $L(\boldsymbol{\theta})$ parametrized through a quantum neural network, i.e., a parametrized quantum circuit, $U(\boldsymbol{\theta})$, to predict labels that closely match those in the dataset.*

The exact form of $L(\boldsymbol{\theta})$, and concomitantly the nature of what we want to "learn" from the dataset depends on the task at hand. For instance, in a binary QML classification task where y_s are labels one can minimize an empirical loss function such as the mean-squared error $L(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}|}\sum_s (y_s - L_s(\boldsymbol{\theta}))^2$, or the hinge-loss where $L(\boldsymbol{\theta}) = \frac{-1}{|\mathcal{S}|}\sum_s y_s L_s(\boldsymbol{\theta})$. Here,

$$L_s(\boldsymbol{\theta}) = \mathrm{Tr}[U(\boldsymbol{\theta})\,|\psi_s\rangle\langle\psi_s|\,U^\dagger(\boldsymbol{\theta})O_s]\,, \tag{2}$$

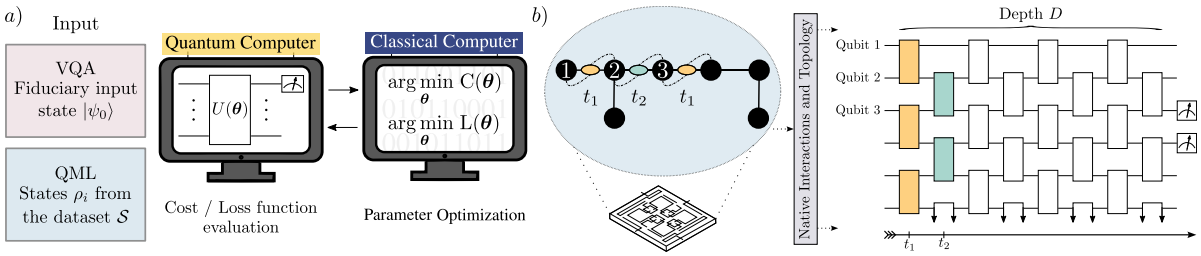

Figure 1: (a) Both VQA and QML models train parametrized quantum circuits $U(\boldsymbol{\theta})$ to minimize either a cost function $C(\boldsymbol{\theta})$ for VQAs, or a loss function $L(\boldsymbol{\theta})$ for QML models. While VQAs start from some fiduciary, easy to prepare state $|\psi_0\rangle$, in QML one uses states from dataset $|\psi_s\rangle \in \mathcal{S}$ as input of the parametrized circuit $U(\boldsymbol{\theta})$. Both models exploit the power of classical optimizers for the minimization task. (b) The architecture of a HEA seeks to minimize the effect of hardware noise by following the topology, and using the native gates, of the physical hardware. Specifically, we consider HEA as a one-dimensional alternating layered ansatz of two-qubit gates organized in a brick-like fashion. In the figure, we show how a first layer of gates is implemented at time t_1 while a second layer at time t_2. At the end of the computation, a local operator is measured.

with O_s being label-dependent Hermitian operator. The parameters in the quantum neural network $U(\boldsymbol{\theta})$ are trained by solving the optimization task $\arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, and the ensuing parameters, along with the loss, are used to make predictions.

While VQAs and QML share some similarities, they also share some differences. Let us first discuss their similarities. First, in both frameworks, one trains a parametrized quantum circuit. This requires choosing an *ansatz* for $U(\boldsymbol{\theta})$ and using a classical optimizer to train its parameters. As for their differences, in a VQA task as described in Definition 1 and Eq. (1), the input state to the parametrized quantum circuit $U(\boldsymbol{\theta})$ is usually an easy-to-prepare state $|\psi_0\rangle$ such as the all-zero state, or some physically motivated product state (e.g., the Hartree-Fock state in quantum chemistry [5, 54]). On the other hand, in a QML task as in Definition 2 and Eq. (2), the input states to $U(\boldsymbol{\theta})$ are taken from the dataset $\mathcal{S}$, and thus can be extremely complex quantum states (see Fig. 1(a)).

2.2 Hardware Efficient Ansatz

As previously mentioned, one of the most important aspects of VQAs and QML models is the choice of ansatz for $U(\boldsymbol{\theta})$. Without loss of generality, we assume that the parametrized quantum circuit is expressed as

$$U(\boldsymbol{\theta}) = \prod_l e^{-i\theta_l H_l} V_l \,, \tag{3}$$

where the $\{V_l\}$ are some unparametrized unitaries, $\{H_l\}$ are traceless Pauli operators, and where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$. While recently the field of ansatz design has seen a tremendous amount of interest, here we will focus on the HEA, one of the most widely used ansatzes in the literature. Originally introduced in Ref. [55], the term HEA is a generic name commonly reserved for ansatzes that are aimed at reducing the circuit depth by choosing gates $\{V_l\}$ and generators $\{H_l\}$ from a native gate alphabet determined from the connectivity and interactions to the specific quantum computer being used.

As shown in Fig. 1(b), throughout this work we will consider the most depth-frugal instantiation of the HEA: the one-dimensional alternating layered HEA. Here, one assumes that the physical qubits in the hardware are organized in a chain, where the i-th qubit can be coupled with the $(i-1)$-th and $(i+1)$-th. Then, at each *layer* of the circuit one connects each qubit with its nearest neighbors in an alternating, brick-like, fashion. We will denote as D the *depth*, or the number of layers, of the ansatz. This type of alternating-layered HEA exploits the native connectivity of the device to maximize the number of operations at each layer while preventing qubits to idle. For instance, alternating-layered HEAs are extremely well suited for the IBM quantum hardware topology where only nearest neighbor qubits are directly connected (see e.g. Ref. [56]). We note that henceforth when we use the term HEA, we will refer to the alternating-layered ansatz of Fig. 1.

3 Trainability of the HEA

3.1 Review of the literature

In recent years, several results about the non-trainability of VQAs/QML have been pointed out [42–51]. In particular, it has been shown that quantum landscapes can exhibit the barren plateau phenomenon, which is nowadays consid-

ered to be one of the most challenging bottlenecks for trainability of these hybrid models. We say that the cost function exhibits a barren plateau if, for the cost, or loss function, the optimization landscape becomes exponentially flat with the number of qubits. When this occurs, an exponential number of measurement shots are required to resolve and determine a cost-minimizing direction. In practice, the exponential scaling in the precision due to the barren plateaus erases the potential quantum advantage, as the VQA or QML scheme will have complexity comparable to the exponential scaling of classical algorithms.

Being more concrete, let $f(\boldsymbol{\theta}) = C(\boldsymbol{\theta}), L_s(\boldsymbol{\theta})$, i.e., either the cost function $C(\boldsymbol{\theta})$ of a VQAs, or the s-th term $L_s(\boldsymbol{\theta})$ in the loss function of a QML settings. For simplicity of notation, we will omit the "s" sub-index of O_s and ψ_s when $f(\boldsymbol{\theta}) = L_s(\boldsymbol{\theta})$. In a barren plateau, there are two types of concentration (or flatness) notions that have been explored: deterministic concentration (all landscape is flat) and probabilistic (most of the landscape is flat). Let us first define the deterministic notion of concentration:

Definition 3. *(Deterministic concentration) Let the trivial value of the cost function be* $f_{trv} := \frac{1}{2^n}\mathrm{Tr}[O]$. *Then the function* $f(\boldsymbol{\theta})$ *is* $\epsilon-$*concentrated iff* $\exists\, \epsilon > 0$ *s.t. for any* $\boldsymbol{\theta}$

$$|f(\boldsymbol{\theta}) - f_{trv}| \leq \epsilon\,. \tag{4}$$

The above definition puts forward a necessary condition for trainability. It is clear that if $f(\boldsymbol{\theta})$ is ϵ-concentrated, then $f(\boldsymbol{\theta})$ must be resolved within an error that scales as $\sim \epsilon$, i.e., one must use $\sim \epsilon^{-2}$ measurement shots to estimate $f(\boldsymbol{\theta})$. Thus, we define a VQA/QML model to be trainable if ϵ vanishes no faster than polynomially with n ($\epsilon \in \Omega(1/\mathrm{poly}(n))$). Conversely, if $\epsilon \in \mathcal{O}(2^{-n})$, one requires an exponential number of measurement shots to resolve the quantum landscape, making the model non-scalable to a higher number of qubits. Deterministic concentration was shown in Refs. [57, 58], which study the performance of VQA and QML models in the presence of quantum noise and prove that $|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(q^D)$, where $0 < q < 1$ is a parameter that characterizes the noise. Using the results therein, it can be shown that if the depth D of the HEA is $D \in \mathcal{O}(\mathrm{poly}(n))$, then the noise acting through the circuit leads to an exponential concentration around the trivial value f_{trv}.

Let us now consider the following definition of probabilistic concentration:

Definition 4 (Probabilistic concentration)**.** *Let* $\langle\cdot\rangle_{\boldsymbol{\theta}}$ *be the average with respect to the parameters* $\theta_1 \in \Theta_1, \theta_2 \in \Theta_2, \ldots$, *for a set of parameter domains* $\{\boldsymbol{\Theta}\}$. *Then the function* $f(\boldsymbol{\theta})$ *is probabilistic concentrated if for any* $\boldsymbol{\theta}'$

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}')] \leq \epsilon\,, \tag{5}$$

where the average is taken over the domains $\{\boldsymbol{\Theta}\}$.

Here we make an important remark on the connection between probabilistic concentration and barren plateaus. The barren plateau phenomenon, as initially formulated in Ref. [42] indicates that the cost function gradients are concentrated, i.e., that $\mathrm{Var}[\partial_\nu f(\boldsymbol{\theta})] \equiv \langle[\partial_\nu f(\boldsymbol{\theta})]^2\rangle_{\boldsymbol{\theta}} - \langle\partial_\nu f(\boldsymbol{\theta})\rangle^2_{\boldsymbol{\theta}} \leq \epsilon$, where $\partial_\nu f(\boldsymbol{\theta}) := \partial f(\boldsymbol{\theta})/\partial\theta_\nu$. However one can prove that probabilistic cost concentration implies probabilistic gradient concentration, and vice-versa [47]. According to Definition 4, we can again see that if $\epsilon \in \mathcal{O}(2^{-n})$, one requires an exponential number of measurement shots to navigate through the optimization landscape. As shown in Ref. [42], such probabilistic concentration can occur if the depth is $D \in \mathcal{O}(\mathrm{poly}(n))$, as at the ansatz becomes a 2-design [42, 59, 60].

From the previous, we know that deep HEAs with $D \in \mathcal{O}(\mathrm{poly}(n))$ can exhibit both deterministic cost concentration (due to noise), but also probabilistic cost concentration (due to high expressibility [46]). However, the question still remains open of whether HEA can avoid barren plateaus and cost concentration with sub-polynomial depths. This question was answered in Ref. [43]. Where it was shown that HEAs can avoid barren plateaus and have trainability guarantees if two necessary conditions are met: *locality and shallowness*. In particular, one can prove that if $D \in \mathcal{O}(\log(n))$, then measuring *global* operators – i.e., O is a sum of operators acting non-identically on every qubit – leads to barren plateaus, whereas measuring *local* operators – i.e. O is a sum of operators acting (at most) on k qubits, for $k \in \mathcal{O}(1)$ – leads to gradients that vanish only polynomially in n.

3.2 A new source for untrainability

The discussions in the previous section provide a sort of recipe for avoiding expressibility-induced probabilistic concentration (see Definition 4), and noise-induced deterministic concentration: *Use local cost measurement operators and keep the depth of the quantum circuit shallow enough.*

Unfortunately, the previous is still not enough to guarantee trainability. As there are other sources of untrainability which are usually less explored. To understand what those are, we will recall a simplified version of the main result in Theorem 2 of Ref. [43]. First, let O act non-trivially

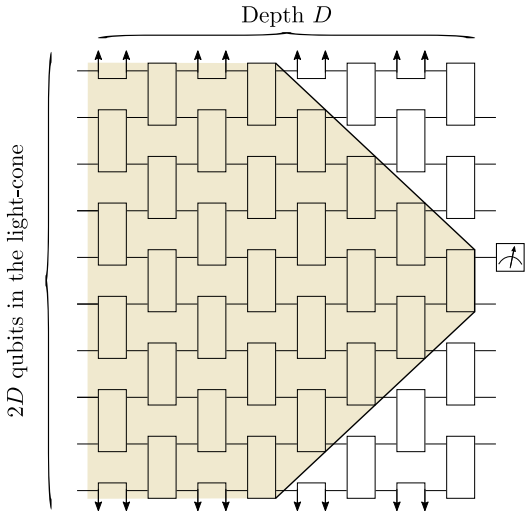

Figure 2: The sketch shows the light-cone of a local measurement operator at the end of a shallow HEA. Here we can see how the support of the local operator grows with the depth D of the HEA. Since the HEA gates act on neighboring qubits, the support increases no more than $2D$.

only on two adjacent qubits, one of them being the $\lfloor \frac{n}{2} \rfloor$-th qubit, and let us study the partial derivative $\partial_\nu f(\boldsymbol{\theta})$ with respect to a parameter in the last gate acting before O (see Fig. 2). The variance of $\partial_\nu f(\boldsymbol{\theta})$ is lower bounded as [43]

$$G_n(D, |\psi\rangle) \leq \mathrm{Var}[\partial_\nu f(\boldsymbol{\theta})]\,, \tag{6}$$

where we recall that D is the depth of the HEA, $|\psi\rangle$ is the input state, and with

$$G_n(D, |\psi\rangle) = \left(\frac{2}{5}\right)^{2D} \frac{2}{225}\, \eta(O) \sum_{\substack{k,k'=n/2-D \\ k'>k}}^{n/2+D} \eta(\psi_{k,k'})\,. \tag{7}$$

Here $\eta(M) = \left\| M - \mathrm{Tr}[M]\frac{\mathbb{1}}{d_M} \right\|_2$, is the Hilbert-Schmidt distance between M and $\mathrm{Tr}[M]\frac{\mathbb{1}}{d_M}$, where d_M is the dimension of the matrix M. Moreover, here we defined $\psi_{k,k'} = \mathrm{Tr}_{1,\ldots,k-1,k'+1,\ldots,n}[|\psi\rangle\langle\psi|]$ as the reduced density matrix on the qubits with index $i \in [k, k']$. As such, $\psi_{k,k'}$ correspond to the reduced states of all possible combinations of adjacent qubits in the *light-cone* generated by O (see Fig. 2). Here, by light-cone we refer to the set of qubit indexes that are causally related to O via $U(\boldsymbol{\theta})$, i.e., the set of indexes over which $U^\dagger(\boldsymbol{\theta}) O U(\boldsymbol{\theta})$ acts non-trivially.

Equation (7) provides the necessary condition to guarantee trainability, i.e., to ensure that the gradients do not vanish exponentially. First, one recovers the condition on the HEA that $D \in \mathcal{O}(\log(n))$. However, a closer inspection of the above formula reveals that both the initial state $|\psi\rangle$ and the measurement operator O also play a key role. Namely one needs that O, as well as *any* of the reduced density matrices of $|\psi\rangle$ an *any* set adjacent qubits in the light-cone, to not be close (in Hilbert-Schmidt distance) to the (normalized) identity matrix. This is due to the fact that if $\eta(O) \in \mathcal{O}(2^{-n})$, or if $\eta(\psi_{k,k'}) \in \mathcal{O}(2^{-n})$ $\forall k, k'$, then $G_n(D, |\psi\rangle) \in \mathcal{O}(2^{-n})$ and the trainability guarantees are lost (the lower bound in Eq. (6) becomes trivial).

The previous results highlight that one should pay close attention to the measurement operator and the input states. Moreover, these results make intuitive sense as they say that extracting information by measuring an operator O that is exponentially close to the identity will be exponentially hard. Similarly, training an ansatz with local gates on a state whose marginals are exponentially close to being maximally mixed will be exponentially hard.

Here we remark that in a practical scenario of interest, one does not expect O to be exponentially close to the identity. For a VQA, one is interested in finding the ground state of O (see Eq. (1)), and as such, it is reasonable to expect that O is non-trivially close to the identity [5, 6]. Then, for QML there is additional freedom in choosing the measurement operators O_s in (2), meaning that one simply needs to choose an operator with non-exponentially vanishing support in non-identity Pauli operators.

In the following sections, we will take a closer look at the role that the input state can have in the trainability of shallow-depth HEA.

4 Entanglement and information scrambling

Here we will briefly recall two fundamental concepts: that of states satisfying an area law of entanglement, and that of states satisfying a volume law of entanglement. Then, we will relate the concept of area law of entanglement with that of scrambling.

First, let us rigorously define what we mean by area and volume laws of entanglement.

Definition 5 (Volume law vs. area law)**.** *Let $|\psi\rangle$ be a state in a bipartite Hilbert space $\mathcal{H}_\Lambda \otimes \mathcal{H}_{\bar{\Lambda}}$. Let Λ be a subsystem composed of $|\Lambda|$ qubits, and let $\bar{\Lambda}$ be its complement set. Let $\psi_\Lambda = \mathrm{Tr}_{\bar{\Lambda}}[|\psi\rangle\langle\psi|]$ be the reduced density matrix on Λ. Then, the state $|\psi\rangle$ possesses volume law for the entanglement within*

Λ *and* $\bar{\Lambda}$ *if*

$$|\Lambda| - S(\psi_\Lambda) \in \mathcal{O}\left(\frac{1}{2^{|\bar{\Lambda}|-|\Lambda|}}\right), \tag{8}$$

where $S(\rho) = -\mathrm{Tr}[\rho \log(\rho)]$ *is the entropy of entanglement. Conversely, the state possesses area law for the entanglement within* Λ *and* $\bar{\Lambda}$ *if*

$$|\Lambda| - S(\psi_\Lambda) \in \Omega\left(\frac{1}{\mathrm{poly}(|\Lambda|)}\right). \tag{9}$$

Note that the above definition of area vs. volume law of entanglement is nonstandard. In particular, it is completely agnostic to the geometry on which the given state resides. However, as we will show, it is the relevant one to look at for the purpose of this work.

From the definition of volume law of entanglement, the concept of scrambling of quantum information can be easily defined; the information contained in $|\psi\rangle$ is said to be scrambled throughout the system if the state $|\psi\rangle$ follows a volume law for the entropy of entanglement according to Definition 5 across any bipartition such that $|\Lambda| \in \mathcal{O}(\log(n))$.

Here we further recall that an information-theoretic measure of the quantum information that can be extracted by a subsystem Λ is

$$\mathcal{I}_\Lambda(\psi) \coloneqq \left\| \psi_\Lambda - \frac{\mathbb{1}_\Lambda}{2^{|\Lambda|}} \right\|_1, \tag{10}$$

which quantifies the maximum distinguishability between the reduced density matrix ψ_Λ and the maximally mixed state $\mathbb{1}_\Lambda/2^{|\Lambda|}$. To see this, let O_Λ be a local operator acting non-trivially on Λ, whose maximum eigenvalue is one, then $|\langle\psi|O_\Lambda|\psi\rangle - 2^{-n}\mathrm{Tr}[O_\Lambda]| \leq \mathcal{I}_\Lambda(\psi)$. Thus, if $\mathcal{I}_\Lambda$ is exponentially small, then the measurement of the local operator O_Λ is not able to efficiently reveal any information contained in $|\psi\rangle$. We thus lay down the following formal definition of information scrambling:

Definition 6 (Information scrambling)**.** *The quantum information in* $|\psi\rangle$ *is scrambled iff for any subset of qubits* Λ *such that* $|\Lambda| \in \mathcal{O}(\log(n))$*:*

$$\mathcal{I}_\Lambda(\psi) \in \mathcal{O}(2^{-cn}). \tag{11}$$

for some constant $c > 0$*.*

The definition of scrambling of quantum information easily follows from the definition of volume law for entanglement in Definition 5. Indeed, given a subsystem Λ, one has the following bound

$$\mathcal{I}_\Lambda(\psi) \leq \sqrt{2(|\Lambda| - S(\psi_\Lambda))}, \tag{12}$$

and thus, if $|\psi\rangle$ follows a volume law of entanglement according to Definition 5 for any bipartition $\Lambda \cup \bar{\Lambda}$ with $|\Lambda| \in \mathcal{O}(\log(n))$, one indeed has the exponential suppression of the information contained in Λ, i.e., $\mathcal{I}_\Lambda(\psi) \in \mathcal{O}(2^{-n})$. This motivates us to propose the following alternative definition for states following volume and area law of entanglement

Definition 7 (Volume law vs. area law)**.** *Let* $|\psi\rangle$ *be a state in a bipartite Hilbert space* $\mathcal{H}_\Lambda \otimes \mathcal{H}_{\bar{\Lambda}}$*. Let* Λ *be a subsystem composed of* $|\Lambda|$ *qubits, and let* $\bar{\Lambda}$ *be its complement set. Let* $\psi_\Lambda = \mathrm{Tr}_{\bar{\Lambda}}[|\psi\rangle\langle\psi|]$ *be the reduced density matrix on* Λ*. Then the state* $|\psi\rangle$ *possesses volume law for the entanglement within* Λ *and* $\bar{\Lambda}$ *if*

$$\mathcal{I}_\Lambda(\psi) \in \mathcal{O}\left(\frac{1}{2^{cn}}\right), \tag{13}$$

for some $c > 0$*. Conversely, the state possesses area law for the entanglement within* Λ *and* $\bar{\Lambda}$ *if*

$$\mathcal{I}_\Lambda(\psi) \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right). \tag{14}$$

5 HEA and the role of entanglement

Let us consider a VQA or QML task from Definitions 1–2, where the ansatz for the parametrized quantum circuit $U(\boldsymbol{\theta})$ is a shallow HEA with depth $D \in \mathcal{O}(\log(n))$ (see Fig. 1(b)). Moreover, let the function $f(\boldsymbol{\theta}) = C(\boldsymbol{\theta}), L_s(\boldsymbol{\theta})$, be either the cost or loss function in Eqs. (1) and (2).

5.1 HEA and volume law of entanglement

As we show in this section, a shallow HEA will be untrainable if the input state satisfies a volume law of entanglement according to Definition 5. Before going deeper into the technical details, let us sketch the idea behind our statement, with the following warm-up example.

5.1.1 A toy model

Let us consider for simplicity the case when O is a local operator acting non-trivially on a single qubit, and let us recall that $f(\boldsymbol{\theta}) = \mathrm{Tr}[U(\boldsymbol{\theta})\,|\psi\rangle\langle\psi|\,U^\dagger(\boldsymbol{\theta})O]$. In the Heisenberg picture we can interpret $f(\boldsymbol{\theta})$ as the expectation value of the backwards-in-time evolved operator $O(\boldsymbol{\theta}) = U^\dagger(\boldsymbol{\theta})OU(\boldsymbol{\theta})$ over the initial state $|\psi\rangle$. Thanks to the brick-like structure of the HEA, one can see from a simple geometrical argument that the operator $O(\boldsymbol{\theta})$ will act non-trivially only on a set Λ containing (at most) $2D$ qubits (see Fig. 2, and

also see below for the rigorous proof). Thus, we can compute $f(\boldsymbol{\theta})$ as

$$f(\boldsymbol{\theta}) = \mathrm{Tr}_{\Lambda}[\mathrm{Tr}_{\bar{\Lambda}}[|\psi\rangle\langle\psi|]O(\boldsymbol{\theta})] \,, \tag{15}$$

where $\bar{\Lambda}$ is the complement set of Λ, and where we assume $|\Lambda| \ll |\bar{\Lambda}|$. Since the HEA is shallow, the cost function is evaluated by tracing out the majority of qubits, i.e. $|\bar{\Lambda}| \sim n$. If the input state $|\psi\rangle$ is highly entangled, since $|\Lambda| \ll |\bar{\Lambda}|$, and thanks to the monogamy of entanglement [61], we can assume that there is a subset of $\bar{\Lambda}$, say Λ', maximally entangled with Λ, i.e., $|\psi\rangle = \sqrt{\frac{1-\epsilon^2}{2^{|\Lambda|}}} \sum_{i=1} |i_{\Lambda}\rangle \otimes |i_{\Lambda'}\rangle \otimes |rest\rangle + \epsilon\, |\phi\rangle$, where $\epsilon \ll 1$ and $|\phi\rangle$ being orthogonal to the rest. Neglecting terms in ϵ^2, and choosing $\|O\|_{\infty} = 1$, this results in a function 2ϵ-concentrated around its trivial value f_{trv}

$$|f(\boldsymbol{\theta}) - f_{trv}| \leq 2\epsilon \,. \tag{16}$$

The previous shows that a highly entangled state, such as the one presented above which satisfies a volume law-of entanglement, will lead to a landscape that exhibits a deterministic exponential concentration according to Definition 3.

5.1.2 Formal statement

In this section, we will present a deterministic concentration result. To begin, let us introduce the following definition:

Definition 8 (Support of a Pauli operator)**.** *Let P be a Pauli operator, we define the support* $\mathrm{supp}(P)$ *as the ordered set of natural numbers q_i labeling the qubits on which P acts non-trivially.*

$$\mathrm{supp}(P) = \{q_1, \ldots, q_{S_i} \,|\, q_1 < \ldots < q_S, q_k \in [1, n]\} \tag{17}$$

with S the number of qubits on which P acts non-trivially.

We are finally ready to state a deterministic concentration result based on the information-theoretic measure $\mathcal{I}_{\Lambda}(\psi)$ for HEA circuits and for $f(\boldsymbol{\theta}) = C(\boldsymbol{\theta}), L_s(\boldsymbol{\theta})$ being the VQA cost function or the QML loss function.

Theorem 1 (Concentration and measurement operator support)**.** *Let $U(\boldsymbol{\theta})$ be HEA with depth D, and $O = \sum_i c_i P_i$, with P_i Pauli operators. Now, let $\Lambda_i \coloneqq \mathrm{supp}(U^{\dagger}(\boldsymbol{\theta}) P_i U(\boldsymbol{\theta}))$:*

$$|f(\boldsymbol{\theta}) - f_{trv}| \leq \sum_i |c_i| \mathcal{I}_{\Lambda_i}(\psi) \tag{18}$$

moreover, $\sum_i |c_i| \leq (3n)^{\max_i |\mathrm{supp}(P_i)|} \|O\|_{\infty}$. Let $\Lambda \coloneqq \max_i \Lambda_i$. Then, the following bound on the size of Λ holds:

$$|\Lambda| \leq 2D \times \max_i |\mathrm{supp}(P_i)| \,. \tag{19}$$

See App. A for the proof.

Let us discuss the implications of Theorem 1. First, we find that the difference between the training function $f(\boldsymbol{\theta})$, and its trivial value f_{trv} depends on the information-theoretic measure of information scrambling $\mathcal{I}_{\Lambda}(\psi)$ for $|\Lambda| = \max_i |\Lambda_i|$ (see Definition 6). From Eq. (19) it is clear that the size of Λ is determined by two factors: (i) the depth of the circuit D, and (ii) the locality of the operator O. As soon as either the depth D or $\max_i \mathrm{supp}(P_i)$ starts scaling with the number of qubits n, the bound in Eq. (18) becomes trivial as one can obtain information by measuring a large enough subsystem with $|\Lambda| \in \Theta(n)$. However, as explained above in Sec. 3, we already know that this regime is precluded as the necessary requirements to ensure the trainability of the HEA (and thus to ensure trainability of the VQA/QML model) are: (i) the depth of the HEA circuit must not exceed $\mathcal{O}(\log(n))$, and (ii) the operator O must have local support on at most $\mathcal{O}(\log(n))$ qubits. Hence, the trainability of the model is solely determined by the scaling of $\mathcal{I}_{\Lambda}(\psi)$.

From Theorem 1, we can derive the following corollaries

Corollary 1. *Let the depth D of the HEA be $D \in \mathcal{O}(\log(n))$, and let $\max_i |\mathrm{supp}(P_i)| \in \mathcal{O}(\log(n))$. Then, if $|\psi\rangle$ satisfies a volume law, or alternatively if the information contained in $|\psi\rangle$ is scrambled, i.e., if $\mathcal{I}_{\Lambda}(\psi) \in \mathcal{O}(2^{-cn})$ for some $c > 0$, and if $\|O\|_{\infty} \in \Omega(1)$, then:*

$$|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(2^{-nc/2}) \,. \tag{20}$$

Here we can see that if the information contained in $|\psi\rangle$ is too scrambled throughout the system, one has deterministic exponential concentration of cost values according to Definition 3.

Theorem 1 puts forward another important necessary condition for trainability and to avoid deterministic concentration: *the information in the input state must not be too scrambled throughout the system.* When this occurs, the information in $|\psi\rangle$ cannot be accessed by local measurements, and hence one cannot train the shallow depth HEA.

At this point, we ask the question of how typical is for a state to contain information scrambled throughout the system, hidden in non-local degrees of freedom, and resulting in $\mathcal{I}_{\Lambda}(\psi) \in \mathcal{O}(2^{-n})$. To answer this question, we use tools of the Haar measure and show that for the overwhelming majority of states, the information cannot be accessed by local measurements as their information is too scrambled.

Corollary 2. *Let $|\psi\rangle \sim \mu_{Haar}$ be a Haar random state, $D \in \mathcal{O}(\log(n))$, and $\max_i \mathrm{supp}(P_i) \in$*

$\mathcal{O}(\log(n))$. *Here* μ_{Haar} *is the uniform Haar measure over the states in the Hilbert space. Then* $\mathcal{I}_\Lambda(|\psi\rangle) \leq 2^{|\Lambda|-n/3}$, *with overwhelming probability* $\geq 1 - e^{-c2^{n/3-1}+(2n+1)\ln 2}$ *with* $c = (18\pi^3)^{-1}$, *and thus*

$$|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(2^{-n/6})\,. \tag{21}$$

See App. A for a proof. Note that henceforth, we will refer to *overwhelming probability* as a probability 1 up to a exponentially (in n, the size of the system) decaying correction. In many tasks, multiple copies of a quantum state $|\psi\rangle$ are used to predict important properties, such as entanglement entropy [62–64], quantum magic [65–67], or state discrimination [68]. Thus, it is worth asking whether a function of the form $f(\boldsymbol{\theta}) = \mathrm{Tr}[U(\boldsymbol{\theta})\,|\psi\rangle\langle\psi|^{\otimes 2}\,U^\dagger(\boldsymbol{\theta})O]$ can be trained when $U(\boldsymbol{\theta})$ is a shallow HEA acting on $2n$ qubits. In the following corollary, we prove that for the overwhelming majority of states, $\mathcal{I}_\Lambda(\psi^{\otimes 2}) \in \mathcal{O}(2^{-n})$, and one has deterministic concentration according to Definition 3 even if one has access to two copies of a quantum state.

Corollary 3. *Suppose one has access to* 2 *copies of a Haar random state* $|\psi\rangle$ *and one computes the function* $f(\boldsymbol{\theta}) = \mathrm{Tr}[U(\boldsymbol{\theta})\,|\psi\rangle\langle\psi|^{\otimes 2}\,U^\dagger(\boldsymbol{\theta})O]$. *Let* $D \in \mathcal{O}(\log(n))$ *be the depth of a HEA* $U(\boldsymbol{\theta})$, *and* $\max_i \mathrm{supp}(P_i) \in \mathcal{O}(\log(n))$. *Then* $\mathcal{I}_\Lambda(|\psi\rangle^{\otimes 2}) \leq 2^{|\Lambda|-n/3}$, *with overwhelming probability* $\geq 1 - e^{-c2^{n/3-4}+(2n+1)\ln 2}$, *where* $c = (18\pi^3)^{-1}$, *and one has:*

$$|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(2^{-n/6})\,. \tag{22}$$

See App. A for a proof. Note that the generalization to more copies is straightforward.

The above results show us that there is indeed a no-free-lunch for the shallow HEA. The majority of states in the Hilbert space, follow a volume law for the entanglement entropy and thus have quantum information hidden in highly non-local degrees of freedom, which cannot be accessed through local measurement at the output of a shallow HEA.

5.2 HEA and area law of entanglement

The previous results indicate that shallow HEAs are untrainable for states with a volume law of entanglement, i.e., they are untrainable for the vast majority of states. The question still remains of whether shallow HEA can be used if the input states follow an area law of entanglement as in Definition 5. Surprisingly, we can show that in this case there is no concentration, as the following result holds

Theorem 2 (Anti-concentration of expectation values)**.** *Let* $U(\boldsymbol{\theta})$ *be a shallow HEA with depth* $D \in \mathcal{O}(\log(n))$ *where each local two-qubit gate forms a* 2*-design on two qubits. Then, let* $O = \sum_i c_i P_i$ *be the measurement composed of, at most, polynomially many traceless Pauli operators* P_i *having support on at most two neighboring qubits, and where* $\sum_i c_i^2 \in \mathcal{O}(\mathrm{poly}(n))$. *If the input state follows an area law of entanglement, for any set of parameters* $\boldsymbol{\theta_B}$ *and* $\boldsymbol{\theta_A} = \boldsymbol{\theta_B} + \hat{\boldsymbol{e}}_{AB} l_{AB}$ *with* $l_{AB} \in \Omega(1/\mathrm{poly}(n))$, *then*

$$\mathrm{Var}_{\boldsymbol{\theta}_B}[f(\boldsymbol{\theta}_A) - f(\boldsymbol{\theta}_B)] \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)\,. \tag{23}$$

See App. B for the proof.

Theorem 2 shows that if the input states to the shallow HEA follow an area law of entanglement, then the function $f(\boldsymbol{\theta})$ anti-concentrates. That is, one can expect that the loss function values will differ (at least polynomially) at sufficiently different points of the landscape. This naturally should imply that the cost function does not have barren plateaus or exponentially vanishing gradients. In fact, we can prove this intuition to be true as it can be formalized in the following result.

Proposition 1. *Let* $f(\boldsymbol{\theta})$ *be a VQA cost function or a QML loss function where* $U(\boldsymbol{\theta})$ *is a shallow HEA with depth* $D \in \mathcal{O}(\log(n))$. *If the values of* $f(\boldsymbol{\theta})$ *anti-concentrate according to Theorem 2 and Eq. (23), then* $\mathrm{Var}_{\boldsymbol{\theta}}[\partial_\nu f(\boldsymbol{\theta})] \in \Omega(1/\mathrm{poly}(n))$ *for any* $\theta_\nu \in \boldsymbol{\theta}$ *and the loss function does not exhibit a barren plateau. Conversely, if* $f(\boldsymbol{\theta})$ *has no barren plateaus, then the cost function values anti-concentrate as in Theorem 2 and Eq. (23).*

See App. B for the proof.

Taken together, Theorem 2 and Proposition 1 suggest that shallow HEAs are ideal for processing states with area law of entanglement, as the loss landscape is immune to barren plateaus. Evidently, the previous shallow HEAs are capable of achieving a quantum advantage. However, determining whether a quantum advantage is feasible or not, for such ansatzes is beyond the scope of this work (as it requires a detailed analysis of properties beyond the absence of barren plateaus such as quantifying the presence of local minima) we can still further identify scenarios where a quantum advantage could potentially exist.

First, let us rule out certain scenarios where a provable quantum advantage will be unlikely. These correspond to cases where the input state $|\psi\rangle$ satisfies an area law of entanglement but also admits an efficient classical representation [69–73].

The key issue here is that if the input state admits a classical decomposition, then the expectation value $f(\boldsymbol{\theta})$ for $U(\boldsymbol{\theta})$ being a shallow HEA can be efficiently classically simulated [74]. For instance, one can readily show that the following result holds.

Proposition 2 (Cost of classically computing $f(\boldsymbol{\theta})$)**.** *Let $U(\boldsymbol{\theta})$ be an alternating layered HEA of depth D, and $O = \sum_i c_i P_i$. Let $|\psi\rangle$ be an input stat that admits a Matrix Product State [75] (MPS) description with bond-dimension χ. Then, there exists a classical algorithm that can estimate $f(\boldsymbol{\theta})$ with a complexity which scales as $\mathcal{O}((\chi \cdot 4^D)^3)$.*

The proof of the above proposition can be found in [75]. From the previous theorem, we can readily derive the following corollary.

Corollary 4. *Shallow depth HEAs with depth $D \in \mathcal{O}(\log(n))$, and with an input state with a bond-dimension $\chi \in \mathcal{O}(\text{poly}(n))$ can be efficiently classically simulated with a complexity that scales as $\mathcal{O}(\text{poly}(n))$.*

Note that Proposition 2 and its concomitant Corollary 4 do not preclude the possibility that shallow HEA can be useful even if the input state admits an efficient classical description. This is due to the fact that, while requiring computational resources that scale polynomially with n (if χ is at most polynomially large with n), the order of the polynomial can still lead to prohibitively large (albeit polynomially growing) computational resources. Still, we will not focus on discussing this fine line, instead, we will attempt to find scenarios where a quantum advantage can be achieved.

In particular, we highlight the seminal work of Ref. [76], which indicates that while states satisfying an area law of entanglement constitute just a very small fraction of all the states (which is expected from the fact that Haar random states –the vast majority of states– satisfy a volume law), the subset of such area law of entanglement states that admit an efficient classical representation is exponentially small. This result can be better visualized in Fig. 3. The previous gives hope that one can achieve a quantum advantage with *area law classically-unsimulable states*.

6 Implications of our results

Let us here discuss how our results can help identify scenarios where shallow HEA can be useful, and scenarios where they should be avoided.

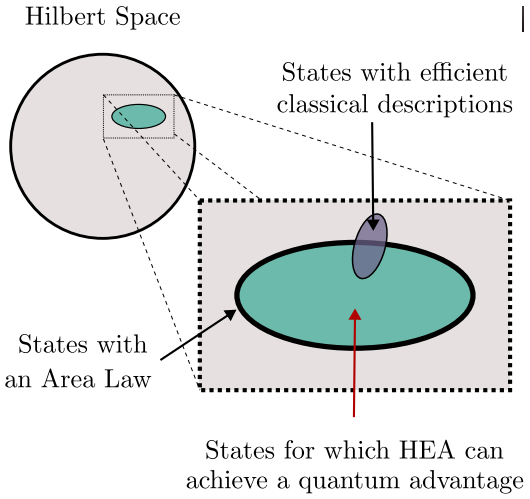

Figure 3: Schematic representation of the Hilbert space. The vast majority of states satisfy a volume law, and hence a shallow HEA cannot be used to extract information from them. From the set of states satisfying an area law, only a very small subset admits an efficient classical representation. For these states, the effect of a shallow HEA can be efficiently simulated. As such, there exists a Goldilocks regime where HEA can potentially be used to achieve a quantum advantage: non-classically-simulable area law states.

6.1 Implications to VQAs

As indicated in Definition 1, in a VQA one initializes the circuit to some easy-to-prepare fiduciary quantum state $|\psi_0\rangle$. For instance, in a variational quantum eigensolver [7] quantum chemistry application such an initial state is usually the unentangled mean-field Hartree-Fock state [54]. Similarly, when solving a combinatorial optimization task with the quantum optimization approximation algorithm [9] the initial state is an equal superposition of all elements in the computational basis $|+\rangle^{\otimes n}$. In both of these cases, the initial states are separable, satisfy an area law, and admit an efficient classical decomposition. This means that while the shallow HEA will be trainable, it will also be classically simulable. This situation will arise for most VQA implementations as it is highly uncommon to prepare non-classically simulable initial states. From the previous, we can see that shallow HEA should likely be avoided in VQA implementations if one seeks to find a quantum advantage.

6.2 Implications to QML

In a QML task according to Definition 2, one sends input states from some datasets into the shallow HEA. As shown in Fig. 4, the input states are

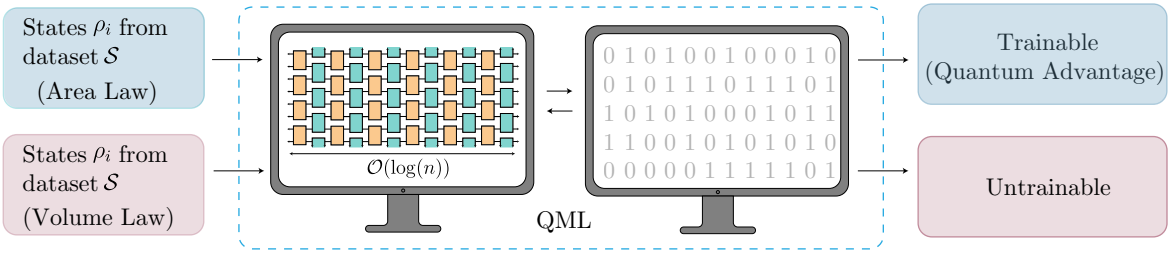

Figure 4: Schematic representation of the trainability of two QML tasks with two different datasets $\mathcal{S}$. The first task is trainable since the dataset $\mathcal{S}$ is composed of states ρ_i possessing an area law of entanglement. Conversely, the second task is untrainable being the dataset $\mathcal{S}$ composed of states ρ_i possessing volume law of entanglement. We remark that the first task can enjoy a quantum advantage since not all the area-law states are classically simulable, see Ref. [76] and Fig. 3.

problem-dependent, implying that the usability of the HEA depends on the task at hand. Our results indicate that HEAs should be avoided when the input states satisfy a volume law of entanglement or when they follow an area law but also admit an efficient classical description. In fact, it is clear that while the HEA is widely used in the literature, most cases where it is employed fall within the cases where the HEA should be avoided [45]. As such, we expect that many proposals in the literature should be revised. However, the trainability guarantees pointed out in this work, narrow down the scenarios where the HEA should be used, and leave the door open for using shallow HEAs in QML tasks to analyze non-classically simulable area-law states. In the following section, we give an explicit example, based on state discrimination between area law states having no MPS decomposition, with a possible achievable quantum advantage.

7 Random Hamiltonian discrimination

7.1 General framework

In this section, we present an application of our results in a QML setting based on Hamiltonian Discrimination. The QML problem is summarized as follows: the data contains states that are obtained by evolving an initial state either by a general Hamiltonian or by a Hamiltonian possessing a given symmetry. The goal is to train a QML model to distinguish between states arising from these two evolutions. In the example below, we show how the role of entanglement governs the success of the QML algorithm.

Let us begin by formally stating the problem. Consider two Hamiltonians H_G, H_S, and a local operator S being a symmetry for H_S, i.e. $[H_S, S] = 0$. Let $\mathcal{V}_S \coloneqq \text{span}\{|z\rangle \in \mathbb{C}^{2\otimes n} \,|\, S\,|z\rangle = |z\rangle\}$, i.e. the linear subspace filled by eigenvectors of S. Let us assume that $\dim(\mathcal{V}_S) \geq 2$. Then, we build the dataset following Algorithm 1.

Algorithm 1 Build the dataset $\mathcal{S}$

$\mathcal{S} = \emptyset$
s=0
while $s \leq N$ **do**
  take $|z_s\rangle \in \mathcal{V}_S$
  $|\psi_s\rangle_t = \exp(-iH_S t)\,|z_s\rangle$
  $(1, |\psi_s^{H_S}\rangle_t) \in \mathcal{S}$
  $|\psi_{s+1}\rangle_t = \exp(-iH_G t)\,|z_s\rangle$
  $(0, |\psi_{s+1}^{H_G}\rangle_t) \in \mathcal{S}$
  s=s+2
end while

We consider the case when $U(\boldsymbol{\theta})$ is a parametrized shallow HEA, and is O a local operator measured at the output of the circuit. We define

$$L_s(H, \boldsymbol{\theta}, t) \coloneqq \text{Tr}[|\psi_s^H\rangle_t\langle\psi_s^H|\, O(\boldsymbol{\theta})]\,, \qquad (24)$$

where $|\psi_s^H\rangle_t \in (y_s, |\psi_s^H\rangle_t) \in \mathcal{S}$ for $H \in \{H_G, H_S\}$, and $O(\boldsymbol{\theta}) \coloneqq U^\dagger(\boldsymbol{\theta})OU(\boldsymbol{\theta})$. Here, $y_s = 1(0)$ if $H = H_S(H_G)$. In the following, we will drop the superscript in $|\psi_s\rangle \in \mathcal{S}$ to light the notation, unless necessary. Then, the goal is to minimize the empirical loss function:

$$L(\boldsymbol{\theta}, t) = \frac{1}{N} \sum_{(y_s, |\psi_s\rangle) \in \mathcal{S}} \left(y_s - L_s(H, \boldsymbol{\theta}, t)\right)^2 \,, \qquad (25)$$

where N is the size of the dataset $\mathcal{S}$. There are two necessary conditions for the success of the algorithm: (i) the parameter landscape is not exponentially concentrated around its trivial value, and (ii) there exists $\boldsymbol{\theta}_0$ such that the model outputs

are different for data in distinct classes. For instance, this can be achieved if $U^\dagger(\boldsymbol{\theta}_0)OU(\boldsymbol{\theta}_0) = S$; as here $L_s(H_S, \boldsymbol{\theta}_0, t) = 1$ for any s such that $|\psi_s\rangle \equiv |\psi_s^{H_S}\rangle_t$. Then, one also needs to have $L_s(H_G, \boldsymbol{\theta}_0, t)$ not being close to one with high probability. Note that if the symmetry S is a local operator, and O is chosen to be local, there are cases in which a shallow-depth HEA can find the solution $U^\dagger(\boldsymbol{\theta}_0)OU(\boldsymbol{\theta}_0) = S$. Such an example is shown below.

7.2 Gaussian Diagonal Ensemble Hamiltonian discrimination

Let us now specialize the example to an analytically tractable problem. We first show how the growth of the evolution time t, and thus the entanglement generation, affects the HEA's ability to solve the task. Then, we show that there exists a critical time t^* for which the states in the dataset satisfy an area law, and thus for which the QML algorithm can succeed. Since classically simulating random Hamiltonian evolution is a difficult task, the latter constitutes an example where a QML algorithm can enjoy a quantum speed-up with respect to classical machine learning.

Let H_G be a random Hamiltonian, i.e., $H_G = \sum_i E_i \Pi_i$, where Π_i are projectors onto random Haar states, and E_i are normally distributed around 0 with standard deviation 1/2, (see App. C for additional details). This ensemble of random Hamiltonians is called Gaussian Diagonal Ensemble (GDE), and it is the simplest, non-trivial example where our results apply. In Fig. 5 we explicitly show how the time evolution under such a Hamiltonian can be implemented in a quantum circuit. Generalizations to wider used ensembles, such as Gaussian Unitary Ensemble (GUE), Gaussian Symplectic Ensemble (GSE), Gaussian Orthogonal Ensemble (GOE), or the Poisson Ensemble (P), will be straightforward. We refer the reader to Refs. [77, 78] for more details on these techniques.

Consider a bipartition of n qubits, i.e., $A \cup B$ such that $|A| \ll |B|$. Let H_S a Random Hamiltonian commuting with all the operators on a local subsystem A, i.e. $[H_S, P_A] = 0$ for all P_A. We can choose H_S as $H_S = \mathbb{1}_A \otimes H_B$, with H_B belonging to the GDE in the subsystem B. Let H_G be a random Hamiltonian belonging to the GDE ensemble in the subsystem $A \cup B$. Since the Hamiltonian H_S commutes with all the operators in A, we choose the symmetry S to be $S \equiv P_A \otimes \mathbb{1}_B$, i.e. a Pauli operator with local support on A. To build the data-set $\mathcal{S}$, we thus identify the vector space containing all the eigenvectors with eigenvalue 1 of P_A, $\mathcal{V}_{P_A} = \text{span}\{|z\rangle \in \mathbb{C}^{2\otimes n} \,|\, P_A\,|z\rangle = |z\rangle\}$ and follow Algorithm 1. Note that, with this choice, $\dim(\mathcal{V}_{P_A}) = 2^{|A|-1}$ and thus we take $|A| \geq 2$. The QML task is to distinguish states evolved in time by H_G or by H_S. Let us choose O a Pauli operator having support on a local subsystem. Then, the following proposition holds:

Proposition 3. *Let $L_s(H, \boldsymbol{\theta}, t)$ be the expectation value defined in Eq. (24), for $H \in \{H_G, H_S\}$. If there exist $\boldsymbol{\theta}_0$ such that $U^\dagger(\boldsymbol{\theta}_0)OU(\boldsymbol{\theta}_0) = S$, then*

$$L_s(H_S, \boldsymbol{\theta}_0, t) = 1, \quad \forall t\,, \tag{26}$$

$$\mathbb{E}_{\text{GDE}}[L_s(H_G, \boldsymbol{\theta}, t)] = e^{-t^2/4} + \mathcal{O}(2^{-n}), \quad \forall \boldsymbol{\theta}\,. \tag{27}$$

See App. C for the proof.

Notably, the symmetry of H_S ensures that if the HEA is able to find $\boldsymbol{\theta}_0$, then the output $L_s(H_S, \boldsymbol{\theta}_0, t) = 1$ is distinguishable from the expected value of $L_s(H_G, \boldsymbol{\theta}_0, t)$ which is exponentially suppressed in t. While in principle it is possible to minimize the loss function $L(\boldsymbol{\theta}, t)$ for any t, the following theorem states that as the time t grows, the parameter landscape get more and more concentrated, according to Definition 3.

Theorem 3 (Concentration of loss for GDE Hamiltonians)**.** *Let $L_s(H, \boldsymbol{\theta}, t)$ be the expectation value defined in Eq. (24) for $H \in \{H_G, H_S\}$. For random* GDE *Hamiltonians one has*

$$\Pr\left(|L_s(H, \boldsymbol{\theta}, s)| \leq \epsilon\right) \geq 1 - \epsilon^{-1}\, e^{-t^2/4} + \mathcal{O}(2^{-n})\,. \tag{28}$$

See App. C for the proof.

Note the above concentration bound holds for both H_S and H_G, provided that $|A| \ll |B|$. From the above, one can readily derive the following corollary:

Corollary 5. *Let $L_s(H, \boldsymbol{\theta}, t)$ be the expectation value defined in Eq. (24) for $H \in \{H_G, H_S\}$. Then for $t \geq 4\alpha\sqrt{n}/\log_2 e$, $\alpha > 0$, and $\epsilon = e^{-\beta n}$, then:*

$$\Pr\left(|L_s(H, \boldsymbol{\theta}, s) - L_{trv}| \leq 2^{-\beta n}\right) \geq 1 - 2^{-(\alpha-\beta)n} \tag{29}$$

for any $\beta < \alpha$. See App. C

Taken together, Theorem 3 and Corollary 5, provide a no-go theorem for the success of the QML task, as they indicate that beyond $t \sim \sqrt{n}$ one encounters deterministic concentration with overwhelming probability. Crucially, the role of the entanglement generated by $H \in \{H_G, H_S\}$ is hidden in the variable t of the bound in Theorem 3. Indeed, as shown in Refs. [77, 78] the entanglement for random GDE Hamiltonians is monotonically growing with t.

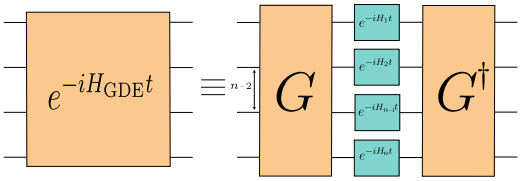

Figure 5: Circuit representation of the exponentiation of a GDE Hamiltonian. The light-blue gates $e^{-iH_k t}$, are single qubit gates, exponentiation of $H_k \equiv \epsilon_k |0_k\rangle\langle 0_k| + \epsilon_{k+1} |1_k\rangle\langle 1_k|$. Labeling $i_k \in \{0_k, 1_k\}$, the numbers $E_{\mathrm{dec}(\boldsymbol{i})} = \sum_{k=0}^{n-1} \epsilon_{k+i_k}$ are 2^n random numbers sampled from $\mathcal{N}(0, 1/2)$, and corresponds to the eigenvalues of H_{GDE}; where $\mathrm{dec}(\boldsymbol{i}) = \sum_k 2^k i_k$. G is a deep-random quantum circuit, building the eigenvectors of H_{GDE}.

While the previous results indicate that a HEA-based QML model will fail on the random Hamiltonian discrimination QML task for $t \sim \sqrt{n}$ (due to high levels of entanglement), this does not preclude the possibility of the model succeeding for smaller evolution times. Notably, here we can show that for $t = \mathcal{O}(\sqrt{\log(n)})$ the conditions are ideal for a quantum advantage: the states in the dataset will satisfy an area law of entanglement, and since GDE Hamiltonians are built out of a very deep random quantum circuit, their time evolution can be classically hard. In particular, the following theorem holds.

Theorem 4 (Area law and short-time evolution under GDE Hamiltonians)**.** *Let H_G be a* GDE *Hamiltonian, and let $|\psi_0\rangle$ be any factorized state over the bipartition $\Lambda \cup \bar{\Lambda}$, with $|\psi_t\rangle = \exp(-iHt)|\psi_0\rangle$ being its unitary evolution under H_G. Let Λ be a subsystem, where $|\Lambda| = \mathcal{O}(\log(n))$. Then,*

$$\Pr\left[\mathcal{I}_\Lambda(\psi_t) \geq \frac{e^{-t^2/8}}{2}\sqrt{1-2^{-|\Lambda|}}\right] \geq 1 - 2^{-n/2}\,, \tag{30}$$

thus for $t = \mathcal{O}(\sqrt{\log(n)})$, the evolved state satisfies area law of the entanglement with overwhelming probability

$$\Pr\left[\mathcal{I}_\Lambda(\psi_t) \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)\right] \geq 1 - 2^{-n/2}\,. \tag{31}$$

See App. C for the proof. Moreover, the following corollary holds

Corollary 6. *Let H_G be a* GDE *Hamiltonian, and let $|\psi_0\rangle$ a factorized state, with $|\psi_t\rangle = \exp(-iHt)|\psi_0\rangle$ being its unitary evolution under H_G. Let Λ be a subsystem, where $|\Lambda| = \mathcal{O}(\log(n))$. Then the probability that the loss function $L(\boldsymbol{\theta}, t)$ anti-concentrating is overwhelming.*

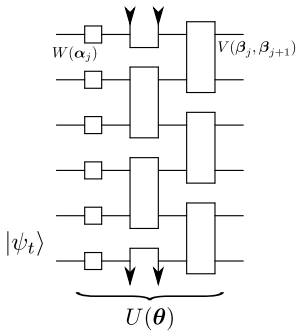

Figure 6: Circuit of HEA used in numerical simulations. The architecture is composed of an initial layer of general single qubit unitaries. We denote these gates as $W(\boldsymbol{\alpha}_j) \equiv e^{-i\boldsymbol{\sigma}_j \cdot \boldsymbol{\alpha}_j}$, and we parametrize them by 3 angles $\boldsymbol{\alpha}_j \equiv (\alpha_j^x, \alpha_j^y, \alpha_j^z)$. Here $\boldsymbol{\sigma}_j = (X_j, Y_j, Z_j)$. Then, two-qubit gates act on neighboring pairs of qubits. Each two-qubit gate denoted as $V_{j,j+1}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j+1})$, is composed of a CNOT followed by a general single qubit gate on each qubit. That is, $V_{j,j+1}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j+1}) = W(\boldsymbol{\beta}_j) W(\boldsymbol{\beta}_{j+1})\,\mathrm{CNOT}_{j,j+1}$.

Proof. The corollary easily descends from Proposition 4, Theorem 2 and Proposition 1. □

As shown above, for $t \in \mathcal{O}(\sqrt{\log(n)})$, the states generated by the time evolution of GDE Hamiltonians obey to area law of the entanglement with overwhelming probability. Thanks to Theorem 2, we also have that the loss function $L(\boldsymbol{\theta}, t)$ anti-concentrates, giving strong evidence of the success of the Hamiltonian Discrimination QML task.

8 Numerical simulations

In this section, we present numerical results which further explore the connection between the entanglement in the input state, and the phenomenon of gradient concentration. In particular, we are interested in showing how the parameter landscape of a QML problem becomes more and more concentrated as the entanglement in the input state grows.

To create n-qubit states with different amounts of entanglement, we will consider time-evolved states of the form

$$|\psi_t\rangle := \exp(-iHt)|\psi_0\rangle\,, \tag{32}$$

where $|\psi_0\rangle$ is a random product state, and where H is the Heisenberg model with first-neighbor interactions

$$H = \sum_{i=1}^{n} X_i X_{i+1} + Y_i Y_{i+1} + 2Z_i Z_{i+1} + X_i\,, \tag{33}$$

with periodic boundary conditions ($n+1 \equiv 1$). Here, σ_i with $\sigma = X, Y, Z$, denotes a Pauli operator acting on qubit i. As we will see below, as t increases, so does the entanglement in $|\psi_t\rangle$.

Next, we will consider a learning task where we want to minimize a cost-function of the form

$$L(\boldsymbol{\theta}) = 1 - \mathrm{Tr}[U(\boldsymbol{\theta})\,|\psi_t\rangle\langle\psi_t|\,U^\dagger(\boldsymbol{\theta})O_Z]\,, \qquad (34)$$

where $O_Z = \sum_i Z_i$ (i.e., O_Z is a sum of 1-local operators), and where $U(\boldsymbol{\theta})$ is a shallow HEA. Specifically, we employ the HEA architecture shown in Fig. 6 which is composed of an initial layer of general single qubit rotations, followed by two-qubit gates on alternating pairs of qubits. The two-qubit gates are themselves composed of a CNOT gate followed by general single-qubit gates on each qubit.

In Figs. 7(a,b) we show averaged norm of the gradient $\partial_\mu L(\boldsymbol{\theta})$, i.e. $\|\nabla L\|_\infty \equiv \max_\mu |\partial_\mu L(\boldsymbol{\theta})|$, as a function of the evolution time t used to prepare the input state of the HEA for different problem sizes. Gradients are computed by averaging over 400 random product states $|\psi_0\rangle$, and two sets of random parameters in the HEA for each initial state. Here we can see that for small evolution times the cost exhibits large gradients independently of the system size. This result is expected as we recall that in the limit $t \to 0$ the input state $|\psi_0\rangle$ is a tensor product state, which, along with 1-local measurements and the HEA structure, leads to the gradients which norms are independent of n. As t increases, we can see that the gradient norm decrease until a saturation value G_{sat} is achieved. Moreover, we can see that the value of G_{sat} depends on the number of qubits in the system. In fact, as shown in Fig. 7(c), G_{sat} decays polynomially with n. We can further understand this behavior by noting that as t increases, the time-evolution $\exp(-iHt)$ produces larger amounts of entanglement in the input state, and concomitantly smaller gradients (as indicated by our main results above). To see that this is the case, we compute rescaled entropy $S(\rho_2) = -\mathrm{Tr}[\rho_2 \log(\rho_2)]/2$ where ρ_2 is the reduced state on two nearest-neighbor qubits for a sufficiently large time t such that G_{sat} is achieved. Results are shown in Fig. 7(c). It shows a positive correlation between the decay of gradients and the increase in reduced state entropy. Thus, the more entanglement in the input state, the smaller the gradients, and the more concentrated the landscape.

9 Discussion and conclusions

Understanding the capabilities and limitations of VQA and QML algorithms is crucial to developing strategies that can be used to achieve a quantum advantage. One of the most relevant ingredients in ensuring the success of a VQA/QML model is the choice of ansatzes for the parametrized quantum circuit. In this work, we focused our attention on the shallow HEA, as it can avoid barren plateaus, and since it is perhaps one of the most NISQ-friendly ansatzes. Currently, the HEA is widely used for a plethora of problems, irrespective of whether it is well-fit for the task and data at hand. In a sense, the HEA is still a "*solution in search of a problem*" as there was no rigorous study of the tasks where it should, or should not be used. In this work, we establish rigorous results, showing how, and in which contexts, HEAs are (and are not) useful and can eventually provide a signature of quantum advantage.

We first review relevant results from the literature, discussing the notion of cost and loss function concentration and necessary conditions for trainability of HEAs – i.e. shallowness and locality of measurements. Here we highlight the existence of a new source of untrainability of shallow HEAs: the entanglement of the input states. On one hand, we proved that HEAs are untrainable if the input states satisfy a volume law of entanglement, as the cost function is deterministically concentrated around its trivial value. On the other hand, if the input states follow an area law of entanglement, the HEA is trainable. In fact, here we prove that the loss function anti-concentrates, i.e., it differs, at least polynomially, at sufficiently different points of the parameters landscape.

While the role of entanglement in the trainability of VQA and QML models has been explored in Refs. [50, 51], the results found therein are conceptually different from ours. Namely, in these references the authors point out that deep parametrized quantum circuit ansatzes create volume law for the entanglement entropy, making the parameter landscape exponentially flat in the number of qubits and thus giving rise to entanglement-induced barren plateaus. As such, these results study the entanglement created *during* the circuit, but not that *already present* in the input states. For instance, the shallow HEA cannot create volume law of entanglement, yet, it is still untrainable if such entanglement exists in the input state. Hence, our work provides a new source of untrainability for certain datasets.

Next, we also analyzed the still open question of whether the HEA is able to achieve a quan-

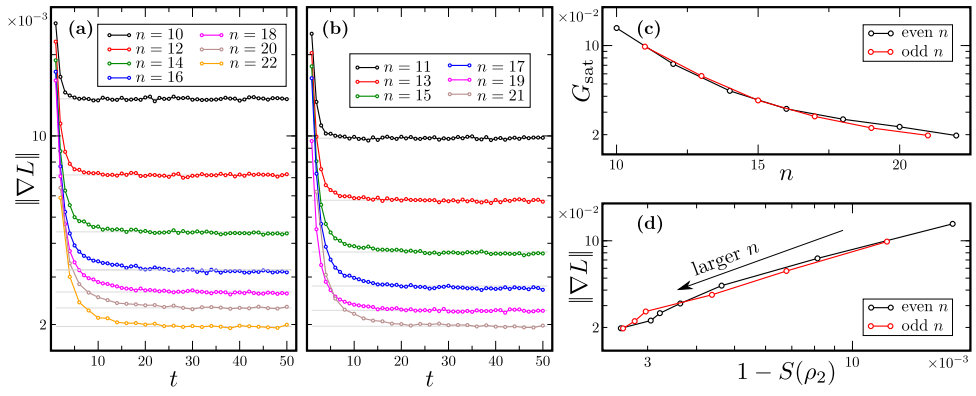

Figure 7: Numerical results. We consider a problem where the input states $|\psi_t\rangle$ of the HEA are determined by Eqs. (32) and (33) and where the loss function is given by (34). In panels (a) and (b) we respectively show (averaged over $|\psi_t\rangle$ and $\boldsymbol{\theta}$) norm of the gradient $\partial_\mu L(\boldsymbol{\theta})$ as a function of the evolution time t for different system sizes with n even and odd. Panel (c) shows norm saturation value G_{sat} of the results presented in (a) and (b) versus system size n. Panel (d) shows the norm of the gradient versus $1 - S(\rho_2)$, where $S(\rho_2)$ is the entropy of two-qubit subsystem for an evolution time such that the saturation value G_{sat} is achieved. Different points correspond to different values of n. The results are averaged over 400 initial states $|\psi_0\rangle$ in Eq. (32) and two sets of angles $\boldsymbol{\theta}$ for every initial state.

tum advantage in a VQA/QML setting. While the answer is far beyond the scope of the paper, we identified regimes in which the HEA can or cannot provide quantum speed-ups. Here we proved that thanks to the shallowness of HEAs, input states with bond dimensions at most polynomially in the number of qubits can be simulated with only a polynomial overhead on classical machines. This result rules out the use of HEA in VQAs: as many examples show, the typical input state for a VQA is an easy-to-prepare product state, thus allowing an efficient classical decomposition. Conversely, for QML algorithms the question still remains open: the portion of area law states admitting an efficient classical description is exponentially small [76]. While this is not a guarantee for achieving quantum advantage, this is definitely the window to look at for applications beyond those solvable by classical capabilities.

We indeed push forward the latter intuition and provide an example to which our results apply. Namely, we present a Hamiltonian discrimination QML problem, where initial product states are evolved in time by two types of Hamiltonians, one possessing a given local symmetry, and one completely general. We show that, while the task becomes less and less feasible if the evolution time is long (as entanglement growing in time), for a given time window (scaling logarithmically with the number of qubits) such states possess area law of entanglement, ensuring the absence of barren plateaus in the loss landscape.

Such an example serves as a pivotal one for future, and hopefully fruitful, usages of the HEA. Our recipe to prepare a Barren-plateau-free QML problem is the following: consider entangled enough quantum input data, pass it through a shallow-depth circuit, and then measure with local operators. Importantly, if one wishes to use this scheme for classical data, it will be extremely important to find data-embedding schemes that lead to area-law states, but which are not themselves classically simulable. We expect that the search for such entangled-tamed embedding could be a fruitful research direction.

There are still several directions to be explored after the analysis of the present work. We indeed emphasize that, while for unstructured HEAs, this paper definitely rules out volume law states as input states of QML algorithms with HEA ansatzes, there is the fascinating possibility that, with even a little prior knowledge of the input states, a problem-aware HEA could avoid exponential concentration in the parameter landscapes. Indeed, the choice of some structured, and problem-dependent ansatz can avoid barren plateaus: a prominent example is the Geometric Quantum Machine Learning, which exploits the geometric symmetries of the input data-set to design symmetry-aware parametrized ansatzes [35–41].

Acknowledgements

This work was supported by the U.S.Department of Energy (DOE) through a quantum computing program sponsored by the Los Alamos National Laboratory Information Science & Technology Institute. L.L and S.F.E.O. acknowledge support from NSF award no. 2014000 and by the Cen-

ter for Nonlinear Studies at Los Alamos National Laboratory (LANL). L.C. was partially supported by the U.S. DOE, Office of Science, Office of Advanced Scientific Computing Research, under the Accelerated Research in Quantum Computing (ARQC) program. M.C. was initially supported by the Laboratory Directed Research and Development (LDRD) program of LANL under project number 20230049DR. The authors also acknowledge support by the U.S. DOE through a quantum computing program sponsored by the LANL Information Science & Technology Institute.

References

[1] John Preskill. "Quantum computing in the nisq era and beyond". Quantum **2**, 79 (2018).

[2] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, et al. "Quantum supremacy using a programmable superconducting processor". Nature **574**, 505–510 (2019).

[3] Yulin Wu, Wan-Su Bao, Sirui Cao, Fusheng Chen, et al. "Strong Quantum Computational Advantage Using a Superconducting Quantum Processor". Physical Review Letters **127**, 180501 (2021).

[4] Lars S Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob FF Bulmer, Filippo M Miatto, Leonhard Neuhaus, Lukas G Helt, Matthew J Collins, et al. "Quantum computational advantage with a programmable photonic processor". Nature **606**, 75–81 (2022).

[5] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. "Variational quantum algorithms". Nature Reviews Physics **3**, 625–644 (2021).

[6] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. "Noisy intermediate-scale quantum algorithms". Reviews of Modern Physics **94**, 015004 (2022).

[7] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. "A variational eigenvalue solver on a photonic quantum processor". Nature Communications **5**, 1–7 (2014).

[8] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Sergio Boixo, Michael Broughton, Bob B Buckley, David A Buell, et al. "Hartree-fock on a superconducting qubit quantum computer". Science **369**, 1084–1089 (2020).

[9] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm" (2014). url: https://arxiv.org/abs/1411.4028.

[10] Matthew P. Harrigan, Kevin J. Sung, Matthew Neeley, Kevin J. Satzinger, et al. "Quantum approximate optimization of non-planar graph problems on a planar superconducting processor". Nature Physics **17**, 332–336 (2021).

[11] Carlos Bravo-Prieto, Ryan LaRose, Marco Cerezo, Yigit Subasi, Lukasz Cincio, and Patrick J Coles. "Variational quantum linear solver". Quantum **7**, 1188 (2023).

[12] Hsin-Yuan Huang, Kishor Bharti, and Patrick Rebentrost. "Near-term quantum algorithms for linear systems of equations with regression loss functions". New Journal of Physics **23**, 113021 (2021).

[13] Xiaosi Xu, Jinzhao Sun, Suguru Endo, Ying Li, et al. "Variational algorithms for linear algebra". Science Bulletin **66**, 2181–2188 (2021).

[14] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. "Quantum machine learning". Nature **549**, 195–202 (2017).

[15] Maria Schuld and Francesco Petruccione. "Machine Learning with Quantum Computers". Springer International Publishing. Cham, Switzerland (2021).

[16] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. "Supervised learning with quantum-enhanced feature spaces". Nature **567**, 209–212 (2019).

[17] Louis Schatzki, Andrew Arrasmith, Patrick J. Coles, and M. Cerezo. "Entangled datasets for quantum machine learning" (2021). url: https://arxiv.org/abs/2109.03400.

[18] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, et al. "Unsupervised machine learning on a hybrid quantum computer" (2017). url: https://arxiv.org/abs/1712.05771.

[19] Sofiene Jerbi, Casper Gyurik, Simon Marshall, Hans Briegel, et al. "Parametrized Quantum Policies for Reinforcement Learning". Advances in Neural Information Processing Systems **34**, 28362–

28375 (2021). url: https://proceedings.neurips.cc/paper/2021/hash/eec96a7f788e88184c0e713456026f3f-Abstract.html.

[20] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. "Data re-uploading for a universal quantum classifier". Quantum **4**, 226 (2020).

[21] Iris Cong, Soonwon Choi, and Mikhail D Lukin. "Quantum convolutional neural networks". Nature Physics **15**, 1273–1278 (2019).

[22] Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. "Generalization in quantum machine learning from few training data". Nature Communications**13** (2022).

[23] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. "Provably efficient machine learning for quantum many-body problems". Science **377**, eabk3333 (2022).

[24] M Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. "Challenges and opportunities in quantum machine learning". Nature Computational Science (2022).

[25] Linghua Zhu, Ho Lun Tang, George S Barron, Nicholas J Mayhall, Edwin Barnes, and Sophia E Economou. "An adaptive quantum approximate optimization algorithm for solving combinatorial problems on a quantum computer" (2020). url: https://arxiv.org/abs/2005.10258.

[26] Ho Lun Tang, VO Shkolnikov, George S Barron, Harper R Grimsley, Nicholas J Mayhall, Edwin Barnes, and Sophia E Economou. "qubit-adapt-vqe: An adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor". PRX Quantum **2**, 020310 (2021).

[27] Zi-Jian Zhang, Thi Ha Kyaw, Jakob S. Kottmann, Matthias Degroote, and Alán Aspuru-Guzik. "Mutual information-assisted adaptive variational quantum eigensolver". Quantum Science and Technology **6**, 035001 (2021).

[28] Matias Bilkis, Marco Cerezo, Guillaume Verdon, Patrick J Coles, and Lukasz Cincio. "A semi-agnostic ansatz with variable structure for variational quantum algorithms". Quantum Machine Intelligence **5**, 43 (2023).

[29] Arthur G Rattew, Shaohan Hu, Marco Pistoia, Richard Chen, and Steve Wood. "A domain-agnostic, noise-resistant, hardware-efficient evolutionary variational quantum eigensolver" (2019). url: https://arxiv.org/abs/1910.09694.

[30] Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G Rieffel, Davide Venturelli, and Rupak Biswas. "From the quantum approximate optimization algorithm to a quantum alternating operator ansatz". Algorithms **12**, 34 (2019).

[31] Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. "Exploring entanglement and optimization within the hamiltonian variational ansatz". PRX Quantum **1**, 020319 (2020).

[32] Juneseo Lee, Alicia B Magann, Herschel A Rabitz, and Christian Arenz. "Progress toward favorable landscapes in quantum combinatorial optimization". Physical Review A **104**, 032401 (2021).

[33] Guillaume Verdon, Trevor McCourt, Enxhell Luzhnica, Vikash Singh, Stefan Leichenauer, and Jack Hidary. "Quantum graph neural networks" (2019). url: https://arxiv.org/abs/1909.12264.

[34] Johannes Bausch. "Recurrent quantum neural networks". In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems. Volume 33, pages 1368–1379. Curran Associates, Inc. (2020). url: https://proceedings.neurips.cc/paper/2020/file/0ec96be397dd6d3cf2fecb4a2d627c1c-Paper.pdf.

[35] Martín Larocca, Frédéric Sauvage, Faris M. Sbahi, Guillaume Verdon, Patrick J. Coles, and M. Cerezo. "Group-invariant quantum machine learning". PRX Quantum **3**, 030341 (2022).

[36] Johannes Jakob Meyer, Marian Mularski, Elies Gil-Fuster, Antonio Anna Mele, Francesco Arzani, Alissa Wilms, and Jens Eisert. "Exploiting symmetry in variational quantum machine learning". PRX Quantum **4**, 010328 (2023).

[37] Andrea Skolik, Michele Cattelan, Sheir Yarkoni, Thomas Bäck, and Vedran Dunjko. "Equivariant quantum circuits for learning on weighted graphs" (2022). url: https://arxiv.org/abs/2205.06109.

[38] Frederic Sauvage, Martin Larocca, Patrick J Coles, and Marco Cerezo. "Building spatial symmetries into parameterized quantum circuits for faster training". Quantum Science and Technology **9**, 015029 (2024).

[39] Michael Ragone, Quynh T. Nguyen, Louis Schatzki, Paolo Braccia, Martin Larocca, Frederic Sauvage, Patrick J. Coles, and M. Cerezo. "Representation theory for geometric quantum machine learning" (2022). url: https://arxiv.org/abs/2210.07980.

[40] Quynh T. Nguyen, Louis Schatzki, Paolo Braccia, Michael Ragone, Patrick J. Coles, Frédéric Sauvage, Martín Larocca, and M. Cerezo. "Theory for equivariant quantum neural networks". PRX Quantum **5**, 020328 (2024).

[41] Louis Schatzki, Martín Larocca, Quynh T. Nguyen, Frédéric Sauvage, and M. Cerezo. "Theoretical guarantees for permutation-equivariant quantum neural networks". npj Quantum Information**10** (2024).

[42] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Barren plateaus in quantum neural network training landscapes". Nature Communications **9**, 1–6 (2018).

[43] M Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. "Cost function dependent barren plateaus in shallow parametrized quantum circuits". Nature Communications **12**, 1–12 (2021).

[44] Kunal Sharma, Marco Cerezo, Lukasz Cincio, and Patrick J Coles. "Trainability of dissipative perceptron-based quantum neural networks". Physical Review Letters **128**, 180505 (2022).

[45] Supanut Thanasilp, Samson Wang, Nhat A. Nghiem, Patrick J. Coles, and M. Cerezo. "Subtleties in the trainability of quantum machine learning models". Quantum Machine Intelligence **5**, 21 (2023).

[46] Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J Coles. "Connecting ansatz expressibility to gradient magnitudes and barren plateaus". PRX Quantum **3**, 010313 (2022).

[47] Andrew Arrasmith, Zoë Holmes, Marco Cerezo, and Patrick J Coles. "Equivalence of quantum barren plateaus to cost concentration and narrow gorges". Quantum Science and Technology **7**, 045015 (2022).

[48] Arthur Pesah, M. Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. "Absence of barren plateaus in quantum convolutional neural networks". Physical Review X **11**, 041011 (2021).

[49] AV Uvarov and Jacob D Biamonte. "On barren plateaus and cost function locality in variational quantum algorithms". Journal of Physics A: Mathematical and Theoretical **54**, 245301 (2021).

[50] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. "Entanglement-induced barren plateaus". PRX Quantum **2**, 040316 (2021).

[51] Taylor L Patti, Khadijeh Najafi, Xun Gao, and Susanne F Yelin. "Entanglement devised barren plateau mitigation". Physical Review Research **3**, 033090 (2021).

[52] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles. "Quantum-assisted quantum compiling". Quantum **3**, 140 (2019).

[53] David H Wolpert and William G Macready. "No free lunch theorems for optimization". IEEE transactions on evolutionary computation **1**, 67–82 (1997).

[54] Jonathan Romero, Ryan Babbush, Jarrod R McClean, Cornelius Hempel, Peter J Love, and Alán Aspuru-Guzik. "Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz". Quantum Science and Technology **4**, 014008 (2018).

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

[56] "IBM Q 16 Rueschlikon backend specification" (2018).

[57] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. "Noise-induced barren plateaus in variational quantum algorithms". Nature Communications **12**, 1–11 (2021).

[58] Daniel Stilck França and Raul Garcia-Patron. "Limitations of optimization algorithms on noisy quantum devices". Nature Physics **17**, 1221–1227 (2021).

[59] Fernando GSL Brandao, Aram W Harrow, and Michał Horodecki. "Local random quantum circuits are approximate polynomial-designs". Communications in Mathematical Physics **346**, 397–434 (2016).

[60] Aram W Harrow and Saeed Mehraban. "Approximate unitary t-designs by short random quantum circuits using nearest-neighbor and long-range gates". Communications in Mathematical Physics **401**, 1531–1626 (2023).

[61] Valerie Coffman, Joydip Kundu, and William K Wootters. "Distributed entanglement". Physical Review A **61**, 052306 (2000).

[62] Dmitry A. Abanin and Eugene Demler. "Measuring entanglement entropy of a generic

many-body system with a quantum switch". Physical Review Letters **109**, 020504 (2012).

[63] Steph Foulds, Viv Kendon, and Tim Spiller. "The controlled SWAP test for determining quantum entanglement". Quantum Science and Technology **6**, 035002 (2021).

[64] Jacob L. Beckey, N. Gigena, Patrick J. Coles, and M. Cerezo. "Computable and operationally meaningful multipartite entanglement measures". Phys. Rev. Lett. **127**, 140501 (2021).

[65] Lorenzo Leone, Salvatore F. E. Oliviero, and Alioscia Hamma. "Stabilizer Rényi entropy". Physical Review Letters **128**, 050402 (2022).

[66] Salvatore F. E. Oliviero, Lorenzo Leone, Alioscia Hamma, and Seth Lloyd. "Measuring magic on a quantum processor". npj Quantum Inf **8**, 1–8 (2022).

[67] Tobias Haug and MS Kim. "Scalable measures of magic resource for quantum computers". PRX Quantum **4**, 010301 (2023).

[68] Min-Sung Kang, Jino Heo, Seong-Gon Choi, Sung Moon, and Sang-Wook Han. "Implementation of swap test for two unknown states in photons via cross-kerr nonlinearities under decoherence effect". Scientific reports **9**, 1–14 (2019).

[69] Román Orús. "A practical introduction to tensor networks: Matrix product states and projected entangled pair states". Annals of Physics **349**, 117–158 (2014).

[70] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". Annals of Physics **326**, 96–192 (2011).

[71] Frank Verstraete, Valentin Murg, and J Ignacio Cirac. "Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems". Advances in physics **57**, 143–224 (2008).

[72] Norbert Schuch, Michael M Wolf, Frank Verstraete, and J Ignacio Cirac. "Entropy scaling and simulability by matrix product states". Physical review letters **100**, 030504 (2008).

[73] Frank Verstraete and J Ignacio Cirac. "Renormalization algorithms for quantum-many body systems in two and higher dimensions" (2004). url: https://arxiv.org/abs/cond-mat/0407066.

[74] Jakob S. Kottmann and Alán Aspuru-Guzik. "Optimized low-depth quantum circuits for molecular electronic structure using a separable-pair approximation". Physical Review A **105**, 032449 (2022).

[75] Román Orús. "A practical introduction to tensor networks: Matrix product states and projected entangled pair states". Annals of Physics **349**, 117–158 (2014).

[76] Yimin Ge and Jens Eisert. "Area laws and efficient descriptions of quantum many-body states". New Journal of Physics **18**, 083026 (2016).

[77] Salvatore F. E. Oliviero, Lorenzo Leone, Francesco Caravelli, and Alioscia Hamma. "Random Matrix Theory of the Isospectral twirling". SciPost Physics **10**, 76 (2021).

[78] Lorenzo Leone, Salvatore F. E. Oliviero, and Alioscia Hamma. "Isospectral twirling and quantum chaos". Entropy**23** (2021).

[79] Sandu Popescu, Anthony J. Short, and Andreas Winter. "Entanglement and the foundations of statistical mechanics". Nature Physics **2**, 754–758 (2006).

[80] Aram W Harrow. "The church of the symmetric subspace" (2013). url: https://arxiv.org/abs/1308.6595.

[81] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. "Quantum circuit learning". Physical Review A **98**, 032309 (2018).

[82] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. "Evaluating analytic gradients on quantum hardware". Physical Review A **99**, 032331 (2019).

[83] Don Weingarten. "Asymptotic behavior of group integrals in the limit of infinite rank". Journal of Mathematical Physics **19**, 999–1001 (1978). arXiv:https://doi.org/10.1063/1.523807.

[84] Benoît Collins. "Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability". International Mathematics Research Notices **2003**, 953–982 (2003).

[85] Benoît Collins and Piotr Śniady. "Integration with respect to the haar measure on unitary, orthogonal and symplectic group". Communications in Mathematical Physics **264**, 773–795 (2006).

[86] Patrick J Coles, M Cerezo, and Lukasz Cincio. "Strong bound between trace distance and hilbert-schmidt distance for low-rank states". Physical Review A **100**, 022103 (2019).

[87] Pavan Hosur, Xiao-Liang Qi, Daniel A. Roberts, and Beni Yoshida. "Chaos in quantum channels". Journal of High Energy Physics **2016**, 4 (2016).

[88] Lorenzo Leone, Salvatore F. E. Oliviero, You Zhou, and Alioscia Hamma. "Quantum Chaos is Quantum". Quantum **5**, 453 (2021).

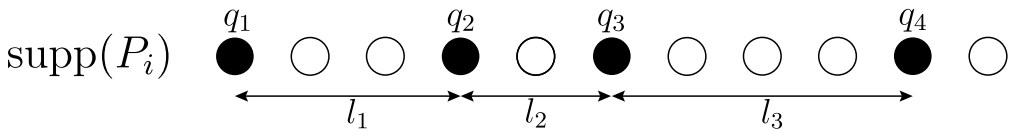

Figure 8: The support $\mathrm{supp}(P_i)$ contains $q_1, \ldots, q_4$ qubits, whose relative distances are $l_1, \ldots, _3$ qubits, used to define clusters when compared to $2D$.

[89] Salvatore F.E. Oliviero, Lorenzo Leone, and Alioscia Hamma. "Transitions in entanglement complexity in random quantum circuits by measurements". Physics Letters A **418**, 127721 (2021).

[90] Dawei Ding, Patrick Hayden, and Michael Walter. "Conditional mutual information of bipartite unitaries and scrambling". Journal of High Energy Physics **2016**, 145 (2016).

[91] Zi-Wen Liu, Seth Lloyd, Elton Zhu, and Huangjun Zhu. "Entanglement, quantum randomness, and complexity beyond scrambling". Journal of High Energy Physics **2018**, 41 (2018).

[92] Jordan Cotler, Nicholas Hunter-Jones, Junyu Liu, and Beni Yoshida. "Chaos, complexity, and random matrices". Journal of High Energy Physics **2017**, 48 (2017).

[93] Zbigniew Puchala and Jaroslaw Adam Miszczak. "Symbolic integration with respect to the haar measure on the unitary groups". Bulletin of the Polish Academy of Sciences Technical Sciences **65**, 21–27 (2017). url: http://journals.pan.pl/dlibra/publication/121307/edition/105697/content.

A Proof of Theorem 1 and Corollaries 1, 2, and 3

A.1 Proof of Theorem 1

Let us consider the following distance $|f(\boldsymbol{\theta}) - f_{trv}|$, where $f(\boldsymbol{\theta}) = C(\boldsymbol{\theta})$ and $f_{trv} = 2^{-n}\,\mathrm{Tr}[O]$. Our task is to upper-bound this distance with the information-theoretic measure of information scrambling $\mathcal{I}_\Lambda(\psi)$. Let us decompose the measurement operator O on the Pauli basis:

$$O = \sum_{i=1}^{N} c_i P_i\,, \tag{35}$$

where $c_i \coloneqq \mathrm{Tr}[OP_i]/2^n$, and $\mathrm{Tr}[P_iP_j] = 2^n\delta_{ij}$. Then, we define the support $\mathrm{supp}(P_i)$ as the ordered set of qubits containing non-identity operators in the tensor product structure of each Pauli operator $P_i = \bigotimes_{j=1}^{n}\sigma_j^{(i)}$, where $\sigma_j^{(i)}$ is a single qubit Pauli operator acting on the j-th qubit. Explicitly, $\mathrm{supp}(P_i) = \{q_1^{(i)}, \ldots, q_{S_i}^{(i)} \,|\, q_1^{(i)} < \ldots < q_{S_i}^{(i)}, q_k^{(i)} \in [1,n]\}$ is an ordered subset of natural numbers labeling the qubits on which P_i acts non-trivially, and S_i labels the number of qubits on which P_i acts, see Fig. 8. From this, we also define the support of O as the ordered subset of qubits given by the union of the supports of each P_i, i.e., $\mathrm{supp}(O) = \bigcup_i \mathrm{supp}(P_i)$. Moreover, we note that $\sum_i |c_i| \leq N \max_i |c_i| \leq N\,\|O\|_\infty$ (which the latter follows from Holder's inequality: $|c_i| = 2^{-n}|\,\mathrm{Tr}(OP_i)| \leq 2^{-n}\|O\|_\infty\|P_i\|_1 = \|O\|_\infty$), and $N \leq (3n)^{\max_i |\,\mathrm{supp}(P_i)|}$ (which follows from a counting argument). Let us define the pairwise relative distance between the qubits q_k and q_{k+1} belonging to $\mathrm{supp}(P_i)$ as:

$$l_k^{(i)} \coloneqq |q_{k+1}^{(i)} - q_k^{(i)}|, \quad k = 1, \ldots, S_i - 1\,. \tag{36}$$

It is now possible to define clusters, i.e., subsets of contiguous qubits whose pairwise relative distance is less than $2D$. The definition can be done recursively: let $C_1^{(i)}$ be the first cluster, then $q_1^{(i)} \in C_1^{(i)}$; if $l_1^{(i)} \leq 2D$ then $q_2^{(i)} \in C_1^{(i)}$, otherwise $q_2^{(i)} \in C_2^{(i)}$. This procedure defines $C_1^{(i)}, \ldots, C_{L_i}^{(i)}$ clusters of qubits for any $i = 1, \ldots, N$, such that $1 \leq L_i \leq S_i$. Note that, we can write the operator O as $O = \sum_i c_i \bigotimes_{\alpha=1}^{L_i} O_\alpha^{(i)}$, with $\mathrm{supp}(O_\alpha^{(i)}) = C_\alpha^{(i)}$. Consequently, we can rewrite the cost function by expanding $O = \sum_i c_i P_i = \sum_i c_i (\bigotimes_\alpha O_\alpha^{(i)})$:

$$f(\boldsymbol{\theta}) = \sum_i c_i\,\mathrm{Tr}[|\psi\rangle\langle\psi|\,U^\dagger(\boldsymbol{\theta})(\bigotimes_\alpha O_\alpha^{(i)})U(\boldsymbol{\theta})]\,. \tag{37}$$

It is possible to evaluate the distance between $f(\boldsymbol{\theta})$ and f_{trv}, that can be rewritten as:

$$f_{trv} = 2^{-n} \sum_i c_i \operatorname{Tr}[U^\dagger(\boldsymbol{\theta})(\bigotimes_\alpha O_\alpha^{(i)})U(\boldsymbol{\theta})]\,. \tag{38}$$

Defining $\Lambda_i \coloneqq \operatorname{supp}(U^\dagger(\boldsymbol{\theta})(\bigotimes_\alpha O_\alpha^{(i)})U(\boldsymbol{\theta}))$, the distance between $f(\boldsymbol{\theta})$ and f_{trv} reads:

$$\begin{aligned} |f(\boldsymbol{\theta}) - f_{trv}| &\leq \sum_i |c_i| |\operatorname{Tr}_{\Lambda_i}[\psi_{\Lambda_i} U^\dagger(\boldsymbol{\theta})(\bigotimes_\alpha O_\alpha^{(i)})U(\boldsymbol{\theta})] - 2^{-|\Lambda_i|}\operatorname{Tr}_{\Lambda_i}[U^\dagger(\boldsymbol{\theta})(\bigotimes_\alpha O_\alpha^{(i)})U(\boldsymbol{\theta})]| \\ &\leq \sum_i |c_i| \left\|O_\alpha^{(i)}\right\|_\infty \left\|\psi_{\Lambda_i} - \frac{\mathbb{1}_{\Lambda_i}}{2^{|\Lambda_i|}}\right\|_1 \equiv \sum_i |c_i| \mathcal{I}_{\Lambda_i}(\psi)\,, \end{aligned} \tag{39}$$

where $\psi_\Lambda \equiv \operatorname{Tr}_{\bar{\Lambda}} |\psi\rangle\langle\psi|$. In the first inequality, we use the triangle inequality of absolute value, while in the second one, we used $|\operatorname{Tr}[AB]| \leq \|A\|_\infty \|B\|_1$, and the fact that unitary operators preserve the trace and any Schatten p-norm. Finally, we used $\|\mathcal{O}_\alpha^{(i)}\|_\infty = 1$. To complete the proof it is still necessary to bound $|\Lambda| \equiv \max_i |\Lambda_i|$.

In what follows we will use the following lemma and corollary, which we will prove at the end of this Appendix (see App. D).

Lemma 1. *Let $U(\boldsymbol{\theta})$ be a HEA with depth D as in Fig. 1(b), and $O = \sum_i c_i P_i = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ be an operator, as in Eq. (35). Then for any $O_\alpha^{(i)}$:*

$$|\operatorname{supp}(U(\boldsymbol{\theta}) O_\alpha^{(i)} U^\dagger(\boldsymbol{\theta}))| \leq \sum_{\substack{k \in C_\alpha^{(i)} \\ l_k^{(i)} < 2D}} l_k^{(i)} + 2D\,. \tag{40}$$

The following corollary descends from the above lemma and the clustering of qubits.

Corollary 7. *Let $U(\boldsymbol{\theta})$ be HEA with depth D, with $D \in \mathcal{O}(\log(n))$ and $O = \sum_i c_i P_i = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$. Then for any $O_\alpha^{(i)}, O_{\alpha'}^{(i)}$, with $\alpha \neq \alpha'$, one finds that asymptotically (in n),*

$$\operatorname{supp}(U(\boldsymbol{\theta}) O_\alpha^{(i)} U^\dagger(\boldsymbol{\theta})) \cap \operatorname{supp}(U(\boldsymbol{\theta}) O_{\alpha'}^{(i)} U^\dagger(\boldsymbol{\theta})) = \emptyset\,. \tag{41}$$

Thanks to Corollary 7, we know that $\operatorname{supp}(U^\dagger(\boldsymbol{\theta}) O_\alpha U(\boldsymbol{\theta})) \cap \operatorname{supp}(U^\dagger(\boldsymbol{\theta}) O_{\alpha'} U(\boldsymbol{\theta})) = \emptyset$, and thus:

$$|\Lambda| = \max_i \sum_\alpha |\operatorname{supp}(U^\dagger(\boldsymbol{\theta}) O_\alpha^{(i)} U(\boldsymbol{\theta}))| \tag{42}$$

then, from Lemma 1:

$$|\Lambda| \leq \max_i \sum_\alpha \sum_{\substack{k \in C_\alpha^{(i)} \\ l_k < mD}} l_k^{(i)} + mD \leq \max_i \sum_\alpha (|C_\alpha^{(i)}| - 1) mD + mD\,, \tag{43}$$

where the second inequality follows by assuming maximum distance between qubits in the cluster. Clearly $\sum_\alpha |C_\alpha^{(i)}| = |\operatorname{supp}(P_i)|$ by definition, and $\sum_\alpha 1 = L_i \geq 1$, thus $mD(1 - \sum_\alpha 1) \leq 0$, and we obtain the following

$$|\Lambda| \leq mD \max_i |\operatorname{supp}(P_i)|\,. \tag{44}$$

A.2 Proof of Corollary 1

By theorem 1, we have

$$|f(\boldsymbol{\theta}) - f_{trv}| \leq \|O\|_\infty (3n)^{\max_i \operatorname{supp}(P_i)} \max_i \mathcal{I}_{\Lambda_i}(\psi)\,, \tag{45}$$

if the information is scrambled, according to Definition 6, we have that, since $\max_i |\Lambda_i| \in \mathcal{O}(\log(n))$, then $\max_i \mathcal{I}_{\Lambda_i}(\psi) \in \mathcal{O}(2^{-cn})$ for some $c > 0$. Then, if $\max_i |\operatorname{supp}(P_i)| \in \mathcal{O}(\log(n))$, there exists a constant c' such that $n^{\max_i \operatorname{supp}(P_i)} \in \mathcal{O}(n^{c' \log(n)})$, and therefore:

$$(3n)^{\max_i \operatorname{supp}(P_i)} \in \mathcal{O}(n^{2c' \log(n)})\,, \tag{46}$$

where we crudely upperbounded $3n < n^2$ (for $n > 3$). Now, we have that $n^{2c' \log(n)} = 2^{2c' \log^2 n} \in \mathcal{O}(2^{kn})$ for any $k \in \Omega(1)$. Choosing $k = c/2$ for simplicity, we finally reach the final bound:

$$|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(2^{-cn/2}) \,. \tag{47}$$

where we considered $\|O\|_\infty \in \Omega(1)$.

A.3 Proof of Corollary 2

The proof of Corollary 2 follows from the proof in the work of Popescu et al. [79]. Here we report it for completeness.

Let $\psi_\Lambda = \mathrm{Tr}_{\bar{\Lambda}}[|\psi\rangle\langle\psi|]$, consider a complete set of (Hermitian) observables $O_\Lambda^{(i)}$ on Λ with $\left\|O_\Lambda^{(i)}\right\|_\infty = 1$, and decompose the state on the complete set of observables $O_\Lambda^{(i)}$:

$$\psi_\Lambda = \frac{1}{d_\Lambda} \sum_i \mathrm{Tr}[\psi_\Lambda O_\Lambda^{(i)}] O_\Lambda^{(i)} \,. \tag{48}$$

Consider the 1-norm distance between ψ_Λ and the completely mixed state $\mathbb{1}_\Lambda d_\Lambda^{-1}$:

$$\begin{aligned} \mathcal{I}_\Lambda(\psi) &\leq \sqrt{d_\Lambda} \left\| \psi_\Lambda - \frac{\mathbb{1}_\Lambda}{d_\Lambda} \right\|_2 \leq \sqrt{\sum_i \left(\mathrm{Tr}\left[\psi_\Lambda O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda} \right)^2} \\ &\leq d_\Lambda \max_i \left| \mathrm{Tr}[\psi_\Lambda O_\Lambda^{(i)}] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda} \right| . \end{aligned} \tag{49}$$

In the first inequality, we have used the norm equivalence, while in the second we have expanded ψ_Λ as in Eq. (48). The third inequality follows by simply taking the maximum over i in the summation.

Thus:

$$\Pr\left(\left\| \psi_\Lambda(\psi) - \mathbb{1}_\Lambda d_\Lambda^{-1} \right\|_1 \geq d_\Lambda \epsilon \right) \leq \Pr\left(\max_i \left| \mathrm{Tr}\left[\psi_\Lambda O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda} \right| \geq \epsilon \right) . \tag{50}$$

Since the expectation values are Lipschitz functions [79], by exploiting Levy's lemma one can easily prove that:

$$\Pr\left(\max_i \left| \mathrm{Tr}\left[\psi_\Lambda O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda} \right| \geq \epsilon \right) \leq 2 d_\Lambda^2 e^{-\frac{C 2^n \epsilon^2}{2}} \,,$$

where $C = (18\pi^3)^{-1}$. By choosing $\epsilon = 2^{-1/3n}$, one gets:

$$\Pr(\mathcal{I}_\Lambda < 2^{|\Lambda| - 1/3n}) \geq 1 - 2e^{-C 2^{n/3-1} + 2|\Lambda| \ln 2} \,. \tag{51}$$

To conclude, in virtue of Theorem 1, we have:

$$|f(\boldsymbol{\theta}) - f_{trv}| \leq \|O\|_\infty \, n^{\max_i \mathrm{supp}(P_i)} \max_i \mathcal{I}_{\Lambda_i}(\psi) \,. \tag{52}$$

Thanks to the above equation, we can bound $\mathcal{I}_{\Lambda_i} < 2^{|\Lambda_i|} 2^{-n/3}$ with overwhelming probability. Thus, $\max_i \mathcal{I}_{\Lambda_i} < 2^{|\Lambda|} 2^{-n/3}$, for Λ being such that $|\Lambda| = \max_i |\Lambda_i|$. Thus, we have $|f(\boldsymbol{\theta}) - f_{trv}| \leq \|O\|_\infty (3n)^{\max_i \mathrm{supp}(P_i)} 2^{\max_i \mathrm{supp}(P_i) + mD} 2^{-n/3}$. Now, given $\max_i \mathrm{supp}(P_i) \in \mathcal{O}(\log(n))$, and $D \in \mathcal{O}(\log(n))$, it means that there exists two constants c, c' such that

$$|f(\boldsymbol{\theta}) - f_{trv}| \leq \|O\|_\infty \, n^{c + c' \log(n)} 2^{-n/3} \,. \tag{53}$$

To conclude, we can loosely write that for any constant k we have $n^{c + c' \log(n)} \in \mathcal{O}(2^{n/k})$. Choosing $k = 6$ for simplicity, we finally reach the final bound:

$$|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(2^{-n/6}) \,, \tag{54}$$

holding with probability

$$\Pr(|f(\boldsymbol{\theta}) - f_{trv}| \in \mathcal{O}(2^{-n/6})) \geq 1 - e^{-C2^{n/3}+(2n+1)\ln 2}\,, \tag{55}$$

where we bound $|\Lambda| < n$.

A.4 Proof of Corollary 3

Now suppose one has as input k copies of a Haar random state $|\psi\rangle$ on n qubits, denote it as:

$$|\Psi^{(k)}\rangle = |\psi\rangle^{\otimes k}\,, \tag{56}$$

and denote $d = 2^n$ the dimension of the Hilbert space in which $|\psi\rangle$ is living. Let us decompose its reduced density matrix $\Psi_\Lambda^{(k)} \coloneqq \mathrm{Tr}_{\bar{\Lambda}} |\Psi^{(k)}\rangle\langle\Psi^{(k)}|$ in a Hermitian operator basis P_i:

$$\Psi_\Lambda^{(k)} = \frac{1}{d_\Lambda} \sum_i \mathrm{Tr}[\Psi_\Lambda^{(k)} P_i] P_i\,. \tag{57}$$

Let us prove that each expectation value on $|\Psi^{(k)}\rangle$ is a Lipschitz function with respect to $|\psi\rangle$. Given a second state $\tilde{\Psi}_\Lambda^{(k)}$, we have

$$\begin{aligned} |\,\mathrm{Tr}[\Psi_\Lambda^{(k)} P_i] - \mathrm{Tr}[\tilde{\Psi}_\Lambda^{(k)} P_i]| &\leq \left\| |\psi\rangle\langle\psi|^{\otimes k} - |\tilde{\psi}\rangle\langle\tilde{\psi}|^{\otimes k} \right\|_1 \\ &\leq k \left\| |\psi\rangle\langle\psi| - |\tilde{\psi}\rangle\langle\tilde{\psi}| \right\|_1 \\ &\leq 2k\sqrt{1 - |\,\langle\psi|\tilde{\psi}\rangle\,|^2} \\ &\leq 2k \left\| |\psi\rangle - |\tilde{\psi}\rangle \right\|\,, \end{aligned} \tag{58}$$

in the first inequality we used the fact that $\|P_i\|_\infty = 1$. In the second line, we used k times the following trick. Denote $\psi = |\psi\rangle\langle\psi|$:

$$\begin{aligned} \left\|\psi^{\otimes k} - \tilde{\psi}^{\otimes k}\right\|_1 &= \left\|\psi^{\otimes k} - \psi\otimes\tilde{\psi}^{\otimes k-1} + \psi\otimes\tilde{\psi}^{\otimes k-1} - \tilde{\psi}^{\otimes k}\right\|_1 \\ &\leq \left\|\psi^{\otimes k-1} - \tilde{\psi}^{\otimes k-1}\right\|_1 + \left\|\psi - \tilde{\psi}\right\|_1\,, \end{aligned} \tag{59}$$

and we used $\left\|\psi^{\otimes k}\right\|_1 = 1$ for any k. Thanks to the fact that $\mathrm{Tr}[\Psi_\Lambda^{(k)} P_i]$ is Lipschitz, and denoting as $\overline{\mathrm{Tr}\left[\Psi_\Lambda^{(k)} P_i\right]}$ the Haar average over the input state, we know that:

$$\Pr\left(\left|\mathrm{Tr}\left[\Psi_\Lambda^{(k)} P_i\right] - \overline{\mathrm{Tr}\left[\Psi_\Lambda^{(k)} P_i\right]}\right| \geq \epsilon\right) \leq e^{-C2^n\epsilon^2/2k^2}\,, \tag{60}$$

where $C = (18\pi^3)^{-1}$. We need to bound the probability that $\mathcal{I}_\Lambda(\psi) \geq \epsilon$. Using the same trick as in Corollary 2, we can write:

$$\Pr\left(\mathcal{I}_\Lambda(\Psi^{(k)}) \geq d_\Lambda\epsilon\right) \leq \Pr\left(\max_i \left|\mathrm{Tr}\left[\Psi_\Lambda^{(k)} O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda}\right| \geq \epsilon\right)\,.$$

Note that the following inequality holds

$$\left|\mathrm{Tr}\left[\Psi_\Lambda^{(k)} O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda}\right| \leq \left|\mathrm{Tr}\left[\Psi_\Lambda^{(k)} O_\Lambda^{(i)}\right] - \overline{\mathrm{Tr}\left[\Psi_\Lambda^{(k)} O_\Lambda^{(i)}\right]}\right| + \left|\overline{\mathrm{Tr}\left[\Psi_\Lambda^{(k)} O_\Lambda^{(i)}\right]} - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda}\right|\,, \tag{61}$$

then

$$\begin{aligned} \left|\overline{\mathrm{Tr}\left[\Psi_\Lambda^{(k)} O_\Lambda^{(i)}\right]} - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda}\right| &= \frac{\left|\mathrm{Tr}\left[\mathrm{Tr}_{\bar{\Lambda}}\,\Pi_{\mathrm{sym}} O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda}\right|}{D_{\mathrm{sym}}} \\ &\leq \left\| \frac{\mathrm{Tr}_{\bar{\Lambda}}[\Pi_{\mathrm{sym}}^{(k)}]}{D_{\mathrm{sym}}^{(k)}} - \frac{\mathbb{1}_\Lambda}{d_\Lambda} \right\|\,, \end{aligned}$$

where the last inequality follows from Holder's inequality. Here we have also used the known relation [80]

$$\frac{\Pi_{\rm sym}^{(k)}}{D_{\rm sym}^{(k)}} = \int \mathrm{d}\psi\, |\psi\rangle\langle\psi|^{\otimes k} = \frac{1}{D_{\rm sym}^{(k)}} \sum_{\pi} T_\pi\,, \tag{62}$$

where T_π are permutation operators between the copies of $|\psi\rangle$. Let us compute $\mathrm{Tr}_{\bar\Lambda}[\Pi_{\rm sym}^{(k)}]/D_{\rm sym}^{(k)}$, for the simplified case of $k=2$. For $k=2$ we have $\Pi_{\rm sym}^{(k)}/D_{\rm sym}^{(k)} = \frac{\mathbb{1}+T}{d(d+1)}$, where T denotes the SWAP operator. Then:

$$\mathrm{Tr}_{\bar\Lambda}\left[\frac{\mathbb{1}+T}{d(d+1)}\right] = \frac{d_{\bar\Lambda}}{d(d+1)}\mathbb{1}_\Lambda + \frac{\mathrm{Tr}_{\bar\Lambda}[T]}{d(d+1)}\,. \tag{63}$$

To compute the partial trace, let us define $\Lambda_{12} \coloneqq \{(i, n+i)\,|\, i\in\Lambda,\, n+i\in\Lambda\}$, and $\Lambda_{12}^c \coloneqq \{(i,n+i)\,|\, i\in\bar\Lambda,\, n+i\in\bar\Lambda\}$, then $\Lambda_1 \coloneqq \{(i,n+i)\,|\, i\in\Lambda,\, n+i\in\bar\Lambda\}$, and $\Lambda_2 \coloneqq \{(i,n+i)\,|\, i\in\bar\Lambda,\, n+i\in\Lambda\}$. Note that $\Lambda_{12}\cap\Lambda_{12}^c = \emptyset$,$\Lambda_{12}\cap\Lambda_1 = \emptyset$ etc., but $\Lambda_{12}\cup\Lambda_{12}^c\cup\Lambda_1\cup\Lambda_2 = 2n$. We can thus write the swap T as:

$$T = T_{\Lambda_{12}} T_{\Lambda_{12}^c} T_{\Lambda_1} T_{\Lambda_2}\,, \tag{64}$$

the partial trace thus can be written as:

$$\mathrm{Tr}_{\bar\Lambda}[T] = T_{\Lambda_{12}}\,\mathrm{Tr}_{12}[T_{\Lambda_{12}^c}]\,\mathrm{Tr}_2[T_{\Lambda_1}]\,\mathrm{Tr}_1[T_{\Lambda_2}]\,, \tag{65}$$

where the labels mean, trace over all the pair of qubits Tr_{12}, trace over the first qubits Tr_1, and trace over the second qubits Tr_2. We have $\mathrm{Tr}_{12}[T_{\Lambda_{12}^c}] = 2^{|\Lambda_{12}^c|} < d_{\bar\Lambda}$, $\mathrm{Tr}_1[T_{\Lambda_2}] = \mathbb{1}_2$, and $\mathrm{Tr}_2[T_{\Lambda_1}] = \mathbb{1}_1$. At the end, we have:

$$\begin{aligned}\left\|\mathrm{Tr}_{\bar\Lambda}[\frac{\mathbb{1}+T}{d(d+1)}] - \frac{\mathbb{1}_\Lambda}{d_\Lambda}\right\|_1 &= \left|\frac{d_{\bar\Lambda}d_\Lambda}{d(d+1)} - 1\right| \left\|\frac{\mathbb{1}_\Lambda}{d_\Lambda}\right\|_1 + \left\|\frac{\mathrm{Tr}_{\bar\Lambda}[T]}{d(d+1)}\right\|_1 \\ &\le \mathcal{O}(d^{-1}) + \frac{d_{\bar\Lambda}d_\Lambda}{d(d+1)} = \mathcal{O}(d^{-1})\,,\end{aligned}$$

where we used the fact that $T_{\Lambda_{12}}\,\mathrm{Tr}_2[T_{\Lambda_1}]\,\mathrm{Tr}_1[T_{\Lambda_2}]$ is unitary and $\|T_{\Lambda_{12}}\,\mathrm{Tr}_2[T_{\Lambda_1}]\,\mathrm{Tr}_1[T_{\Lambda_2}\|_1] = d_\Lambda$, and that $d_\Lambda d_{\bar\Lambda} = d^2$, hence $|d/(d+1) - 1| = \mathcal{O}(d^{-1})$. We thus find that for $\epsilon = 2^{-1/3n}$, one has:

$$\begin{aligned}\Pr\left(\mathcal{I}(\Psi^{(2)}) \ge d_\Lambda\epsilon\right) &\le \Pr\left(\max_i \left|\mathrm{Tr}\left[\Psi_\Lambda^{(2)} O_\Lambda^{(i)}\right] - \frac{\mathrm{Tr}\left[O_\Lambda^{(i)}\right]}{d_\Lambda}\right| \ge \epsilon\right) \\ &\le \Pr\left(\max_i \left|\mathrm{Tr}\left[\Psi^{(2)} O_\Lambda^{(i)}\right] - \overline{\mathrm{Tr}\left[\Psi_\Lambda^{(2)} O_\Lambda^{(i)}\right]}\right| \ge \epsilon\right) \\ &\le 2d_\Lambda^2 e^{-C2^n\epsilon^2/4} \equiv e^{-C2^{1/3n-4}+(2n+1)\ln 2}\,,\end{aligned}$$

where we used the fact that (asymptotically):

$$\Pr\left(\max_i \left|\overline{\mathrm{Tr}\left[\Psi_\Lambda^{(2)} O_\Lambda^{(i)}\right]} - \mathrm{Tr}[\mathcal{O}_\Lambda^{(i)}/d_\Lambda]\right| \ge 2^{-n/3}\right) = 0\,. \tag{66}$$

Thus:

$$\Pr\left(\mathcal{I}_\Lambda(\psi^{\otimes 2}) \le 2^{|\Lambda|-1/3n}\right) \ge 1 - e^{-C2^{1/3n-4}+(2n+1)\ln 2}\,. \tag{67}$$

The calculation becomes more intricate for $k>2$. However, making the assumption that Λ is symmetric between the copies of $|\psi\rangle$, with $|\Lambda| = k\lambda$ qubits, and setting $d_\Lambda \equiv 2^{k\lambda}$, we can use

$$\frac{\mathrm{Tr}_{\bar\Lambda}[\Pi_{\rm sym}^{(k)}]}{D_{\rm sym}^{(k)}} = \frac{\mathbb{1}_\Lambda 2^{(n-\lambda)k}}{D_{\rm sym}} + \sum_{c(\pi)} T_\pi \frac{2^{(n-\lambda)(k-c)}}{D_{\rm sym}^{(k)}}\,, \tag{68}$$

where the sum is over the conjugacy class $c(\pi)$ of the symmetric group S_k. As evidenced by the above formula, the order of the trace depends on the conjugacy class c. For example for $c(\pi) = (12), (23), \ldots$

one has $c = 1$, for $c(\pi) = (1234), (1243), \ldots$ one has $c = 3$ and so on. Finally:

$$\begin{aligned}\left\| \frac{\mathrm{Tr}_{\bar{\Lambda}}[\Pi^{(k)}_{\mathrm{sym}}]}{D^{(k)}_{\mathrm{sym}}} - \frac{\mathbb{1}_\Lambda}{d_\Lambda} \right\| &\leq \left| \frac{2^{(n-\lambda)}}{D^{(k)}_{\mathrm{sym}}} \right| + (k! - 1)\frac{2^{(n-\lambda)(k-1)}}{D_{\mathrm{sym}}} \\ &= \frac{1}{2^{k\lambda}} - \frac{1}{d_\Lambda} + \mathcal{O}(2^{-n}) + \mathcal{O}((k-1)!2^{-n}) \\ &= \mathcal{O}(2^{-n})\,.\end{aligned} \tag{69}$$

Thus, we have:

$$\Pr\left(\left| \mathrm{Tr}\left[\Psi^{(k)}_\Lambda\right] - \frac{\mathrm{Tr}\left[O^{(i)}_\Lambda\right]}{d_\Lambda} \right| \geq \epsilon \right) \leq \Pr\left(\left| \mathrm{Tr}\left[\Psi^{(k)}_\Lambda\right] - \overline{\mathrm{Tr}\left[\Psi^{(k)}_\Lambda O^{(i)}_\Lambda\right]} \right| \geq \epsilon \right), \tag{70}$$

provided that $\epsilon > \mathcal{O}(2^{-n})$. By applying Levy's lemma thanks to the typicality of $\mathrm{Tr}[\Psi^{(k)}_\Lambda]$, choosing $\epsilon = 2^{-1/3n}$ we have:

$$\Pr(\mathcal{I}_\Lambda(\Psi^{(k)}) \leq 2^{-1/3n}) \geq 1 - e^{\tilde{C}2^{n/3}/2k+\lambda}\,. \tag{71}$$

B Proof of Theorem 2 and Proposition 1

B.1 Proof of Theorem 2

Let us consider the following quantity

$$\Delta f_{A,B} = f(\boldsymbol{\theta}_A) - f(\boldsymbol{\theta}_B)\,, \tag{72}$$

where $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ are such that $\boldsymbol{\theta_A} = \boldsymbol{\theta_B} + \hat{\boldsymbol{e}}_{AB} l_{AB}$ with $l_{AB} \in \Omega(1/\operatorname{poly}(n))$. Our goal will be to show that the variance of this quantity is at most polynomially vanishing.

First, we will use the following notation $\boldsymbol{\theta}_n \equiv \boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_B$. Moreover, it is useful to further divide the path in the parameter space by a sequence of single parameter changes so that

$$\boldsymbol{\theta}_0 \to \boldsymbol{\theta}_1 \to \ldots \to \boldsymbol{\theta}_m\,, \tag{73}$$

where $m \in \mathcal{O}(\operatorname{poly}(n))$. Here we have defined

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_0 + \sum_{j=1}^{i} \hat{\boldsymbol{e}}_j l_j = \boldsymbol{\theta}_i + \hat{\boldsymbol{e}}_i l_i\,, \tag{74}$$

where $\hat{\boldsymbol{e}}_i$ is a vector with a single one and where at least one l_i is such that $\sin^2(l_i) \in \Omega(1/\operatorname{poly}(n))$. Note that we can guarantee both $m \in \mathcal{O}(\operatorname{poly}(n))$ and that there exists an l_i such that $\sin^2(l_i) \in \Omega(1/\operatorname{poly}(n))$ from the fact that we have at most a polynomial number of parameters, and that $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta_B}$ are at most polynomially close. The previous allows us to note that

$$\Delta f_{n,0} = \sum_{i=0}^{m-1} f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i) = \sum_{i=0}^{m-1} \Delta f_{i+1,i}\,. \tag{75}$$

where we defined $\Delta f_{i+1,i} \coloneqq f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i)$. Taking the variance of Eq. (75) we have

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}_0}[\Delta f_{n,0}] &= \mathrm{Var}_{\boldsymbol{\theta}_0}[\sum_{i=0}^{m-1} \Delta f_{i+1,i}] \\ &= \sum_{i,j=0}^{n-1} \mathrm{Cov}[\Delta f_{i+1,i}, \Delta f_{j+1,j}]\,.\end{aligned} \tag{76}$$

Here we recall that $\mathrm{Cov}[\Delta f_{i+1,i}, \Delta f_{i+1,i}] = \mathrm{Var}_{\boldsymbol{\theta}_0}[\Delta f_{i+1,i}]$.

In what follows we will use the following lemmas, which we will prove at the end of this appendix(See App. D).

Lemma 2. *Let $U(\boldsymbol{\theta})$ be a shallow HEA with depth $D \in \mathcal{O}(\log(n))$, and $O = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ with $O_\alpha^{(i)}$ being traceless operators having support on at most two neighboring qubits. Then,*

$$\mathbb{E}_{\boldsymbol{\theta}_0}[f(\boldsymbol{\theta}_i)] = 0\,, \quad \forall i\,. \tag{77}$$

Lemma 3. *Let $U(\boldsymbol{\theta})$ be a shallow HEA with depth $D \in \mathcal{O}(\log(n))$, and $O = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ with $O_\alpha^{(i)}$ being traceless operators having support on at most two neighboring qubits. Then,*

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta f_{i+1,i}\Delta f_{j+1,j}] = 0\,, \quad \forall i \neq j\,. \tag{78}$$

Lemma 4. *Let $U(\boldsymbol{\theta})$ be a shallow HEA with depth $D \in \mathcal{O}(\log(n))$ where each local two-qubit gates forms a 2-design on two qubits, and let $O = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ be the measurement composed of, at most, polynomially many traceless Pauli operators $O_\alpha^{(i)}$ having support on at most two neighboring qubits, and where $\sum_i c_i^2 \in \mathcal{O}(\text{poly}(n))$. Then, if the input state follows an area law of entanglement, we have*

$$\text{Var}[\partial_\mu f(\boldsymbol{\theta})] \in \Omega\left(\frac{1}{\text{poly}(n)}\right)\,. \tag{79}$$

Let us now go back to Eq. (76). Note that

$$\text{Cov}[\Delta f_{i+1,i}, \Delta f_{j+1,j}] = \mathbb{E}[\Delta f_{i+1,i}\Delta f_{j+1,j}] - \mathbb{E}[\Delta f_{i+1,i}]\mathbb{E}[\Delta f_{j+1,j}] = 0 \tag{80}$$

where in the second line we have used Lemmas 2 and 3. Thus, we have

$$\text{Var}_{\boldsymbol{\theta}_0}[\Delta f_{n,0}] = \sum_{i=0}^{m-1} \text{Var}[\Delta f_{i+1,i}]\,. \tag{81}$$

Using the fact that $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \hat{\boldsymbol{e}}_i l_i$, and leveraging the parameter shift-rule for computing gradients [81, 82], we then get

$$\Delta f_{i+1,i} = f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i) = 2\sin\left(\frac{l_i}{2}\right)\partial_i f(\boldsymbol{\theta}_i)\,. \tag{82}$$

and

$$\text{Var}_{\boldsymbol{\theta}_0}[\Delta f_{n,0}] = \sum_{i=0}^{m-1} 4\sin^2\left(\frac{l_i}{2}\right)\text{Var}[\partial_i f(\hat{\boldsymbol{\theta}}_i)]\,, \tag{83}$$

where we have defined $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_{i-1} + \hat{\boldsymbol{e}}_i \frac{l_i}{2}$. Then, recalling that $m \in \mathcal{O}(\text{poly}(n))$ and that there exists an l_i such that $\sin^2(l_i) \in \Omega(1/\text{poly}(n))$, we can use Lemma 4 to find

$$\text{Var}_{\boldsymbol{\theta}_0}[\Delta f_{n,0}] \in \Omega\left(\frac{1}{\text{poly}(n)}\right)\,. \tag{84}$$

B.2 Proof of Proposition 1

Here we note that in the previous section, where we have proved Theorem 2 we have shown that the absence of barren plateaus (through Lemma 4) implies cost values anti-concentration. In this section, we prove the converse.

First, we note that by anti-concentration we mean that for any set of parameters $\boldsymbol{\theta_B}$ and $\boldsymbol{\theta_A} = \boldsymbol{\theta_B} + \hat{\boldsymbol{e}}_{AB} l_{AB}$ with $l_{AB} \in \Omega(1/\text{poly}(n))$ one has

$$\text{Var}_{\boldsymbol{\theta}_B}[f(\boldsymbol{\theta}_A) - f(\boldsymbol{\theta}_B)] \in \Omega\left(\frac{1}{\text{poly}(n)}\right)\,. \tag{85}$$

Then, let us use the fact that

$$\text{Var}_{\boldsymbol{\theta}}[\partial_\nu f(\boldsymbol{\theta})] = \frac{1}{4}\text{Var}[f(\boldsymbol{\theta}^+) - f(\boldsymbol{\theta}^-)]\,, \tag{86}$$

where we have used the parameter-shift rule, and where $\boldsymbol{\theta}^+ = \boldsymbol{\theta}^+ \pm \hat{\boldsymbol{e}}_\nu \frac{\pi}{2}$ such that $\hat{\boldsymbol{e}}_\nu$ is a unit vector with a one at the ν-th entry. Then, we have that since the difference between $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ is $\mathcal{O}(1)$, we can use Eq. (85) to find

$$\text{Var}_{\boldsymbol{\theta}}[\partial_\nu f(\theta)] \in \Omega\left(\frac{1}{\text{poly}(n)}\right)\,, \tag{87}$$

which completes the proof of Proposition 1.

C Isospectral twirling and proof of Proposition 3,Thereom 3 and Proposition 4

In this section, we aim to prove Proposition 3, Theorem 3, Corollary 5, and Proposition 4.

C.1 Isospectral twirling

We first, review the useful notion of the Isospectral twirling introduced in Refs. [77, 78]. Consider an Hamiltonian H written in its spectral decomposition $H = \sum_k E_k \Pi_k$, where Π_k are its eigenvectors and E_k are its eigenvalues. Consider the time-evolution generated by H, i.e., $W(t) = \exp(-iHt) = \sum_k e^{-iE_k t}\Pi_k$. Denote $\mathbb{U}(n)$ the unitary group on n qubits. Define the following ensemble of isospectral unitary evolutions:

$$\mathcal{E}_H = \{G^\dagger \exp(-iHt)\, G \,|\, G \in \mathbb{U}(n)\} \tag{88}$$

whose representative element is H. The Isospectral twirling of order k, denoted as $\mathcal{R}^{(2k)}(W(t))$ is the $2k$-fold Haar channel of the operator $W^{\otimes k,k}(t) \coloneqq W^{\otimes k}(t) \otimes W^{\dagger \otimes k}(t)$:

$$\mathcal{R}^{(2k)}(W(t)) \coloneqq \int_{\mathbb{U}(n)} \mathrm{d}G\, G^{\dagger \otimes 2k} W^{\otimes k,k}(t) G^{\otimes 2k}\,. \tag{89}$$

Using the Weingarten functions [83–85], one can compute the isospectral twirling as:

$$\mathcal{R}^{(2k)}(W(t)) = \sum_{\pi\sigma \in S_{2k}} (\Omega^{-1})_{\pi\sigma} \operatorname{Tr}[W^{\otimes k,k}(t) T_\pi] T_\sigma\,, \tag{90}$$

where S_{2k} is the symmetric group of order $2k$, T_π is the unitary representation of the permutation $\pi \in S_{2k}$, and $\Omega_{\pi\sigma} \coloneqq \operatorname{Tr}[T_\pi T_\sigma]$. Note that $\operatorname{Tr}[W^{\otimes k,k}(t) T_\pi]$ are spectral functions of the representative Hamiltonian H. $\pi = e$ (the identity) defines the $2k$-spectral form factor $c_{2k}(t) \equiv |\operatorname{Tr}[W(t)]|^{2k}$:

$$c_{2k}(t) = \left(\sum_{kl} e^{i(E_k - E_l)t}\right)^k \tag{91}$$

which governs the behavior of many figures of merit of random isospectral Hamiltonians, as noted in [77, 78]. While the expression for Isospectral twirling of order k, for $k \geq 2$ is cumbersome to report, we do recall the expression for the Isospectral twirling for $k = 1$:

$$\mathcal{R}^{(2)}(t) = \frac{c_2(t) - 1}{d^2 - 1} \mathbb{1}^{\otimes 2} + \frac{d^2 - c_2(t)}{d^2 - 1} \frac{T_{12}}{d}\,, \tag{92}$$

where T_{12} is the swap operator between the two copies of $\mathcal{H}$, and $c_2(t)$ is the 2-point spectral form factor in Eq. (91). One can consider the isospectral twirling of a scalar function of the time evolution operator $W(t)$ (characterized by the operator of interest O), i.e., $F_O[W(t)]$, which can be written after the isospectral twirling as $\left\langle F_O[G^\dagger W(t) G]\right\rangle_G \coloneqq \operatorname{Tr}[T_\sigma O \mathcal{R}^{(2k)}(t)]$, where T_σ is a particular permutation operator. Its value (depending on the evolution time t) characterizes the average behavior in the ensemble $\mathcal{E}_H$ of all those Hamiltonians sharing the same spectrum of H. Since the Isospectral twirling of a scalar function depends upon the particular choice of the spectrum of H, one then averages over spectra of a given ensemble of Hamiltonians E. Relevant examples are the Gaussian unitary ensemble $E \equiv$ GUE, the Poisson ensemble $E \equiv$ P, or the Gaussian diagonal ensemble $E \equiv$ GDE. As one can see, in this picture, spectra and eigenvectors become completely unrelated, since the average over the full unitary group erases the information about the eigenvectors. Although, in Refs. [77, 78] many ensembles of Hamiltonians have been considered, in this paper, we are particularly interested in the GDE ensemble which is the simplest ensemble of Hamiltonians: the $2k$-point spectral form factors can readily be computed. Let $\mathrm{sp}(H) \coloneqq \{E_k\}_{k=1}^d$, where $d \equiv 2^n$. The GDE ensemble is characterized by the following probability distribution for $\mathrm{sp}(H)$:

$$P_{\mathrm{GDE}}(\{E_k\}) \coloneqq \left(\frac{2}{\pi}\right)^{d/2} e^{-2\sum_k E_k^2}\,, \tag{93}$$

i.e., all the eigenvalues E_k are independent identically distributed Gaussian random variables with zero mean and standard deviation $1/2$. Define $\tilde{c}_{2k}(t) \coloneqq \frac{c_{2k}(t)}{d^{2k}}$, then the average of the normalized $2k$-spectral form factors $\tilde{c}_{2k}(t)$ reads

$$\overline{\tilde{c}_{2k}(t)}^{\mathrm{GDE}} = e^{-kt^2/4} + \mathcal{O}(d^{-1})\,. \tag{94}$$

Let us now build a notation useful throughout the following proofs. Let $F[W(t)]$ be a scalar function of the unitary evolution $W(t) = \exp(-iHt)$. Then we denote with $\mathbb{E}_{\text{GDE}}$ the Isospectral twirling of the scalar function $F[W(t)]$ followed by the average over the GDE ensemble of Hamiltonian, namely:

$$\mathbb{E}_{\text{GDE}}[F[W(t)]] \coloneqq \overline{\int \mathrm{d}G F[G^\dagger W(t) G]}^{\text{GDE}} . \tag{95}$$

In the following section, we use the techniques introduced in order to prove Proposition 3.

C.2 Proof of Proposition 3

While the proof of Eq. (26) is straightforward, here we prove Eq. (27) via the Isospectral twirling technique. Recall $L_s(H_G, \boldsymbol{\theta}, t) = \text{Tr}[W(t) |\psi_s\rangle\langle\psi_s| W^\dagger(t) O(\boldsymbol{\theta})]$, where $|\psi_s\rangle$ is a completely factorized state. We are interested in computing $\mathbb{E}_{\text{GDE}}[L_s(H_G, \boldsymbol{\theta}, t)]$. Note that $L_s(H_G, \boldsymbol{\theta}, t)$ can be written as:

$$L_s(H_G, \boldsymbol{\theta}, t) = \text{Tr}[T_{12} \psi_s \otimes O(\boldsymbol{\theta}) W^{\otimes 1,1}(t)] . \tag{96}$$

The average over GDE Hamiltonian can be readily taken out of Eq. (92), and Eq. (94). By using that $\text{Tr}[O(\boldsymbol{\theta})] = 0$, one gets:

$$\mathbb{E}_{\text{GDE}}[L_s(H_G, \boldsymbol{\theta}_0, t)] = e^{-t^2/4} \text{Tr}[\psi_s O(\boldsymbol{\theta}_0)] + \mathcal{O}(d^{-1}) , \tag{97}$$

to recover Eq. (27) it is sufficient to note that $\text{Tr}[\psi_s O(\boldsymbol{\theta}_0)] = 1$ because $O(\boldsymbol{\theta}_0) \equiv S$, and $S |\psi_s\rangle = |\psi_s\rangle$ by definition.

C.3 Proof of Theorem 3

To prove the theorem, we use the well-known Markov inequality: let x be a positive semi-definite random variable with average μ, then:

$$\Pr(x \geq \epsilon) \leq \frac{\mu}{\epsilon} . \tag{98}$$

To prove the concentration consider $|L_s(H, \boldsymbol{\theta}, t)|$ for $H \in \{H_G, H_S\}$ defined in the main text. Then by Jensen's inequality, we have:

$$\mathbb{E}_{\text{GDE}}[|L_s(H, \boldsymbol{\theta}, t)|] \leq \sqrt{\mathbb{E}_{\text{GDE}}[L_s(H, \boldsymbol{\theta}, t)^2]} . \tag{99}$$

Note that, one can write

$$L_s(H, \boldsymbol{\theta}, t)^2 = \text{Tr}[T_{(13)(24)} \psi_s^{\otimes 2} \otimes O(\boldsymbol{\theta}) W^{\otimes 2,2}(t)] . \tag{100}$$

Then, taking the isospectral twirling, one has:

$$\left\langle L_s(G^\dagger H G, \boldsymbol{\theta}, t)^2 \right\rangle_G = \text{Tr}[T_{(13)(24)} \psi_s^{\otimes 2} \otimes O(\boldsymbol{\theta}) \mathcal{R}^{(4)}(t)] , \tag{101}$$

where the Isospectral twirling operator $\mathcal{R}^{(4)}(t)$ can be found in Eq. (165) of Ref. [77]. Taking the average over the GDE spectra, one has

$$\begin{aligned} \mathbb{E}_{\text{GDE}}[L_s(H_G, \boldsymbol{\theta}, t)^2] &= e^{-t^2/2} \text{Tr}[\psi_s O(\boldsymbol{\theta})]^2 + \mathcal{O}(d^{-1}) \\ \mathbb{E}_{\text{GDE}}[L_s(H_S, \boldsymbol{\theta}, t)^2] &= e^{-t^2/2} \text{Tr}[\psi_s O(\boldsymbol{\theta})]^2 + \mathcal{O}(d^{-1}) , \end{aligned}$$

where the result is obtained after some algebra, with the hypothesis of $\psi \equiv \psi_A \otimes \psi_B$ and that $\text{Tr}[O] = 0$. Note that the above result is obtained for $H \in \{H_G, H_S\}$; we indeed exploited the fact that $|A| \ll |B|$, and thus $d_B = \mathcal{O}(d)$. Using Markov's inequality Eq. (98), the result can be readily derived.

C.4 Proof of Theorem 4

C.4.1 An anti-concentration inequality

To prove the theorem, we make use of Cantelli's inequality: let x be a random variable with average μ and standard deviation σ. Then:

$$\Pr(x \geq \mu - \theta\sigma) \geq \frac{\theta^2}{1+\theta^2} \tag{102}$$

To bound the measure $\mathcal{I}_\Lambda(\psi)$, we make use of the bound proven in [86] between the Hilbert-Schmidt distance and the trace distance, namely:

$$\mathcal{I}_\Lambda(\psi) \geq \frac{1}{2}\sqrt{\left\|\psi_\Lambda - \frac{\mathbb{1}_\Lambda}{d_\Lambda}\right\|_2}, \tag{103}$$

where $\psi_\Lambda \coloneqq \mathrm{Tr}_{\bar{\Lambda}}[|\psi\rangle\langle\psi|]$, and $d_\Lambda = 2^{|\Lambda|}$. Note that

$$\left\|\psi_\Lambda - \frac{\mathbb{1}_\Lambda}{d_\Lambda}\right\|_2 = \sqrt{P(\psi_\Lambda) - d_\Lambda^{-1}}\,, \tag{104}$$

where $P(\psi_\Lambda) \equiv \mathrm{Tr}[\psi_\Lambda^2]$ is the purity of the reduced density matrix ψ_Λ. Let $|\psi_t\rangle = \exp(-iHt)\,|\psi_s\rangle$ the state resulting from the time evolution under a GDE Hamiltonian acting on a completely factorized state. Let Λ be a subsystem such that $|\Lambda| = \mathcal{O}(\log(n))$. Note that if $\Pr(P(\psi_\Lambda) \geq \epsilon) \geq \delta$, then

$$\Pr\left(\left\|\psi_\Lambda - \frac{\mathbb{1}_\Lambda}{d_\Lambda}\right\|_2 \geq \sqrt{\epsilon - d_\Lambda^{-1}}\right) \geq \delta\,, \tag{105}$$

and, because of Eq. (103), one has:

$$\Pr\left(\mathcal{I}_\Lambda(\psi) \geq \frac{1}{2}(\epsilon - d_\Lambda^{-1})^{1/4}\right) \geq \delta\,. \tag{106}$$

Thus, from the above chain of inequality, it is clear that it is sufficient to obtain a bound on the anti-concentration of the purity $P(\psi_\Lambda)$. Denoting $P_{\mathrm{GDE}} \coloneqq \mathbb{E}_{\mathrm{GDE}}[P(\psi_\Lambda)]$, and $Q_{\mathrm{GDE}} = \mathbb{E}_{\mathrm{GDE}}[P(\psi_\Lambda)^2]$, we can use Eq. (102) to write:

$$\Pr\left(P(\psi_\Lambda) \geq P_{\mathrm{GDE}} - \theta\sqrt{Q_{\mathrm{GDE}} - P_{\mathrm{GDE}}^2}\right) \geq \frac{\theta^2}{1+\theta^2}\,. \tag{107}$$

C.4.2 Computing P_{GDE}

First, let us note the following fact. $P(\psi_\Lambda) = \mathrm{Tr}[T_\Lambda \psi_t^{\otimes 2}]$, where $\psi_t = W_G(t)\psi_0 W_G^\dagger(t)$, and where $W_G(t) \equiv G^\dagger W(t) G$. The Isospectral twirling of $P(\psi_\Lambda)$ is the average with respect to $G \in \mathbb{U}(n)$.

$$\langle P(\psi_\Lambda)\rangle_G = \left\langle \mathrm{Tr}[T_\Lambda W_G(t)\psi_0^{\otimes 2} W_G^\dagger(t)]\right\rangle_G\,. \tag{108}$$

Thanks to the right/left invariance of the Haar measure, we can insert unitaries of the form $V \equiv V_\Lambda \otimes V_{\bar{\Lambda}}$ because $[V, T_\Lambda] = 0$. This means that

$$\langle P(\psi_\Lambda)\rangle_G = \langle P(\psi_\Lambda^{(V)})\rangle_G\,, \tag{109}$$

where $\psi_\Lambda^{(V)}$ is defined as $\psi_\Lambda^{(V)} \coloneqq \mathrm{Tr}_{\bar{\Lambda}}\, W_G(t)\psi^{(V)} W_G^\dagger(t)$, where

$$\psi^{(V)} \coloneqq \int_{\mathbb{U}(n_\Lambda)} \mathrm{d}V_\Lambda \int_{\mathbb{U}(n_{\bar{\Lambda}})} \mathrm{d}V_{\bar{\Lambda}} (V_\Lambda \otimes V_{\bar{\Lambda}})^{\otimes 2} \psi_0 (V_\Lambda \otimes V_{\bar{\Lambda}})^{\dagger\otimes 2}\,, \tag{110}$$

i.e., we can compute the average purity over GDE Hamiltonian on the average product state between Λ and $\bar{\Lambda}$. A straightforward computation leads us to

$$\langle P(\psi_\Lambda)\rangle_G = \frac{d(d_\Lambda + d_{\bar{\Lambda}}) + \langle \mathrm{Tr}[T_\Lambda W_G^{\otimes 2}(t) T_\Lambda W_G^{\dagger\otimes 2}(t)]\rangle_G}{d(d_\Lambda+1)(d_{\bar{\Lambda}}+1)} + \frac{\langle \mathrm{Tr}[T_{\bar{\Lambda}} W_G^{\otimes 2}(t) T_\Lambda W_G^{\dagger\otimes 2}(t)]\rangle_G}{d(d_\Lambda+1)(d_{\bar{\Lambda}}+1)}\,. \tag{111}$$

Following the calculations in Sec. 3.3.1 of Ref. [77], the result can be proven to be:

$$P = \frac{1}{d_\Lambda} + e^{-t^2/2}\left(1 - \frac{1}{d_\Lambda}\right) + \mathcal{O}(d^{-1})\,. \tag{112}$$

C.4.3 Computing Q_{GDE}

In this section, we compute the second moment of the purity, i.e. $\langle P(\psi_\Lambda)^2\rangle$. Thanks to the left/right invariance of the Haar measure over G, and from the commutation with T_Λ, we can compute the average of the second moment for any factorized state $|\psi_0\rangle \equiv |\psi_\Lambda\rangle \otimes |\psi_{\bar{\Lambda}}\rangle$, by computing it with the input state $|\psi_0\rangle \equiv |0\rangle^{\otimes n}$. Namely:

$$\left\langle P(\psi_\Lambda)^2\right\rangle = \langle \mathrm{Tr}^2(T_\Lambda W_G^{\otimes 2}(t)\,|0\rangle\langle 0|^{\otimes 2n}\, W_G^{\dagger\otimes 2})(t)\rangle_G \tag{113}$$

Computing the above expression is trickier than computing P_{GDE} because the latter involves averaging over the 8-th fold power of G. It is well known that the permutation group S_8 contains 8! elements, making the calculation too expensive. It is also known that purity and OTOCs possess many similarities (see Refs. [87–91]). Indeed, the strategy to do the calculation will be: (i) reduce each term of Eq. (113) to a sum of high-order OTOCs, and (ii) use the following asymptotic formula proven in [92]: let $A_1,\dots,A_k,B_1,\dots,B_k$ non-identity Pauli operators, and let $B_l(t) \equiv W_G^\dagger(t) B_l W_G(t)$:

$$\left\langle \mathrm{Tr}\left[\prod_l A_l B_l(t)\right]\right\rangle_G = \mathrm{Tr}\left[\prod_l A_l B_l\right]\tilde{c}_{2k}(t) + \mathcal{O}(d^{-1})\,, \tag{114}$$

where $\tilde{c}_{2k}(t)$ is the normalized $2k$-spectral form factor defined in Eq. (91). The strategy is thus to write (113) in terms of 8−OTOCs. First, note that we can write the swap operator as:

$$T_\Lambda = \frac{\mathbb{1}_\Lambda}{d_\Lambda} + \frac{1}{d_\Lambda}\sum_{P_\Lambda\neq\mathbb{1}} P_\Lambda^{\otimes 2}\,, \tag{115}$$

and, in the same fashion, we can write the state $|0\rangle\langle 0|$ as:

$$|0\rangle\langle 0| = \sum_{P\in\{\mathbb{1},Z\}^n} P\,. \tag{116}$$

Thus, we can write:

$$P^2(\psi_\Lambda) = \sum_{P_i,P_\Lambda^a,P_\Lambda^b} \mathrm{Tr}[P_\Lambda^a\tilde{P}_1]\,\mathrm{Tr}[P_\Lambda^a\tilde{P}_2]\,\mathrm{Tr}[P_\Lambda^b\tilde{P}_3]\,\mathrm{Tr}[P_\Lambda^b\tilde{P}_4]\,.$$

Note that the sum is over P_i for $i=1,\dots,4$ and runs over the subgroup $P\in\{\mathbb{1},Z\}^n$; moreover we defined $\tilde{P}_i \coloneqq U_G P_i U_G^\dagger$ for $i=1,\dots,4$ to light the notation. It is useful for the following to split the identity part of the sum, and write $P^2(\psi_\Lambda)$ as:

$$P^2(\psi_\Lambda) = \sum_{P_i,P_\Lambda^a,P_\Lambda^b\neq\mathbb{1}} \mathrm{Tr}[P_\Lambda^a\tilde{P}_1]\,\mathrm{Tr}[P_\Lambda^a\tilde{P}_2]\,\mathrm{Tr}[P_\Lambda^b\tilde{P}_3]\,\mathrm{Tr}[P_\Lambda^b\tilde{P}_4] + \frac{1}{d_A}P(\psi_\Lambda) - \frac{2}{d_A^2}\,, \tag{117}$$

where each term on the first sum is different from $\mathbb{1}$. To write it in terms of OTOCs, let us use the following identity: let A and B two operators, and P Pauli operators then:

$$\frac{1}{d}\sum_P \mathrm{Tr}[APBP] = \mathrm{Tr}[A]\,\mathrm{Tr}[B]\,, \tag{118}$$

where the summation is on the whole Pauli group. Thus

$$P^2(\psi_\Lambda) = \sum_{\substack{P_i,P_\Lambda^a,P_\Lambda^b\neq\mathbb{1}\\ Q,K,L}} \mathrm{Tr}\left[P_\Lambda^a\tilde{P}_1 Q P_\Lambda^a\tilde{P}_2 Q K P_\Lambda^b\tilde{P}_3 K L P_\Lambda^b\tilde{P}_4 L\right] + \frac{1}{d_A}P(\psi_\Lambda) - \frac{2}{d_A^2}\,. \tag{119}$$

Here, Q,K,L are labeling global Pauli operators, running on the whole Pauli group. In order to recover OTOCs, we still need to split the sum of Q,K,L between the identity and the other non-identity Paulis:

this operation gives rise to 8 different terms, labeled $t_1, \dots, t_8$, each of which proportional to the spectral function $\tilde{c}_8(t)$ thanks to Eq. (114). The coefficients are listed below:

$$
\begin{aligned}
t_1 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 P_\Lambda^a P_2 P_\Lambda^b P_3 P_\Lambda^b P_4] \\
t_2 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 Q P_\Lambda^a P_2 Q P_\Lambda^b P_3 P_\Lambda^b P_4] \\
t_3 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 P_\Lambda^a P_2 K P_\Lambda^b P_3 K P_\Lambda^b P_4] \\
t_4 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 P_\Lambda^a P_2 P_\Lambda^b P_3 L P_\Lambda^b P_4 L] \\
t_5 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 Q P_\Lambda^a P_2 Q K P_\Lambda^b P_3 K P_\Lambda^b P_4] \\
t_6 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 Q P_\Lambda^a P_2 Q P_\Lambda^b P_3 L P_\Lambda^b P_4 L] \\
t_7 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 P_\Lambda^a P_2 K P_\Lambda^b P_3 K L P_\Lambda^b P_4 L] \\
t_8 &\equiv \sum \mathrm{Tr}[P_\Lambda^a P_1 Q P_\Lambda^a P_2 Q K P_\Lambda^b P_3 K L P_\Lambda^b P_4 L]
\end{aligned}
\tag{120}
$$

where we adopted the following convention: the sum $\sum$ contains the sum over all the Pauli's appearing on the summation with the exception on the identity, running on their respective support; more precisely, $P_1, \dots, P_4$ runs in the subgroup $P_i \in \{\mathbb{1}, Z\}^n$ but the identity, Q, K, L run over the whole Pauli group but the identity, while P_Λ^a, P_Λ^b run over the whole Pauli group defined on Λ qubits but the identity. Using Eq. (112), from Eq. (119), one thus arrives to

$$
\langle P^2(\psi_\Lambda)\rangle_G = -\frac{1}{d_\Lambda^2} + 2\tilde{c}_4 \frac{d_\Lambda - 1}{d_\Lambda^2} + \tilde{c}_8(t) \sum_{i=1}^{8} t_i + \mathcal{O}(d^{-1})\,.
$$

After further algebraic simplifications, one gets:

$$
\sum_{i=1}^{8} t_i = \frac{(d_\Lambda - 1)^2}{d_\Lambda^2} + \mathcal{O}(d^{-2})\,. \tag{121}
$$

In order to properly use Eq. (114), one needs to be sure to absorb all the term $\mathcal{O}(d^{-1})$. The final result is:

$$
\langle P^2(\psi_\Lambda)\rangle = \frac{1 + 2\tilde{c}_4(d_\Lambda - 1) + \tilde{c}_8(d_\Lambda - 1)^2}{d_\Lambda^2} + \mathcal{O}(d^{-1})\,. \tag{122}
$$

Hence, we can readily compute Q_{GDE} by using Eq. (94), and obtain:

$$
Q_{\mathrm{GDE}} = \frac{1 + 2e^{-t^2/2}(d_\Lambda - 1) + e^{-t^2}(d_\Lambda - 1)^2}{d_\Lambda^2} + \mathcal{O}(d^{-1})\,. \tag{123}
$$

C.4.4 Final bound

Putting Eq. (111) and Eq. (122) together, we obtain the fluctuations of the purity with respect to an isospectral ensemble of Hamiltonian as a function of t:

$$
\Delta_G[P(\psi_\Lambda)] = (\tilde{c}_8(t) - \tilde{c}_4^2(t)) \frac{(d_\Lambda - 1)^2}{d_\Lambda^2} + \mathcal{O}(d^{-1})\,, \tag{124}
$$

where $\Delta_G[P(\psi_\Lambda)] \coloneqq \langle P^2(\psi_\Lambda)\rangle_G - \langle P(\psi_\Lambda)\rangle_G^2$. Note that the above result is completely general, and valid for any isospectral family of Hamiltonians. Further, substituting Eq. (112) and Eq. (123) we get that the purity fluctuations with respect to GDE Hamiltonians are

$$
Q_{\mathrm{GDE}} - P_{\mathrm{GDE}}^2 = \mathcal{O}(d^{-1})\,. \tag{125}
$$

By using Eq. (107), and the fact that the fluctuations in Eq. (125) are $\mathcal{O}(d^{-1})$, we can choose $\theta = \mathcal{O}(d^{1/4})$, to write the following anti-concentration bound:

$$
\begin{aligned}
\Pr[P(\psi_\Lambda) &\geq d_\Lambda^{-1} + e^{-t^2/2}(1 - d_\Lambda^{-1})/2 + \mathcal{O}(d^{-1/4})] \\
&\geq 1 - \mathcal{O}(d^{-1/2})\,.
\end{aligned}
\tag{126}
$$

Now, we can finally use Eq. (106) to prove the theorem:

$$\Pr(\mathcal{I}_\Lambda(\psi) \geq e^{-t^2/8}(1-d_\Lambda^{-1})^{1/4}/2) \geq 1 - \mathcal{O}(d^{-1/2})\,. \tag{127}$$

Thus choosing $t = \mathcal{O}(\sqrt{\log(n)})$, and recalling that $|\Lambda| = \mathcal{O}(\log(n))$:

$$\Pr\left[\mathcal{I}_\Lambda(\psi) \in \Omega\left(\frac{1}{\text{poly}(n)}\right)\right] \geq 1 - \mathcal{O}(2^{-n/2})\,, \tag{128}$$

which concludes the proof.

D Useful Lemmas

Before proceeding to prove Lemmas, we recall that if a given two-qubit gate V in the HEA forms a 2-design, one can employ the following element-wise formula of the Weingarten calculus in Refs. [85, 93] to explicitly evaluate averages over V up to the second moment:

$$\begin{aligned} \int dV\, v_{\boldsymbol{ij}} v^*_{\boldsymbol{i'j'}} &= \frac{\delta_{\boldsymbol{ii'}}\delta_{\boldsymbol{jj'}}}{4} \\ \int dV\, v_{\boldsymbol{i}_1\boldsymbol{j}_1} v_{\boldsymbol{i}_2\boldsymbol{j}_2} v^*_{\boldsymbol{i'}_1\boldsymbol{j'}_1} v^*_{\boldsymbol{i'}_2\boldsymbol{j'}_2} &= \frac{1}{15}\left(\Delta_1 - \frac{\Delta_2}{4}\right) \end{aligned} \tag{129}$$

where $v_{\boldsymbol{ij}}$ are the matrix elements of V, and

$$\begin{aligned} \Delta_1 &= \delta_{\boldsymbol{i}_1\boldsymbol{i'}_1}\delta_{\boldsymbol{i}_2\boldsymbol{i'}_2}\delta_{\boldsymbol{j}_1\boldsymbol{j'}_1}\delta_{\boldsymbol{j}_2\boldsymbol{j'}_2} + \delta_{\boldsymbol{i}_1\boldsymbol{i'}_2}\delta_{\boldsymbol{i}_2\boldsymbol{i'}_1}\delta_{\boldsymbol{j}_1\boldsymbol{j'}_2}\delta_{\boldsymbol{j}_2\boldsymbol{j'}_1}\,, \\ \Delta_2 &= \delta_{\boldsymbol{i}_1\boldsymbol{i'}_1}\delta_{\boldsymbol{i}_2\boldsymbol{i'}_2}\delta_{\boldsymbol{j}_1\boldsymbol{j'}_2}\delta_{\boldsymbol{j}_2\boldsymbol{j'}_1} + \delta_{\boldsymbol{i}_1\boldsymbol{i'}_2}\delta_{\boldsymbol{i}_2\boldsymbol{i'}_1}\delta_{\boldsymbol{j}_1\boldsymbol{j'}_1}\delta_{\boldsymbol{j}_2\boldsymbol{j'}_2}\,. \end{aligned} \tag{130}$$

Here the integration is taken over $\mathbb{U}(2)$, the unitary group of degree 2.

Lemma 1. *Let $U(\boldsymbol{\theta})$ be a HEA with depth D as in Fig. 1(b), and $O = \sum_i c_i P_i = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ be an operator, as in Eq. (35). Then for any $O_\alpha^{(i)}$:*

$$|\operatorname{supp}(U(\boldsymbol{\theta})O_\alpha^{(i)}U^\dagger(\boldsymbol{\theta}))| \leq \sum_{\substack{k\in C_\alpha^{(i)} \\ l_k^{(i)}<2D}} l_k^{(i)} + 2D\,. \tag{131}$$

Proof. Given $O = \sum_i c_i P_i = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ (from the clustering of qubits), to prove the statement, one has to look at a given operator $O_\alpha^{(i)}$, having support on the cluster $C_\alpha^{(i)}$ defined in Sec. A.

As a starting point, let us introduce the local Pauli operator $\sigma_j^{(i)}$ possessing support only on the $j-$qubit, and then write $U(\boldsymbol{\theta})$ as in Eq. (3) layer by layer, as

$$U(\boldsymbol{\theta}) = V_1 \ldots V_{D-1}V_D \tag{132}$$

where V_k are the unitaries acting on each layer $k = 1, \ldots, D$. For sake of simplicity, we drop the parameter dependence. Each V_k can be further decomposed as:

$$V_k = V_1^{(k)} \otimes \ldots \otimes V_{\frac{n}{2m}}^{(k)} \otimes V_{\frac{n}{2m}-1}^{(k)} \tag{133}$$

where each unitaries V_α, $\alpha = 1, \ldots, n/2m$, acts on m qubits, being n a multiple of m (m even), with periodic boundary conditions; so that either (i) V_1 acts on the first m qubits, V_2 acts on the second m qubits, up to $V_{n/2m}$ acting on the last m qubits; or (ii) V_1 is acting from qubit $m/2+1$ to qubit $3m/2$, V_2 is acting from qubit $3m/2+1$ to qubit $2m$, up to $V_{n/2m}$ acting from qubit $n-m/2+1$ to qubit $m/2$. At the first layer, only one $V_\alpha^{(1)}$ for some α is acting on $\sigma_j^{(i)}$:

$$V_1^\dagger \sigma_j^{(i)} V_1 = V_\alpha^{(1)\dagger}\sigma_j^{(i)}V_\alpha^{(1)} \tag{134}$$

increasing the support to (at most) m qubits, $\text{supp}(V_\alpha^{(k)\dagger} P_i V_\alpha^{(k)}) \leq m$. At the second layer, there will be (at most) two $V_\beta^{(2)}$, $V_\gamma^{(2)}$ for some β, γ acting non-trivially on $V_\alpha^{(1)\dagger}\sigma_j^{(i)} V_\alpha^{(1)}$:

$$V_2^\dagger V_1^\dagger \sigma_j^{(i)} V_1 V_2^\dagger = V_\gamma^{(2)\dagger} V_\beta^{(2)\dagger} V_\alpha^{(1)\dagger} \sigma_j^{(i)} V_\alpha^{(1)} V_\beta^{(2)} V_\gamma^{(2)} \tag{135}$$

increasing the support to at most $2m$. After the second layer, more unitaries will act on it trivially, but the support can only increase by a factor m at each layer. This is because, at the k-th layer, only the unitaries acting on the boundary of $\text{supp}(V_{k-1}^\dagger \sigma_j^{(i)} V_{k-1})$ can increase the size of a factor $m/2$ each. Having shown the statement for an ultra-local operator, now the proof follows easily. Indeed, O_α has support on the cluster C_α, which contains non-first-neighboring qubits whose relative distance is less than mD. The support of each operator living on the $q-$th qubit of C_α will increase by a factor mD after the action of $U(\boldsymbol{\theta})$, which means that the supports are overlapping after the evolution, and thus one can upper-bound the total support of $U^\dagger(\boldsymbol{\theta}) O_\alpha U(\boldsymbol{\theta})$ as:

$$|\,\text{supp}(U^\dagger(\boldsymbol{\theta}) O_\alpha U(\boldsymbol{\theta}))| \leq \sum_{k \in \mathcal{C}_\alpha} l_k + mD \tag{136}$$

i.e. by summing all the relative distances between the qubits within C_α. This concludes the proof. □

Corollary 5. *Let $U(\boldsymbol{\theta})$ be HEA with depth D, with $D \in \mathcal{O}(\log(n))$ and $O = \sum_i c_i P_i = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$. Then for any $O_\alpha^{(i)}, O_{\alpha'}^{(i)}$, with $\alpha \neq \alpha'$, one finds that asymptotically,*

$$\text{supp}(U(\boldsymbol{\theta}) O_\alpha^{(i)} U^\dagger(\boldsymbol{\theta})) \cap \text{supp}(U(\boldsymbol{\theta}) O_{\alpha'}^{(i)} U^\dagger(\boldsymbol{\theta})) = \emptyset\,. \tag{137}$$

Proof. The corollary easily descends from Lemma 1 and relative proof. Indeed, each operator O_α has support on a cluster C_α. This means that two different operators $O_\alpha, O_{\alpha'}$ have support with relative distance greater than mD. Since the support of each operator $O_\alpha, O_{\alpha'}$ can increase no more than $2D$ away from any qubit $q \in C_\alpha$, we have:

$$\text{supp}(U^\dagger(\boldsymbol{\theta}) O_\alpha U(\boldsymbol{\theta})) \cap \text{supp}(U^\dagger(\boldsymbol{\theta}) O_{\alpha'} U(\boldsymbol{\theta})) = \emptyset \tag{138}$$

by construction. It is important to remark that the above result is only valid asymptotically. Indeed, for small n, say n_0, it can be the case that $n \in \mathcal{O}(\text{poly}(\log(n)))$. This concludes the proof. □

Lemma 2. *Let $U(\boldsymbol{\theta})$ be a shallow HEA with depth $D \in \mathcal{O}(\log(n))$, and $O = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ with $O_\alpha^{(i)}$ being traceless operators having support on at most two neighboring qubits. Then,*

$$\mathbb{E}_{\boldsymbol{\theta}_0}[f(\boldsymbol{\theta}_i)] = 0\,, \quad \forall i\,. \tag{139}$$

Proof. Let us recall that $f(\boldsymbol{\theta}_i) = \text{Tr}[U(\boldsymbol{\theta}_i)\,|\psi\rangle\langle\psi|\,U^\dagger(\boldsymbol{\theta}_i) O]$, where $U(\boldsymbol{\theta}_i)$ is a HEA where each two-qubit V_i forms a local 2-design (see Fig. 1). From the previous, we know that

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\cdot] = \prod_{V_i} \int_{\mathbb{U}(2)} d\mu(V_i)(\cdot)\,. \tag{140}$$

In particular, given a single two-qubit gate V_μ, let us compute

$$\int_{\mathbb{U}(2)} dV_\mu f(\hat{\boldsymbol{\theta}}_i) = \int_{\mathbb{U}(2)} dV_\mu \text{Tr}[V_\mu \widetilde{\rho}_\mu V_\mu^\dagger \widetilde{O}_\mu]\,, \tag{141}$$

where we have defined

$$\widetilde{\rho}_\mu = U_B^\mu\,|\psi\rangle\langle\psi|\,(U_B^\mu)^\dagger \tag{142}$$

$$\widetilde{O}_\mu = (U_A^\mu)^\dagger O U_A^\mu \tag{143}$$

with $U(\boldsymbol{\theta}) = U_A^\mu V_\mu U_B^\mu$. In Fig. 9(a) we schematically show the tensor representation of contraction of Eq. (141). A direct calculation using Eq. (129) shows that

$$\int_{\mathbb{U}(2)} dV_\mu f(\boldsymbol{\theta}_i) = \frac{1}{4}\text{Tr}[\text{Tr}_\mu[\widetilde{\rho}_\mu]\text{Tr}_\mu[\widetilde{O}_\mu]]\,, \tag{144}$$

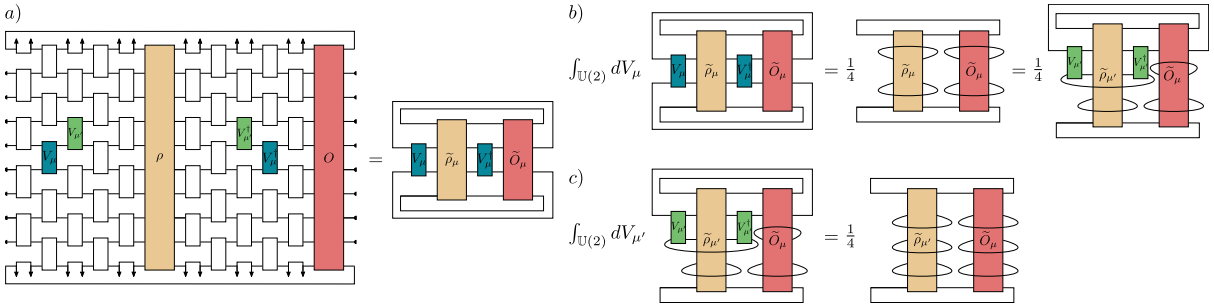

Figure 9: a) Tensor network representation of the function $f(\boldsymbol{\theta}_i) = \mathrm{Tr}[U(\boldsymbol{\theta}_i)\,|\psi\rangle\langle\psi|\,U^\dagger(\boldsymbol{\theta}_i)O]$. In (b) and (c) we show the results obtained by averaging over V_μ and $V_{\mu'}$, assuming that they form 2-designs.

where Tr_μ indicates the partial trace on the qubit over which V_μ acts. In Fig. 9(b) we show the tensor representation of this integral.

Next, we want to integrate the next gate $V_{\mu'}$ such that $U_B^\mu = U_B^{\mu'} V_{\mu'}$ as shown in Fig. 9(c). Here we define $\widetilde{\rho}_{\mu'} = U_B^{\mu'} \rho (U_B^{\mu'})^\dagger$ Using Eq. (129) we find that

$$\int_{\mathbb{U}(2)} dV_\mu \mathrm{Tr}[\mathrm{Tr}_\mu[V_{\mu'}\widetilde{\rho}_{\mu'}(V_{\mu'})^\dagger]\mathrm{Tr}_\mu[\widetilde{O}_\mu]] \tag{145}$$

$$= \frac{1}{4}\mathrm{Tr}[\mathrm{Tr}_{\mu,\mu'}[\widetilde{\rho}_{\mu'}]\mathrm{Tr}_{\mu,\mu'}[\widetilde{O}_\mu]]\,, \tag{146}$$

where Tr_μ indicates the partial trace on the qubit over which V_μ and $V_{\mu'}$ act. In Fig. 9(c) we show the tensor representation of this integral. Iteratively repeating this procedure and integrating over each two-qubit gate in the circuit leads to

$$\mathbb{E}_{\boldsymbol{\theta}_0}[f(\hat{\boldsymbol{\theta}}_i)] \propto \mathrm{Tr}[O]\mathrm{Tr}[|\psi\rangle\langle\psi|] = 0\,, \quad \forall i\,. \tag{147}$$

Here we have used the fact that O is a traceless operator.

□

Lemma 3. *Let $U(\boldsymbol{\theta})$ be a shallow HEA with depth $D \in \mathcal{O}(\log(n))$, and $O = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ with $O_\alpha^{(i)}$ being traceless operators having support on at most two neighboring qubits. Then,*

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\Delta f_{i+1,i}\Delta f_{j+1,j}] = 0\,, \quad \forall i \neq j\,. \tag{148}$$

Proof. First, let us recall that

$$\Delta f_{i+1,i} = f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i)$$
$$= 2\sin\left(\frac{l_i}{2}\right)\partial_i f(\hat{\boldsymbol{\theta}}_i)\,. \tag{149}$$

where we have defined $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_{i-1} + \hat{\boldsymbol{e}}_i \frac{l_i}{2}$ and used the fact we can apply the parameter shift-rule for computing gradients [81, 82]. Then, we have that

$$\mathbb{E}[\Delta f_{i+1,i}\Delta f_{j+1,j}] = \kappa_{i,j}\mathbb{E}[\partial_i f(\hat{\boldsymbol{\theta}}_i)\partial_j f(\hat{\boldsymbol{\theta}}_j)]\,, \tag{150}$$

where $\kappa_{i,j} = 4\sin\left(\frac{l_i}{2}\right)\sin\left(\frac{l_j}{2}\right)$.

Assuming without loss of generality that the gate V_i containing parameter θ_i acts before the gate V_j containing parameter θ_j, and explicitly computing the partial derivatives leads to

$$\partial_i f(\boldsymbol{\theta}_i) = i\mathrm{Tr}[V_B^i \widetilde{\rho}_i (V_B^i)^\dagger X_i]\,. \tag{151}$$

Here, if V_i is the two-qubit gate containing parameter θ_i then we define $U(\boldsymbol{\theta}) = U_A^i V_i U_B^i$ (see Fig. 10(a)), and we denote $\partial_i V_i = -iV_A^i H_i V_B^i$, such that $\partial_i U(\boldsymbol{\theta}) = -iU_A^i V_A^i H_i V_B^i U_B^i$, where U_A^i and U_B^i contain all

other two-qubit gates in the HEA except for V_i. In what follows we will assume that both V_A^i and V_B^i also form 2-designs if V_i forms a 2-design. Then, we defined

$$\widetilde{\rho}_i = U_B^i \left|\psi\right\rangle\!\left\langle\psi\right| (U_B^i)^\dagger \tag{152}$$

$$X_i = [H_i, (V_A^i)^\dagger (U_A^i)^\dagger O U_A^i V_A^i] \,. \tag{153}$$

Then, one can similarly define

$$\partial_j f(\boldsymbol{\theta}_j) = i\mathrm{Tr}[V_A^i V_B^i \widetilde{\rho}_i (V_B^i)^\dagger (V_A^i)^\dagger X_j] \,, \tag{154}$$

where

$$X_j = U_{ij}^\dagger (V_B^j)^\dagger [H_j, (V_A^j)^\dagger (U_A^j)^\dagger O U_A^j (V_A^j)] V_B^j U_{ij} \,. \tag{155}$$

Here we have defined $U_A^i = U_A^j V_j U_{ij}$ (see Fig. 10(b)). That is U_{ij} contains all two-qubit gates in the HEA after V_i but before V_j. Note here that the parametrized gates in $\partial_i f(\hat{\boldsymbol{\theta}}_i)$ and $\partial_j f(\hat{\boldsymbol{\theta}}_j)$ are not necessarily evaluated at the same set of parameters. While this is also true for V_i one can also use the right-invariance of the Haar measure to assume that V_A^i and V_B^i have the same set of parameters in $f(\hat{\boldsymbol{\theta}}_i)$ and $f(\hat{\boldsymbol{\theta}}_j)$ up to some unitary that can be absorbed into U_{ij} and U_B^i, respectively.

From the previous, we want to compute $\mathbb{E}[\partial_i f(\hat{\boldsymbol{\theta}}_i)\partial_j f(\hat{\boldsymbol{\theta}}_j)]$, which results in computing the expectation value $\mathbb{E}[\mathrm{Tr}[V_B^i \widetilde{\rho}_i (V_B^i)^\dagger X_i]\mathrm{Tr}[V_A^i V_B^i \widetilde{\rho}_i (V_B^i)^\dagger (V_A^i)^\dagger X_j]]$. Using the fact that V_i acts non-trivially only on two qubits, we have the following identity (valid for any operators A and B [43])

$$\mathrm{Tr}\left[(\mathbb{1}_{\bar{i}} \otimes V_i) A (\mathbb{1}_{\bar{i}} \otimes V_i^\dagger) B\right] = \sum_{\boldsymbol{p},\boldsymbol{q}} \mathrm{Tr}\left[V_i A_{\boldsymbol{qp}} V_i^\dagger B_{\boldsymbol{pq}}\right] \,, \tag{156}$$

where the summation runs over all bitstrings $\boldsymbol{p}, \boldsymbol{q}$ of length $(n-2)$ and where

$$A_{\boldsymbol{qp}} = \mathrm{Tr}_{\bar{i}}\left[(\left|\boldsymbol{p}\right\rangle\!\left\langle\boldsymbol{q}\right| \otimes \mathbb{1}_4) A\right] \tag{157}$$

$$B_{\boldsymbol{pq}} = \mathrm{Tr}_{\bar{i}}\left[(\left|\boldsymbol{q}\right\rangle\!\left\langle\boldsymbol{p}\right| \otimes \mathbb{1}_4) B\right] \,. \tag{158}$$

Here $\mathbb{1}_2$ denotes the 4×4 identity and $\mathrm{Tr}_{\bar{i}}$ refers to the trace of all qubits except those over which V_i acts on. We refer the reader to [43] for proof of Eq. (156). Then, an explicit integration over V_B^i leads to

$$\int dV_B^i \mathrm{Tr}[V_B^i \widetilde{\rho}_i (V_B^i)^\dagger X_i]\mathrm{Tr}[V_A^i V_B^i \widetilde{\rho}_i (V_B^i)^\dagger (V_A^i)^\dagger X_j] = \frac{1}{15} \sum_{\substack{\boldsymbol{p},\boldsymbol{q} \\ \boldsymbol{p}',\boldsymbol{q}'}} \Delta\Psi_{\boldsymbol{pq}}^{\boldsymbol{p}'\boldsymbol{q}'} \mathrm{Tr}[\Gamma_{\boldsymbol{pq}} \Gamma'_{\boldsymbol{p}'\boldsymbol{q}'}] \,.$$

Here we have defined

$$\Delta\Psi_{\boldsymbol{pq}}^{\boldsymbol{p}'\boldsymbol{q}'} = \mathrm{Tr}[\Psi_{\boldsymbol{pq}} \Psi_{\boldsymbol{p}'\boldsymbol{q}'}] - \frac{\mathrm{Tr}[\Psi_{\boldsymbol{pq}}]\mathrm{Tr}[\Psi_{\boldsymbol{p}'\boldsymbol{q}'}]}{2^m} \,, \tag{159}$$

with

$$\Psi_{\boldsymbol{pq}} = \mathrm{Tr}_{\bar{i}}\left[(\left|\boldsymbol{p}\right\rangle\!\left\langle\boldsymbol{q}\right| \otimes \mathbb{1}_4)\widetilde{\rho}_i\right] \,, \tag{160}$$

$$\Psi_{\boldsymbol{p}'\boldsymbol{q}'} = \mathrm{Tr}_{\bar{i}}\left[(\left|\boldsymbol{p}'\right\rangle\!\left\langle\boldsymbol{q}'\right| \otimes \mathbb{1}_4)\widetilde{\rho}_i\right] \,, \tag{161}$$

$$\Gamma_{\boldsymbol{qp}} = \left[H_i, (V_A^i)^\dagger \Omega_{\boldsymbol{qp}} V_A^i\right] \,, \tag{162}$$

$$\Gamma'_{\boldsymbol{q}'\boldsymbol{p}'} = (V_A^i)^\dagger \Omega'_{\boldsymbol{q}'\boldsymbol{p}'} V_A^i \,, \tag{163}$$

and where

$$\Omega_{\boldsymbol{qp}} = \mathrm{Tr}_{\bar{i}}\left[((\left|\boldsymbol{q}\right\rangle\!\left\langle\boldsymbol{p}\right| \otimes \mathbb{1}_4)(U_A^i)^\dagger O U_A^i\right] \,, \tag{164}$$

$$\Omega'_{\boldsymbol{q}'\boldsymbol{p}'} = \mathrm{Tr}_{\bar{i}}\left[((\left|\boldsymbol{q}'\right\rangle\!\left\langle\boldsymbol{p}'\right| \otimes \mathbb{1}_4) X_j\right] \,. \tag{165}$$

Using Eqs. (162) and (163) we have that

$$\mathrm{Tr}[\Gamma_{\boldsymbol{pq}} \Gamma'_{\boldsymbol{p}'\boldsymbol{q}'}] = \mathrm{Tr}[H_i (V_A^i)^\dagger \Omega_{\boldsymbol{qp}} \Omega'_{\boldsymbol{q}'\boldsymbol{p}'} V_A^i] - \mathrm{Tr}[\Omega_{\boldsymbol{qp}} V_A^i H_i (V_A^i)^\dagger \Omega'_{\boldsymbol{q}'\boldsymbol{p}'}] \,. \tag{166}$$

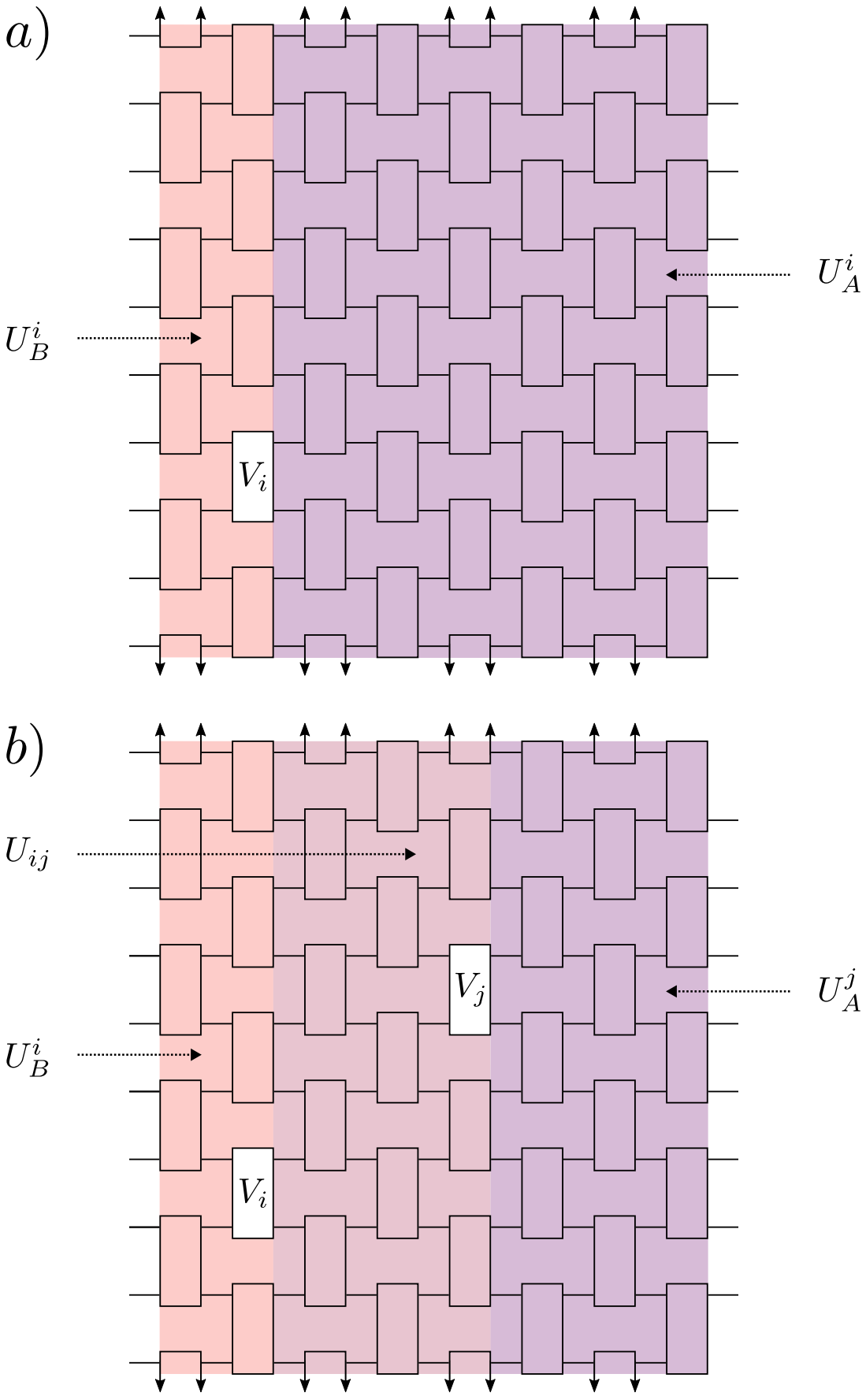

Figure 10: For the proofs, it is useful to divide the HEA into different parts: acting before or after V_i (a), or before, after and in-between V_i and V_j (b).

Then, integrating over V_A^i leads to

$$\int dV_A^i \mathrm{Tr}[\Gamma_{\boldsymbol{pq}}\Gamma'_{\boldsymbol{p'q'}}] = 0\,, \tag{167}$$

as the integration will lead to terms of the form $\mathrm{Tr}[H_i]$ which are equal to zero since H_i are traceless operators.

Combining Eqs. (150), (151), (154) and (167) shows that

$$\mathbb{E}[\Delta f_{i+1,i}\Delta f_{j+1,j}] = 0\,, \quad \forall i \neq j\,. \tag{168}$$

□

Lemma 4. *Let $U(\boldsymbol{\theta})$ be a shallow HEA with depth $D \in \mathcal{O}(\log(n))$ where each local two-qubit gates forms a 2-design on two qubits, and let $O = \sum_i c_i \bigotimes_\alpha O_\alpha^{(i)}$ be the measurement composed of, at most, polynomially many traceless Pauli operators $O_\alpha^{(i)}$ having support on at most two neighboring qubits, and where $\sum_i c_i^2 \in \mathcal{O}(\mathrm{poly}(n))$. Then, if the input state follows an area law of entanglement, we have*

$$\mathrm{Var}[\partial_\mu f(\boldsymbol{\theta})] \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right) \tag{169}$$

Proof. This lemma leverages the results in Theorem 2 of Ref. [43]. Namely, therein it was proved that for a parameter θ_ν in gate V of the HEA

$$\mathrm{Var}[\partial_\nu f(\boldsymbol{\theta})] \geq G_n(D, |\psi\rangle)\,, \tag{170}$$

with

$$G_n(D, |\psi\rangle) = \left(\frac{1}{5}\right)^{2D} \frac{2}{225} \sum_{i \in i_{\mathcal{L}}} \sum_{\substack{(k,k') \in k_{\mathcal{L}_\mathrm{B}} \\ k' \geq k}} c_i^2 \eta(\psi_{k,k'}) \eta(\tilde{O}_i)\,, \tag{171}$$

where $i_{\mathcal{L}}$ is the set of i indices whose associated operators $\tilde{O}_i = \bigotimes_\alpha O_\alpha^{(i)}$ act on qubits in the forward light-cone $\mathcal{L}$ of V, and $k_{\mathcal{L}_\mathrm{B}}$ is the set of k indices whose associated subsystems S_k are in the backward light-cone $\mathcal{L}_\mathrm{B}$ of V. Moreover, we recall that we defined $\eta(M) = \left\| M - \mathrm{Tr}[M]\frac{\mathbb{1}}{d_M} \right\|_2$, as the Hilbert-Schmidt distance between M and $\mathrm{Tr}[M]\frac{\mathbb{1}}{d_M}$, where d_M is the dimension of the matrix M. In addition, $\psi_{k,k'}$ is the partial trace of the input state $|\psi\rangle$ down to the subsystems $S_k S_{k+1} ... S_{k'}$.

Let us note that since $\tilde{O}_i$ is a tensor product of two single-qubit Pauli operators acting on adjacent qubits, then $\eta(\tilde{O}_i) = 4$. Then, if the input state $|\psi\rangle$ follows an area law of entanglement, we have that

$$I_\Lambda(\psi) \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)\,. \tag{172}$$

In particular, noting that $I_\Lambda(\psi)^2 \leq 2^{\Lambda-1}\eta_{\psi_\Lambda}$ [86] and recalling that in our case Λ is the subsystems of qubits $S_k S_{k+1} ... S_{k'}$ containing at most a logarithmic number of qubits, one has that

$$\eta_{\psi_\Lambda} \in \Omega(1/\mathrm{poly}(n)). \tag{173}$$

Finally, since $D \in \mathcal{O}(\log(n))$ we find from Eq. (171) that

$$G_n(D, |\psi\rangle) \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)\,, \tag{174}$$

which in turn implies

$$\mathrm{Var}[\partial_\nu f(\boldsymbol{\theta})] \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)\,. \tag{175}$$

□