# Hardware-efficient ansatz without barren plateaus in any depth

Chae-Yeun Park, Minhyeok Kang, 2, 3, 4 and Joonsuk Huh<sup>1, 2, 3, 4</sup>

<sup>1</sup> Xanadu, Toronto, ON, M5G 2C8, Canada

<sup>2</sup> Department of Chemistry, Sungkyunkwan University, Suwon 16419, Korea

<sup>3</sup> SKKU Advanced Institute of Nanotechnology (SAINT), Sungkyunkwan University, Suwon 16419, Korea

<sup>4</sup> Institute of Quantum Biophysics, Sungkyunkwan University, Suwon 16419, Korea

(Dated: March 11, 2024)

Variational quantum circuits have recently gained much interest due to their relevance in realworld applications, such as combinatorial optimizations, quantum simulations, and modeling a probability distribution. Despite their huge potential, the practical usefulness of those circuits beyond tens of qubits is largely questioned. One of the major problems is the so-called barren plateaus phenomenon. Quantum circuits with a random structure often have a flat cost-function landscape and thus cannot be trained efficiently. In this paper, we propose two novel parameter conditions in which the hardware-efficient ansatz (HEA) is free from barren plateaus for arbitrary circuit depths. In the first condition, the HEA approximates to a time-evolution operator generated by a local Hamiltonian. Utilizing a recent result by [Park and Killoran, Quantum 8, 1239 (2024)], we prove a constant lower bound of gradient magnitudes in any depth both for local and global observables. On the other hand, the HEA is within the many-body localized (MBL) phase in the second parameter condition. We argue that the HEA in this phase has a large gradient component for a local observable using a phenomenological model for the MBL system. By initializing the parameters of the HEA using these conditions, we show that our findings offer better overall performance in solving many-body Hamiltonians. Our results indicate that barren plateaus are not an issue when initial parameters are smartly chosen, and other factors, such as local minima or the expressivity of the circuit, are more crucial.

By combining huge neural networks and parameter optimization techniques, machine learning has achieved great successes in diverse tasks such as image classification [1], defeating human level in playing games [2], natural language processing [3], and predicting protein structures [4]. Inspired by those successes, the same principle has been applied to constructing quantum algorithms. Variational quantum algorithms (VQAs) [5] (including quantum machine learning [6]), which optimize parameters of quantum circuits instead of classical neural networks, have emerged as a new method for solving real-world problems.

Despite their promises, the practical usefulness of the VQAs is largely questioned. One of the main problems is the trainability of quantum circuits. When parameters are randomly sampled, quantum circuits often have flat cost-function landscapes and cannot be efficiently trained. This problem, dubbed barren plateaus [7], is expected to prevail among sufficiently expressive quantum circuit ansätze [8]. Thus, a deep understanding of barren plateaus is essential for devising an efficient variational algorithm.

For this purpose, a number of studies have suggested quantum circuit ansätze without barren plateaus [9–14]. However, less is known about how expressive these ansätze are. Most such circuits are even classically simulable [15], and implementing them on quantum hardware is not straightforward [16–18], either.

An alternative (and intuitive) solution is initializing circuits' parameters to provide large gradients. Indeed, finding good initial parameters is one of the most effective solutions to the vanishing gradient problem in classical

neural networks [19, 20]. Likewise, studies have shown that a quantum model can also have large initial gradients when parameters are initialized smartly [21–28]. However, most of the suggested initialization methods cannot be easily applied to large and deep circuits as they rely on heuristics developed from small circuits [22, 23] or the proven lower bounds of gradient magnitudes are still too small for a deep circuit [25, 26].

In this paper, we propose two novel parameter conditions such that the hardware efficient ansatz (HEA) [29] has large gradients. Since the HEA utilizes a natural entangling gate provided by the hardware, it is the most suitable quantum circuit ansatz for noisy quantum devices [30]. Interestingly, the HEA with a Clifford entangling gate is also preferable to a fault-tolerant quantum computer, as Clifford and single-qubit gates can be implemented with logical qubits rather easily than a parameterized entangling gate [31]. Still, as it is not problemtailored, the HEA is often expected to be more prone to barren plateaus [32, 33]. Existing solutions for suppressing barren plateaus in the HEA indeed have certain limitations. For example, Ref. [25] showed that barren plateaus do not exist for a local observable when parameters are sampled from a Gaussian distribution with a small variance. However, the authors only considered the HEA with the one-dimensional connectivity, and their lower bound of gradient magnitudes still decays exponentially with the number of qubits for a global observable.

In contrast, the parameter regimes we present here give constant gradient magnitudes regardless of circuit depth or the geometry of a circuit. Our first condition is based on Park and Killoran [28], which obtained a condition

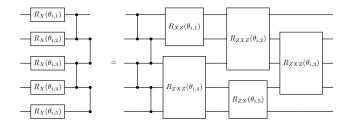


FIG. 1. Circuit identity used for removing CZ gates from the HEA. Using the property that the CZ gate is a Clifford gate, we can move CZ gates in each block to the beginning of the block.

for large gradients when all circuit gates are parameterized. By removing non-parameterized entangling gates from the HEA using circuit identities, we prove that the same condition applies to the HEA with both local and global observables. Our second parameter regime is built upon a recent finding [34] that interprets the HEA to a many-body localized (MBL) system. We utilize a phenomenological theory of MBL systems [35] to show that the HEA in the MBL phase does not have barren plateaus for a local observable. Our second parameter condition may have an advantage over the first one as it allows large initial parameters for the diagonal gates.

Hardware efficient ansatz.—We consider the HEA for a system with N qubits defined on a finite-dimensional lattice. Let us define a vector of all parameters  $\boldsymbol{\theta} = \{\theta_{i,j}\}$ . Then, the output state of the HEA is given by

$$|\psi(\boldsymbol{\theta})\rangle = V(\boldsymbol{\theta}_{p,:}) \cdots V(\boldsymbol{\theta}_{1,:}) |\psi_0\rangle,$$
 (1)

where  $\boldsymbol{\theta}_{i,:} = \{\theta_{i,1}, \dots, \theta_{i,2N}\}$  is a subvector of  $\boldsymbol{\theta}$  and p is a parameter determining the total depth of the circuit. In addition,  $V(\boldsymbol{\theta}_{i::})$  is a unitary operator defined as

$$V(\boldsymbol{\theta}_{i,:}) = \prod_{\langle j,j' \rangle \in E} W_{j,j'} \prod_{j=1}^{N} e^{-iZ_{j}\theta_{i,j+N}/2} \prod_{j=1}^{N} e^{-iX_{j}\theta_{i,j}/2},$$
(2)

where E is the set of edges in the lattice,  $W_{j,j'}$  is a twoqubit gate between sites j and j', and  $\{X_j, Y_j, Z_j\}$  are Pauli operators acting on the j-th qubit. Throughout the paper, we mainly consider  $W = \mathrm{CZ} = \mathrm{diag}(1,1,1,-1)$ , which is a natural entangling gate for major quantum computing platforms [36–41]. Still, our arguments can be extended to the HEA with other mutually commuting Clifford entangling gates.

In the VQAs, a cost function is typically given by  $C(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | O | \psi(\boldsymbol{\theta}) \rangle$  where O is an observable. A parameterized circuit has barren plateaus with respect to a given parameter distribution  $\mathcal{D}$  if  $\mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[(\partial_{ij}C)^2]$  is exponentially small with N for all i,j, where  $\partial_{ij}C := \partial C/\partial \theta_{i,j}$ . For completely random parameters, i.e., each  $\theta_{i,j}$  is sampled from  $\mathcal{U}_{[-\pi,\pi]}$ , the HEA has barren plateaus when one of the following conditions is satisfied: (1) p = poly(N) and O is acting on a constant number of qubits, (2)  $p = \Omega(1)$  and O is acting on  $\Theta(N)$  sites [33].

Large gradients with small parameters.—Assuming that at least one of the gradient components is constant when all parameters are zero, we prove that a parameter constraint exists such that the gradient can be bounded below by a constant.

**Theorem 1.** Let  $C(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | O | \psi(\boldsymbol{\theta}) \rangle$  be the cost function where O is either a Pauli string or k-local Hamiltonian. Suppose that there exist n, m such that  $|\partial_{n,m}C|_{\boldsymbol{\theta}=0} = \Omega(1)$ . Then, there exists a constant  $\gamma > 0$  such that  $|\partial_{n,m}C| = \Omega(1)$  when  $0 \le \theta_{i,j} \le \gamma/(pN)$  is satisfied for all i and j.

See Appendix A for a proof. Theorem 1 is based on Ref. [28], which proved the existence of a parameter condition such that a circuit has a constant gradient component when all gates are parameterized. We use a circuit identity given in Fig. 1 to translate the HEA into a circuit without non-parameterized entangling gates. The circuit identity enables us to move the CZ layer in the 2i-th block to the beginning of the block, which cancels the CZ layer in the 2i-1-th block. The resulting circuit only has parameterized gates (for even p) generated by, at most, k-local operators (acting on at most k nearby sites), where k is determined by the connectivity of the original circuit. This procedure recovers a setup used in Ref. [28]. The same argument also works for odd p, which remains a single layer of CZ gates acting on the initial state (see Appendix A). In addition, we note that one can easily find a product state  $|\psi_0\rangle$  satisfying  $|\partial_{n,m}C|_{\boldsymbol{\theta}=0}=\Omega(1)$  when O is a Pauli string.

Theorem 1 can be compared to the main result of Ref. [25], which proved that the magnitudes of the gradient could be lower bounded by  $\Theta[(pS)^{-S}]$  when the parameters are drawn from  $\mathcal{N}(0, [1/\sqrt{4pS}]^2)$  (in our notation). Here, S is the weight of the observable O, which counts the number of qubits that O acts non-trivially (e.g.,  $X_1Z_5Z_7$  has S=3). Three major differences are listed as follows. First, the parameters are smaller for Theorem 1,  $\Theta[1/(pN)]$  versus  $\Theta(1/\sqrt{pN})$ , but the lower bound is much bigger,  $\Theta(1)$  versus  $\Theta[(pS)^{-S}]$ . Second, Ref. [25] only considered the 1D HEA, whereas our theorem applies to any finite-dimensional lattices, including the heavy-hexagon lattice upon which IBM's recent quantum processors are implemented [42]. Finally, Theorem 1 allows additional gates applied to the initial state, such as data-encoding gates widely used in quantum machine learning setup, as long as the circuit has large gradients when all trainable parameters are zero (see Appendix E).

Floquet-MBL initialization for a large gradient.— One of the potential limitations of the previous parameter condition is that the parameters are too small when applied to a circuit with a large depth. In this case, the parameters between each instance are nearly the same, and an advantage of the randomness in initial parameters [43] may be lost.

Our second parameter condition overcomes this problem by allowing the parameters for the RZ gates to be random in  $\mathcal{U}_{[-\pi,\pi]}$ . Formally, the parameter condition is written as follows:

$$\vartheta_i = \theta_{i,j} \text{ for all } 1 \le j \le N \text{ and } 0 \le \vartheta_i \le \vartheta_c \text{ for all } i, 
\theta_{i,j} \sim \mathcal{U}_{[-\pi,\pi]} \text{ for all } N+1 \le j \le 2N,$$
(3)

where  $\vartheta_c$  is the critical point between the chaotic and the MBL phases. For the 1D HEA, we find  $0.13 \lesssim \vartheta_c \lesssim 0.16$  (see Appendix B).

Our circuit is in the MBL phase when the parameters satisfy this condition. A phenomenological theory of the MBL [35] suggests that one can find a Hamiltonian  $H_{\rm MBL}$  such that

$$V(\boldsymbol{\theta}_{k,:}) \cdots V(\boldsymbol{\theta}_{1,:}) = e^{-iH_{\text{MBL}}kT}, \tag{4}$$

for any  $k \geq 1$  and a constant T, where  $H_{\text{MBL}}$  can be written as

$$H_{\text{MBL}} = \sum_{i=1}^{N} J_i \tau_i^z + \sum_{i \neq j} J_{ij} \tau_i^z \tau_j^z + \sum_{\text{all distinct } i, j, k} J_{ijk} \tau_i^z \tau_j^z \tau_k^z + \cdots$$
 (5)

Here,  $\tau_i^z$  is a local integral of motion, which has a finite overlap with  $Z_i$ .

Let us consider the gradient component for  $\theta_{p,1}$  when  $O = Y_1$  and  $|\psi_0\rangle = |0^N\rangle$  is used. From the definition of the cost function, we obtain

$$\frac{\partial C}{\partial \theta_{p,1}} = \frac{i}{2} \langle 0^N | U_{[1:p-1]}^{\dagger} [X_1, \widetilde{Y}_1] U_{[1:p-1]} | 0^N \rangle, \quad (6)$$

where  $U_{[1:p-1]} = V(\boldsymbol{\theta}_{p-1,:}) \cdots V(\boldsymbol{\theta}_{1,:})$  is a subcircuit of the HEA and  $\widetilde{Y}_1 := V(\boldsymbol{\theta}_{p,:})^{\dagger} Y_1 V(\boldsymbol{\theta}_{p,:})$ . After some steps, the following expression is obtained:

$$[X_1, \widetilde{Y_1}] = 2i\cos(\theta_{p,N+1})[\cos(\theta_p)Z_1 + \sin(\theta_p)Y_1] \times \prod_{j \in \mathcal{N}(1)} [\cos(\theta_p)Z_i + \sin(\theta_p)Y_i], \tag{7}$$

where  $\mathcal{N}(i)$  is a set of all neighbors of i in a given lattice (see Appendix C for details).

In summary, the gradient is expressed as the sum of multi-point correlation functions. From the fact that  $Z_i$  has a finite overlap with  $\tau_z^i$ , and the correlation functions involving Pauli-Y operators are relatively small in the MBL systems [44], we obtain

$$\frac{\partial C}{\partial \theta_{p,N+1}} \approx -\cos(\theta_{p,N+1}) \left[ A^2 \cos(\vartheta_p) \right]^{1+|\mathcal{N}(1)|} \tag{8}$$

for sufficiently large p (see Appendix C for details). Here,  $A = \text{Tr}[\tau_z^i Z_i]/2^N$  quantifies the overlap between two operators and is independent of N and p. Hence, we obtain  $\mathbb{E}_{\boldsymbol{\theta}}[(\partial C/\partial \theta_{p,N+1})^2] \approx [A^2 \cos(\vartheta_p)]^{1+|\mathcal{N}(1)|}/2$ . This implies that the HEA in the MBL phase does not have barren plateaus in any depth for this observable. One

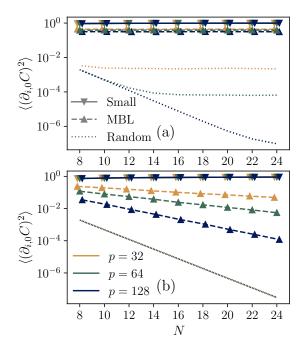


FIG. 2. Averaged squared gradients as functions of N for  $p \in [32,64,128]$ . Observables (a)  $O = Y_1$  and (b)  $O = Y_1 \prod_{j=2}^N Z_j$  are used. Each data point presents the averaged gradient components for the RX gate acting on the first qubit,  $\sum_{j=1}^n (\partial_{i,0}C)^2/p$ . For each parameter initialization scheme, results are averaged over  $2^{10}$  randomly sampled parameters. For the Small initialization, the gradient magnitudes do not decay with N regardless of the observable. On the other hand, the MBL initialization shows  $\Theta(1)$  gradient magnitudes when a local observable is used, whereas they decay exponentially for a global observable.

can also repeat the same calculation for other observables. For a global observable  $O = Y_1 \prod_{j=2}^N Z_j$ , we obtain  $\mathbb{E}_{\boldsymbol{\theta}}[(\partial C/\partial \theta_{p,N+1})^2] \approx \left[A^2 \cos(\vartheta_p)\right]^{N-|\mathcal{N}(1)|}/2$ , which decays exponentially with N (see Appendix C for details).

There is a subtlety in applying our argument here to the HEA in a higher dimensional lattice. Recent studies have claimed that the MBL phase does not exist in the thermodynamic limit when the dimension is larger than one (see, e.g., Ref. [45]). However, as one can still observe a signature of the MBL transition for a finite-size system [46], we expect that our MBL parameter condition would work even in a higher dimensional HEA for system sizes tractable to intermediate-scale quantum computers.

Numerical simulations.— We numerically compare the magnitudes of the gradients when parameters are randomly sampled from the following distributions. (1) **Small**: All parameters are drawn from  $\mathcal{U}_{[0,\pi/(pN)]}$ , (2) **MBL**: Parameters follow Eq. (3) with  $\vartheta_i \in \mathcal{U}_{[0,0.1]}$ , and (3) **Random**: All parameters are completely random, i.e.,  $\theta_{i,j} \sim \mathcal{U}_{[0,2\pi]}$  for all i,j.

We use two observables  $O = Y_1$  and  $O = Y_1 \prod_{j=2}^{N} Z_j$ , and the initial state given by  $|\psi_0\rangle = |0^N\rangle$ . Simple compu-

tation gives that  $\partial_{i,0}C|_{\pmb{\theta}=0}=-1$  for both the observables regardless of i. From Theorem 1, we expect that our first parameter scheme (Small) could give large gradients [47]. On the other hand, the MBL parameter scheme will give a constant magnitude of  $\partial_{p,0}C$  for  $O=Y_1$ , whereas the magnitude would decay exponentially with N for sufficiently large p when  $O=Y_1\prod_{j=1}^N Z_j$  is used. The scaling behaviors of gradients for  $O=Y_1$  from

The scaling behaviors of gradients for  $O=Y_1$  from these schemes are plotted in Fig. 2(a). We observe that gradient magnitudes do not decay with N when parameters are initialized following the small or MBL scheme, which is consistent with our theoretical investigations. On the other hand, when parameters are completely random, gradient magnitudes decay exponentially with N for small N, and they saturate after  $N=N_0(p)$ . This is consistent with previous observations in Refs. [7, 28, 33].

We also plot results for  $O = Y_1 \prod_{j=1}^N Z_j$  in Fig. 2(b), which shows that the magnitudes of the gradients are constant for the small parameter scheme, whereas they decay exponentially with N for the MBL distribution. These also agree with our theoretical expectations. In addition, the random parameter scheme gives exponential decay of the gradient magnitudes regardless of p. This is also consistent with Ref. [33], which proved that barren plateaus appear in  $\Omega(1)$  depth for a global observable. Lastly, while both the MBL and random parameter schemes show exponentially decaying gradients, the decaying rate is much lower for the MBL scheme. This suggests that the MBL parameter scheme may still give practical advantages even for a global observable when p and N are not too large.

In Appendix D, we present additional numerical results for the 2D HEA. We also compare our parameter conditions and the random Gaussian initialization suggested in Ref. [14].

Solving quantum many-body Hamiltonians.— We now solve the ground state problem of two Hamiltonians by simulating variational quantum eigensolvers (VQEs). We consider the one-dimensional Heisenberg and cluster models with external fields given by

$$H_{1} = \sum_{i=1}^{N-1} \left[ X_{i} X_{i+1} + Y_{i} Y_{i+1} + Z_{i} Z_{i+1} \right] + h_{1} \sum_{i=1}^{N} Z_{i}$$
 (9)  

$$H_{2} = -\sum_{i=2}^{N-2} Z_{i-1} X_{i} Z_{i+1} - X_{1} Z_{2} - Z_{N-1} X_{N} - h_{2} \sum_{i=1}^{N} Z_{i},$$
 (10)

with the strength of external fields  $h_1 = h_2 = 1$ .

We use the cost functions given by the expectation value of the Hamiltonians for the output states of circuits (i.e.,  $C = \langle H_{1,2} \rangle$ ) and optimize the parameters using Adam [48] with exactly computed gradients. We choose  $|\psi_0\rangle = |y;+\rangle^{\otimes N}$  as the initial state of the circuit since the gradient components for RX gates are  $\Omega(1)$  when all parameters are zero with this choice, i.e.,  $|\partial_{n,m}C|_{\boldsymbol{\theta}=0} = \Omega(1)$  for all  $n \in \{1, \cdots, N\}$  and  $1 \leq m \leq N$ .

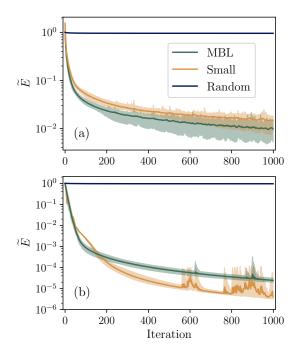


FIG. 3. Normalized energies  $\widetilde{E}=(\langle H_{1,2}\rangle-E_{\rm GS})/|E_{\rm GS}|$  as functions of optimization steps for (a) the Heisenberg model  $(H_1)$  and (b) the cluster model  $(H_2)$  with external fields. The HEA with N=20 and p=256 is used. We optimize the parameters using Adam [48] with learning rates (a)  $\eta=0.005$  and (b)  $\eta=0.001$ , which are chosen from hyperparameter optimizations. For each initialization scheme, we run 16 independent VQE instances. Solid curves show the averaged values for each step, while the shaded regions indicate the range between the worst and best-performing instances.

For the 1D HEA with N=20 and p=256, the learning curves are plotted in Fig. 3. We choose such relatively large p to ensure that gradients are sufficiently small when parameters are completely random. The results show that the circuits initialized following our parameter schemes, Small and MBL, provide much better convergence than Random. However, the best initialization methods depend on the Hamiltonians: The MBL initialization outperforms the Small initialization for  $H_1$ , but the opposite holds for  $H_2$ . We also observe that the HEA finds the nearly exact ground state for  $H_2$  (with  $\widetilde{E} \approx 3 \times 10^{-6}$ ), whereas the results for  $H_1$  are relatively poor. These differences are from the expressivity of the HEA for the target problems. The ground state of  $H_1$ lies within the subspace  $J_z = 0$  where  $J_z = \sum_{i=1}^{N} Z_i$  is the total spin operator. However, the HEA cannot capture this symmetry, and the output state always overlaps with subspaces with other  $J_z$  values. On the other hand, the HEA is a natural ansatz [49–51] for  $H_2$ , which is from our circuit identity (Fig. 1). These results indicate that our parameter initialization sufficiently avoids barren plateaus, and the expressivity of the circuit or local minima can be more critical problems determining the performance of quantum variational algorithms.

In Appendix E, we present additional results solving a machine-learning problem using the HEA. We show that our parameter initialization schemes, Small and MBL, offer better performance also for a binary classification task, a typical supervised learning problem.

Conclusion and outlook.—We found two novel parameter conditions where the hardware-efficient ansatz (HEA) does not have barren plateaus. In contrast to other known conditions, our ones provide a constant gradient regardless of circuit depth.

In addition to the practical advantage of initializing the HEA, our second parameter condition can be a counterexample to a recent claim [15] that all barren plateaus free ansätze are classically simulable [52]. This is because there is no known classically efficient algorithm for simulating the MLB system for an exponentially long time, and our argument based on the phenomenological theory of the MBL systems [35] guarantees a gradient component with a constant magnitude even at an exponentially long time. Still, rigorous proof of this argument should be addressed in future work as it requires careful complexity analysis.

Our MBL parameter condition also raises an interesting question on the role of entanglement in barren plateaus. It is often assumed that entanglement volume-law implies the onset of barren plateaus [24, 53, 54]. However, our results suggest that it is not always the case as a long-time-evolved state in the MBL system follows the volume-law of entanglement [55], whereas our argument

guarantees a constant magnitude gradient component at any time.

Note added.— In the days prior to the submission of this manuscript, a preprint [56] proposing an initialization scheme for the HEA that provides  $\Theta(1)$  gradient norm is uploaded. However, their bound is only proved for the HEA with the one-dimensional connectivity.

Acknowledgements.—CYP thanks Kyunghyun Baek for the initial discussion and Joseph Bowles, Nathan Killoran, Korbinian Kottmann, and Maria Schuld for helpful comments. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award NERSC DDR-ERCAP0025705. This work was partly supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (NRF-2022M3H3A106307411, NRF-2023M3K5A1094805, and NRF-2023M3K5A109481311) and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government(MSIP) (No. 2019-0-00003, Research and Development of Core technologies for Programming, Running, Implementing and Validating of Fault-Tolerant Quantum Computing System). JH acknowledges Xanadu for hosting his sabbatical year visit. Numerical simulations were performed using PennyLane [57] software package with LIGHTNING [58] plugin.

- A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60, 84 (2017).
- [2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529, 484 (2016).
- [3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in Advances in neural information processing systems, Vol. 33 (2020) pp. 1877–1901.
- [4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature 596, 583 (2021).
- [5] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, et al., Variational quantum algorithms, Nat. Rev. Phys. 3, 625 (2021).
- [6] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemporary Physics 56, 172 (2015).
- [7] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Bab-

- bush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. 9, 1 (2018).
- [8] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum 3, 010313 (2022).
- [9] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, Phys. Rev. X 11, 041011 (2021).
- [10] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, Diagnosing barren plateaus with tools from quantum optimal control, Quantum 6, 824 (2022).
- [11] Z. Liu, L.-W. Yu, L.-M. Duan, and D.-L. Deng, Presence and absence of barren plateaus in tensor-network based machine learning, Phys. Rev. Lett. 129, 270501 (2022).
- [12] E. C. Martín, K. Plekhanov, and M. Lubasch, Barren plateaus in quantum tensor network optimization, Quantum 7, 974 (2023).
- [13] T. Barthel and Q. Miao, Absence of barren plateaus and scaling of gradients in the energy optimization of isometric tensor network states, arXiv preprint arXiv:2304.00161 (2023).
- [14] H.-K. Zhang, S. Liu, and S.-X. Zhang, Absence of barren plateaus in finite local-depth circuits with long-range entanglement, arXiv preprint arXiv:2311.01393 (2023).

- [15] M. Cerezo, M. Larocca, D. García-Martín, N. Diaz, P. Braccia, E. Fontana, M. S. Rudolph, P. Bermejo, A. Ijaz, S. Thanasilp, et al., Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing, arXiv preprint arXiv:2312.09121 (2023).
- [16] A. Skolik, M. Cattelan, S. Yarkoni, T. Bäck, and V. Dunjko, Equivariant quantum circuits for learning on weighted graphs, npj Quantum Information 9, 47 (2023).
- [17] R. D. East, G. Alonso-Linaje, and C.-Y. Park, All you need is spin: Su (2) equivariant variational quantum circuits based on spin networks, arXiv preprint arXiv:2309.07250 (2023).
- [18] L. Schatzki, M. Larocca, F. Sauvage, and M. Cerezo, Theoretical guarantees for permutation-equivariant quantum neural networks, npj Quantum Information 10, 12 (2023).
- [19] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings* of the thirteenth international conference on artificial intelligence and statistics (JMLR Workshop and Conference Proceedings, 2010) pp. 249–256.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE interna*tional conference on computer vision (2015) pp. 1026– 1034.
- [21] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, Quantum 3, 214 (2019).
- [22] N. Jain, B. Coyle, E. Kashefi, and N. Kumar, Graph neural network initialisation of quantum approximate optimisation, Quantum 6, 861 (2022).
- [23] A. A. Mele, G. B. Mbeng, G. E. Santoro, M. Collura, and P. Torta, Avoiding barren plateaus via transferability of smooth solutions in a Hamiltonian variational ansatz, Phys. Rev. A 106, L060401 (2022).
- [24] S. H. Sack, R. A. Medina, A. A. Michailidis, R. Kueng, and M. Serbyn, Avoiding barren plateaus using classical shadows, PRX Quantum 3, 020365 (2022).
- [25] K. Zhang, L. Liu, M.-H. Hsieh, and D. Tao, Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits, in *Advances in Neu*ral Information Processing Systems, Vol. 35 (2022) pp. 18612–18627.
- [26] Y. Wang, B. Qi, C. Ferrie, and D. Dong, Trainability enhancement of parameterized quantum circuits via reduced-domain parameter initialization, arXiv preprint arXiv:2302.06858 (2023).
- [27] M. S. Rudolph, J. Miller, D. Motlagh, J. Chen, A. Acharya, and A. Perdomo-Ortiz, Synergistic pretraining of parametrized quantum circuits via tensor networks, Nature Communications 14, 8367 (2023).
- [28] C.-Y. Park and N. Killoran, Hamiltonian variational ansatz without barren plateaus, Quantum 8, 1239 (2024).
- [29] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. 5, 1 (2014).
- [30] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum 2, 79 (2018).
- [31] K. Fujii, Quantum Computation with Topological Codes: from qubit to topological fault-tolerance, Vol. 8 (Springer,

- 2015).
- [32] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the Hamiltonian variational ansatz, PRX Quantum 1, 020319 (2020).
- [33] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. 12, 1 (2021).
- [34] O. Shtanko, D. S. Wang, H. Zhang, N. Harle, A. Seif, R. Movassagh, and Z. Minev, Uncovering local integrability in quantum many-body dynamics, arXiv preprint arXiv:2307.07552 (2023).
- [35] D. A. Huse, R. Nandkishore, and V. Oganesyan, Phenomenology of fully many-body-localized systems, Phys. Rev. B 90, 174202 (2014).
- [36] J. I. Cirac and P. Zoller, Quantum computations with cold trapped ions, Phys. Rev. Lett. 74, 4091 (1995).
- [37] D. Jaksch, J. I. Cirac, P. Zoller, S. L. Rolston, R. Côté, and M. D. Lukin, Fast quantum gates for neutral atoms, Phys. Rev. Lett. 85, 2208 (2000).
- [38] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, et al., Simple allmicrowave entangling gate for fixed-frequency superconducting qubits, Phys. Rev. Lett. 107, 080502 (2011).
- [39] C. Figgatt, A. Ostrander, N. M. Linke, K. A. Landsman, D. Zhu, D. Maslov, and C. Monroe, Parallel entangling operations on a universal ion-trap quantum computer, Nature 572, 368 (2019).
- [40] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, et al., Evidence for the utility of quantum computing before fault tolerance, Nature 618, 500 (2023).
- [41] S. J. Evered, D. Bluvstein, M. Kalinowski, S. Ebadi, T. Manovitz, H. Zhou, S. H. Li, A. A. Geim, T. T. Wang, N. Maskara, H. Levine, G. Semeghini, M. Greiner, V. Vuletić, and M. D. Lukin, High-fidelity parallel entangling gates on a neutral-atom quantum computer, Nature 622, 268–272 (2023).
- [42] The IBM Quantum heavy hex lattice, https://research.ibm.com/blog/heavy-hex-lattice.
- [43] L. F. Wessels and E. Barnard, Avoiding false local minima by proper initialization of connections, IEEE transactions on neural networks 3, 899 (1992).
- [44] M. Serbyn, Z. Papić, and D. A. Abanin, Quantum quenches in the many-body localized phase, Phys. Rev. B 90, 174302 (2014).
- [45] T. Thiery, F. Huveneers, M. Müller, and W. De Roeck, Many-body delocalization as a quantum avalanche, Phys. Rev. Lett. 121, 140601 (2018).
- [46] T. B. Wahl, A. Pal, and S. H. Simon, Signatures of the many-body localized regime in two dimensions, Nat. Phys. 15, 164 (2019).
- [47] The value of  $\gamma$  obtained from the proof of Theorem 1 can be smaller than  $\pi$ . However, since many inequalities we have used for the proof are not tight, we regard Theorem 1 as a rough estimation of the order of such a parameter condition. The exact scaling of the parameter condition can be further investigated numerically. See also Appendix D.
- [48] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in 3rd International Conference on Learn-

- ing Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015).
- [49] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A 92, 042303 (2015).
- [50] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms 12, 34 (2019).
- [51] C.-Y. Park, Efficient ground state preparation in variational quantum eigensolver with symmetry breaking layers, APL Quantum 1, 016101 (2024).
- [52] In contrast, the HEA within the first parameter condition is classically simulable up to an inverse polynomial additive error using algorithms developed in Refs. [59, 60]. However, we expect that one can extend Theorem 1 to an arbitrary graph (instead of a *D*-dimensional lattice), which makes those algorithms cannot be applied.
- [53] C. O. Marrero, M. Kieferová, and N. Wiebe, Entanglement-induced barren plateaus, PRX Quantum 2, 040316 (2021).
- [54] L. Leone, S. F. Oliviero, L. Cincio, and M. Cerezo, On the practical usefulness of the hardware efficient ansatz, arXiv preprint arXiv:2211.01477 (2022).
- [55] J. H. Bardarson, F. Pollmann, and J. E. Moore, Unbounded growth of entanglement in models of many-body localization, Phys. Rev. Lett. 109, 017202 (2012).
- [56] X. Shi and Y. Shang, Avoiding barren plateaus via gaussian mixture model, arXiv preprint arXiv:2402.13501 (2024).
- [57] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, et al., Pennylane: Automatic differentiation of hybrid quantumclassical computations (2018), arXiv:1811.04968 [quantph].
- [58] A. Asadi, A. Dusko, C.-Y. Park, V. Michaud-Rioux, I. Schoch, S. Shu, T. Vincent, and L. J. O'Riordan, Hybrid quantum programming with pennylane lightning on HPC platforms (2024), arXiv:2403.02512 [quant-ph].
- [59] S. Bravyi, D. Gosset, and R. Movassagh, Classical algorithms for quantum mean values, Nat. Phys. 17, 337 (2021).
- [60] N. J. Coble and M. Coudron, Quasi-polynomial time approximation of output probabilities of geometricallylocal, shallow quantum circuits, in 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science

- (FOCS) (IEEE, 2022) pp. 598-609.
- [61] T. Kuwahara, T. Mori, and K. Saito, Floquet–Magnus theory and generic transient dynamics in periodically driven many-body quantum systems, Annals of Physics 367, 96 (2016).
- [62] J. Eisert, M. Friesdorf, and C. Gogolin, Quantum manybody systems out of equilibrium, Nat. Phys. 11, 124 (2015).
- [63] C. Gogolin and J. Eisert, Equilibration, thermalisation, and the emergence of statistical mechanics in closed quantum systems, Rep. Prog. Phys 79, 056001 (2016).
- [64] R. Nandkishore and D. A. Huse, Many-body localization and thermalization in quantum statistical mechanics, Annu. Rev. Condens. Matter Phys. 6, 15 (2015).
- [65] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics, Adv. Phys. 65, 239 (2016).
- [66] B. L. Altshuler, A. G. Aronov, and P. Lee, Interaction effects in disordered fermi systems in two dimensions, Phys. Rev. Lett. 44, 1288 (1980).
- [67] F. Alet and N. Laflorencie, Many-body localization: An introduction and selected topics, Comptes Rendus Physique 19, 498 (2018).
- [68] M. Žnidarič, T. Prosen, and P. Prelovšek, Many-body localization in the Heisenberg XXZ magnet in a random field, Phys. Rev. B 77, 064426 (2008).
- [69] I. H. Kim, A. Chandran, and D. A. Abanin, Local integrals of motion and the logarithmic lightcone in many-body localized systems, arXiv preprint arXiv:1412.3073 (2014).
- [70] P. Ponte, Z. Papić, F. Huveneers, and D. A. Abanin, Many-body localization in periodically driven systems, Phys. Rev. Lett. 114, 140401 (2015).
- [71] L. Zhang, V. Khemani, and D. A. Huse, A floquet model for the many-body localization transition, Phys. Rev. B 94, 224202 (2016).
- [72] J. A. Kjäll, J. H. Bardarson, and F. Pollmann, Manybody localization in a disordered quantum ising chain, Phys. Rev. Lett. 113, 107204 (2014).
- [73] T. Cubitt and A. Montanaro, Complexity classification of local hamiltonian problems, SIAM Journal on Computing 45, 268 (2016).
- [74] M. Schuld and F. Petruccione, Supervised learning with quantum computers, Vol. 17 (Springer, 2018).

### Appendix A: Parameter constraint for lower bounding the gradient magnitudes by a constant

This section presents a proof of Theorem 1 in the main text. Under the assumption that the hardware-efficient ansatz (HEA) has a gradient component whose magnitude is constant when all parameters are zero, Theorem 1 states that the circuit still has a gradient component with a constant magnitude when parameters satisfy a certain condition.

#### 1. Constant gradient magnitudes for the Hamiltonian variational ansatz

In this subsection, we generalize the main results of Ref. [28], which proved that a parameter condition such that the Hamiltonian variational ansatz (HVA) has large gradients exists. The main difference is that we consider a circuit whose gates are generated by local operators instead of local Hamiltonians as in Ref. [28]. Still, if we restrict operators in each layer to commute mutually, we can group them together and make a local Hamiltonian. Using this process, we interpret the resulting circuit as the HVA and prove the same bound following Ref. [28]. In addition, our proof

here also works for a global observable, in contrast to Ref. [28], which only considered a local observable. Let us consider a parameterized quantum circuit for a system with N qubits, given by

$$|\psi(\boldsymbol{\theta})\rangle = \prod_{i=1}^{D} \left[ \prod_{j=1}^{M_i} e^{-iG^{(i,j)}\theta_{i,j}} \right] |\psi_0\rangle, \qquad (A.1)$$

where generators  $G^{(i,j)}$  for each i are mutually commuting, i.e.,  $[G^{(i,j)}, G^{(i,j')}] = 0$  for all i, j, j'. We also assume that each  $G^{(i,j)}$  is a local operator, acting on at most  $\Theta(1)$  sites and  $M_i = \mathcal{O}(N)$ . Here, two notations  $\prod$  and  $\prod$  are defined as

$$\prod_{i=1}^{k} U_i := U_1 \cdots U_k, \qquad \prod_{i=1}^{k} U_i = U_k \cdots U_1. \tag{A.2}$$

The circuit given in Eq. A.1 can be considered a generalized version of the HVA.

We now consider the cost function is given by

$$C = \langle \psi(\boldsymbol{\theta}) | O | \psi(\boldsymbol{\theta}) \rangle. \tag{A.3}$$

A component of the gradient for C is obtained as

$$\begin{split} \partial_{n,m}C &:= \frac{\partial C}{\partial \theta_{n,m}} \\ &= \Big\langle \psi_0 \Big| \prod_{i=1}^n \big[ \prod_{j=1}^{M_i} e^{iG^{(i,j)}\theta_{i,j}} \big] (iG^{(n,m)}) \prod_{i=n}^D \big[ \prod_{j=1}^{M_i} e^{iG^{(i,j)}\theta_{i,j}} \big] O \bigoplus_{i=1}^{D} \big[ \prod_{j=1}^{M_i} e^{-iG^{(i,j)}\theta_{i,j}} \big] |\psi_0 \Big\rangle \\ &+ \Big\langle \psi_0 \Big| \prod_{i=1}^D \big[ \prod_{j=1}^{M_i} e^{iG^{(i,j)}\theta_{i,j}} \big] O \bigoplus_{i=n}^{D} \big[ \prod_{j=1}^{M_i} e^{-iG^{(i,j)}\theta_{i,j}} \big] (-iG^{(n,m)}) \bigoplus_{i=1}^{D} \big[ \prod_{j=1}^{M_i} e^{-iG^{(i,j)}\theta_{i,j}} \big] \Big| \psi_0 \Big\rangle \\ &= i \, \langle \psi_0 | U_B^{\dagger} [G_{n,m}, U_A^{\dagger} O U_A] U_B |\psi_0 \rangle \,, \end{split}$$

where

$$U_{A} = \prod_{i=n}^{D} \left[ \prod_{j=1}^{M_{i}} e^{-iG^{(i,j)}\theta_{i,j}} \right], \qquad U_{B} = \prod_{i=1}^{n} \left[ \prod_{j=1}^{M_{i}} e^{-iG^{(i,j)}\theta_{i,j}} \right]. \tag{A.4}$$

Under this setup, we have the following lemma.

**Lemma A.1.** For  $\rho_0 = |\psi_0\rangle \langle \psi_0|$ , let us assume that  $|\operatorname{Tr}[\rho_0[G^{(n,m)},O]]| > 0$ , and there exist Hamiltonians  $H_A$ ,  $H_B$  such that  $U_A = e^{-iH_At_A}$  and  $U_B = e^{-iH_Bt_B}$  for some  $t_A, t_B \geq 0$ . Then,

$$|\partial_{n,m}C| \ge |\operatorname{Tr}[\rho_0[G^{(n,m)}, O]]|/2$$
 (A.5)

for  $t_A + t_B \le t_c := |\text{Tr}[\rho_0[G^{(n,m)}, O]]|/(4KQ)$ , where  $K = \max\{\|H_B\|, \|[H_A, O]\|\}$  and  $Q = \max\{\|[G^{(n,m)}, O]\|, \|G^{(n,m)}\|\}$ . Here,  $\|\cdot\|$  is the operator norm.

Proof. Let

$$A(t_1, t_2) = i \operatorname{Tr}[e^{-iH_B t_2} \rho_0 e^{iH_B t_2} [G^{(n,m)}, e^{iH_A t_1} O e^{-iH_A t_1}]].$$
(A.6)

One can see that  $A(0,0) = i \operatorname{Tr}[\rho_0[G^{(n,m)},O]]$  and  $A(t_A,t_B) = \partial_{n,m}C$ . Then,

$$|A(t_A, t_B) - A(0, 0)| \le \int_0^{t_A} dt_1 \left| \frac{\partial A(t_1, t_B)}{\partial t_1} \right| + \int_0^{t_B} dt_2 \left| \frac{\partial A(0, t_2)}{\partial t_2} \right|. \tag{A.7}$$

We further have

$$\left| \frac{dA(0, t_2)}{\partial t_2} \right| = \left| \text{Tr} \left\{ [H_B, \rho_0(t_2)] [G^{(n,m)}, O] \right\} \right| 
\leq 2 \|H_B\| \|[G^{(n,m)}, O]\| \leq 2KQ,$$
(A.8)

and

$$\left| \frac{dA(t_1, t_B)}{\partial t_1} \right| = \left| \text{Tr} \left\{ \rho_0(t_B) [G^{(n,m)}, [H_A, e^{iH_A t_1} O e^{-iH_A t_1}]] \right\} \right| 
\leq 2 \|G^{(n,m)}\| \|[H_A, O]\| \leq 2KQ,$$
(A.9)

where  $\rho_0(t) = e^{-iH_B t} \rho_0 e^{iH_B t}$ .

Integrating both sides, we have

$$|A(t_R, t_L) - A(0, 0)| \le 2KQ(t_R + t_L). \tag{A.10}$$

By entering  $t_R + t_L \le t_c = |A(0,0)|/(4KQ)$ , we obtain the desired inequality.

Note that we used different definitions of K and Q from Ref. [28] to incorporate that  $G^{(n,m)}$  is a local operator (instead of a sum of local operators considered in Ref. [28]).

Although exact values of K and Q depend on the definitions of  $H_A$ ,  $H_B$ , and O, the scaling of K and Q can be obtained under the reasonable assumptions that  $H_A$ ,  $H_B$  are local Hamiltonians (sums of operators acting on at most  $\mathcal{O}(1)$  nearby sites in a given lattice) and  $G^{(n,m)}$  is a local operator. In this case, one can readily see that  $K = \Theta(N)$ ,  $Q = \Theta(1)$  both for O given by (1) a Pauli string and (2) a local Hamiltonian.

Next, we find those Hamiltonians  $H_A$  and  $H_B$  for the parameterized circuit given by Eq. (A.1). For this purpose, we rewrite each layer of the circuit. We introduce a Hamiltonians  $H^{(i)}$  defined as

$$H^{(i)} := \sum_{j=1}^{M_i} G^{(i,j)} \frac{\theta_{i,j}}{\theta_{\max}^{(i)}}, \tag{A.11}$$

where  $\theta_{\max}^{(i)} = \max_{j} \theta_{i,j}$ . Using this Hamiltonian, our circuit can be rewritten as

$$U = \prod_{i=1}^{D} e^{-iH^{(i)}\theta_{\text{max}}^{(i)}}.$$
(A.12)

We also define two quantities,  $H_{\text{max}}$  and J, for the truncated Floquet-Magnus [61] expansion. First,  $H_{\text{max}}$  upper bounds the norm of the Hamiltonian:

$$H_{\text{max}} = \max_{i} ||H^{(i)}||$$
 (A.13)

Second, J upper bounds the local interaction strength. Let supp(O) be the set of sites that O acts on. For example,  $supp(X_1X_2X_N) = \{1, 2, N\}$ . Then, J is defined by

$$J = \max_{i} \max_{a \in \Lambda} \sum_{j: \text{supp}(G^{(i,j)}) \ni a} \left\| G^{(i,j)} \frac{\theta_{i,j}}{\theta_{\text{max}}^{(i)}} \right\|, \tag{A.14}$$

where  $\Lambda = [N] := \{1, \dots, N\}$  is the set of all sites.

Finally, we assume that  $G^{(i,j)}$  acts at most k sites, i.e.,  $\max_{i,j} |\operatorname{supp}(G^{(i,j)})| \leq k$ . Then, the truncated Floquet-Magnus expansion [61] gives the following result:

**Lemma A.2** (Proposition 3 in Ref. [28]). Let  $U_A, U_B$  be the unitary operators defined in Eq. (A.4). For the parameters  $H_{\text{max}}$  and J defined above, we can find a (r+1)k-local Hamiltonian  $H_A^{(r)}$  such that

$$\left\| U_A - e^{-iH_A^{(r)}t_A} \right\| \le 6H_{\max} 2^{-r_0} t_A + \frac{2H_{\max}(2kJ)^{r+1}}{(r+1)^2} (r+1)! t_A^{r+2}, \tag{A.15}$$

with  $t_A = \sum_{i=1}^D \theta_{\max}^{(i)}$  for all  $r \leq r_0 := \lfloor 1/(32kJt_A) \rfloor$ . Also, the same inequality holds for  $U_B$  with  $H_B$  and  $t_B = \sum_{i=1}^n \theta_{\max}^{(i)}$ .

This is a direct application of the main result of Ref. [61] to Eq. (A.12). One can prove the following theorem by combining Lemmas A.1 and A.2.

**Theorem A.1.** For a circuit defined in Eq. (A.1), assume that O is a Pauli-string or a k-local Hamiltonian,  $g := |\operatorname{Tr}[\rho_0[G^{(n,m)},O]]| > 0$ , and  $J = \mathcal{O}(1)$ . Then, there exists  $\gamma > 0$  such that

$$|\partial_{n,m}C| \ge g/4 \tag{A.16}$$

is satisfied for sufficiently large N if  $\sum_{i=1}^{D} \theta_{\max}^{(i)} \leq \gamma/N$ .

Proof. First, from Lemma A.2, we can set r=1 if  $t_A \leq (32kJ)^{-1}$ . Then, the error from the truncated Floquet-Magus expansion, given by the RHS of Eq. (A.15), can be lower bounded by  $\mathcal{O}(H_{\max}t_A^3)$ . In addition, from the definition of  $H^{(i)}$  (see Eq. (A.11)), we obtain  $K=\mathcal{O}(N)$  for Lemma A.1. Likewise, since  $G^{(i,j)}$  are local operators, we have  $Q=\mathcal{O}(1)$ . With the lemma given below, we can find a constant  $\gamma_1>0$  such that the following error bound is satisfied for  $t_{A,B}\geq 0$  and  $t_A+t_B\leq \gamma_1/N$ :

$$\partial_{n,m}C = g/2 + \mathcal{O}(\epsilon \|G^{(n,m)}\|\|O\|),$$
 (A.17)

where  $g := |\operatorname{Tr}[\rho_0[G^{(n,m)}, O]]|$  and  $\epsilon$  is the RHS of Eq. (A.15). Given that we can also find  $\gamma_2 > 0$  such that  $\epsilon = \mathcal{O}(1/N^2)$  for  $t_A + t_B \le \gamma_2/N$ , the desired result can be proven for sufficiently large N with  $\gamma = \min\{\gamma_1, \gamma_2\}$ .  $\square$ 

See also Ref. [28] for more detailed error analysis (without big-O notations).

**Lemma A.3.** Suppose that  $\rho$  is a density matrix satisfying  $\rho \geq 0$  and  $\text{Tr}[\rho] = 1$  and A is an Hermitian operator, and  $U_1, U_2, \tilde{U}_1, \tilde{U}_2$  are unitary operators. We further assume that  $|U_1 - \tilde{U}_1| \leq \epsilon$  and  $|U_2 - \tilde{U}_2| \leq \epsilon$ . Then, the following inequality holds:

$$|i\operatorname{Tr}[U_1\rho U_1^{\dagger}[A, U_2^{\dagger}OU_2]] - i\operatorname{Tr}[\tilde{U}_1\rho\tilde{U}_1^{\dagger}[A, \tilde{U}_2^{\dagger}O\tilde{U}_2]]| \le 8\epsilon ||A|| ||O||.$$
 (A.18)

*Proof.* We first obtain the following inequality:

$$\begin{split} \left| \operatorname{Tr}[U\rho U^{\dagger}B] - \operatorname{Tr}[\tilde{U}\rho\tilde{U}^{\dagger}\tilde{B}] \right| &= \left| \operatorname{Tr}[\rho(U^{\dagger}BU - \tilde{U}^{\dagger}\tilde{B}\tilde{U})] \right| \\ &\leq \left\| U^{\dagger}BU - \tilde{U}^{\dagger}\tilde{B}\tilde{U} \right\| = \left\| U^{\dagger}BU - U^{\dagger}B\tilde{U} + U^{\dagger}B\tilde{U} - \tilde{U}^{\dagger}\tilde{B}\tilde{U} \right\| \\ &\leq \left\| B \right\| \left\| U - \tilde{U} \right\| + \left\| U^{\dagger}B - \tilde{U}^{\dagger}\tilde{B} \right\| \\ &\leq \left\| B \right\| \left\| U - \tilde{U} \right\| + \left\| B - \tilde{B} \right\| + \left\| \tilde{B} \right\| \left\| U^{\dagger} - \tilde{U}^{\dagger} \right\|. \end{split}$$

Entering  $U = U_1$ ,  $\tilde{U} = \tilde{U}_1$ ,  $B = i[A, U_2^{\dagger}OU_2]$ , and  $B = i[A, \tilde{U}_2^{\dagger}O\tilde{U}_2]$  to the above inequality yields

$$\left| i \operatorname{Tr}[U_{1} \rho U_{1}^{\dagger}[A, U_{2}^{\dagger} O U_{2}]] - i \operatorname{Tr}[\tilde{U}_{1} \rho \tilde{U}_{1}^{\dagger}[A, \tilde{U}_{2}^{\dagger} O \tilde{U}_{2}]] \right| \leq 4\epsilon \|A\| \|O\| + \|B - \tilde{B}\|. \tag{A.19}$$

In addition,

$$||B - \tilde{B}|| = ||[A, (U_2^{\dagger}OU_2 - \tilde{U}_2^{\dagger}O\tilde{U}_2)]|| \le 4||A|||O|||U_2 - \tilde{U}_2||. \tag{A.20}$$

Combining all these inequalities, we obtain the desired result.

We conclude this subsection with a remark on the scaling of J. From the definition of  $H^{(i)}$  given in Eq. (A.11),  $J = \Theta(1)$  is naturally obtained without additional assumption when each pair of  $G^{(i,j)}$  and  $G^{(i,j')}$  overlaps a finite number of sites, i.e.,  $|\sup(G^{(i,j)}) \cap \sup(G^{(i,j')})| = O(1)$  regardless of i, j, j'. This condition is naturally satisfied for circuits with geometrically local connectivity (see the next subsection).

#### 2. Converting the hardware efficient ansatz to a circuit with parameterized entangling gates

The main limitation of the results in the previous subsection is that a circuit must be given by Eq. (A.1), all gates of which are parameterized. Thus, to apply Theorem A.1 to the hardware efficient ansatz (HEA), defined in the main

text, we need to remove non-parameterized entangling gates from the HEA. Recall the definition of HEA given in the main text, given by  $U(\boldsymbol{\theta}) = \prod_{i=1}^{p} V(\boldsymbol{\theta}_{i,:})$  where

$$V(\boldsymbol{\theta}_{i,:}) = \prod_{\langle j,j' \rangle \in E} CZ_{j,j'} \prod_{j=1}^{N} e^{-iZ_{j}\theta_{i,j+N}/2} \prod_{j=1}^{N} e^{-iX_{j}\theta_{i,j}/2}.$$
 (A.21)

In the main text, we introduced a circuit identity given by

$$\prod_{\langle j,j'\rangle \in E} \operatorname{CZ}_{j,j'} \prod_{j=1}^{N} e^{-iX_j \theta_{i,j}/2} = \prod_{j=1}^{N} e^{-i\Lambda_j \theta_{i,j}/2} \prod_{\langle j,j'\rangle \in E} \operatorname{CZ}_{j,j'}$$
(A.22)

where  $\Lambda_j = X_j \prod_{l \in \mathcal{N}(j)}$  and  $\mathcal{N}(j) = \{j' : \langle j, j' \rangle \in E\}$  is the neighbors of j in a given interaction graph. Our identity follows from

$$\prod_{\langle j,j'\rangle\in E} \operatorname{CZ}_{j,j'} X_k \prod_{\langle j,j'\rangle\in E} \operatorname{CZ}_{j,j'} = X_k \prod_{l\in\mathcal{N}(k)} Z_l.$$
(A.23)

We now use the above identity to remove the CZ gates from the circuit. Let us first consider the case where p is even. In this case, we move the CZ gates for each 2i-th block to the front. Precisely, we have

$$\begin{split} &V(\pmb{\theta}_{2i,:})V(\pmb{\theta}_{2i-1,:})\\ &= \prod_{\langle j,j'\rangle \in E} \mathrm{CZ}_{j,j'} \prod_{j=1}^N e^{-iZ_j\theta_{2i,j+N}/2} \prod_{j=1}^N e^{-iX_j\theta_{2i,j}/2} \prod_{\langle j,j'\rangle \in E} \mathrm{CZ}_{j,j'} \prod_{j=1}^N e^{-iZ_j\theta_{2i-1,j+N}/2} \prod_{j=1}^N e^{-iX_j\theta_{2i-1,j}/2} \\ &= \prod_{j=1}^N e^{-iZ_j\theta_{2i,j+N}/2} \prod_{\langle j,j'\rangle \in E} \mathrm{CZ}_{j,j'} \prod_{j=1}^N e^{-iX_j\theta_{2i,j}/2} \prod_{\langle j,j'\rangle \in E} \mathrm{CZ}_{j,j'} \prod_{j=1}^N e^{-iZ_j\theta_{2i-1,j+N}/2} \prod_{j=1}^N e^{-iX_j\theta_{2i-1,j}/2} \\ &= \prod_{j=1}^N e^{-iZ_j\theta_{2i,j+N}/2} \prod_{j=1}^N e^{-i\Lambda_j\theta_{2i,j}/2} \prod_{j=1}^N e^{-iZ_j\theta_{2i-1,j+N}/2} \prod_{j=1}^N e^{-iX_j\theta_{2i-1,j}/2}. \end{split}$$

Thus, the HEA with p blocks can be converted to the circuit given by Eq. (A.1) with D=2p. In addition, each generator acts on at most  $k=1+\max_{i\in[n]}|\mathcal{N}(i)|$  sites where  $[n]=\{1,\cdots,N\}$  is a set of all qubits. When the HEA is defined on a finite-dimensional lattice, k is also constant (independent of N). Thus, Theorem A.1 can be directly applied to the resulting circuit.

On the other hand, when p is odd, we move the CZ gates for each 2i - 1-th block to the front. In this case, the resulting circuit will be

$$U(\boldsymbol{\theta}) = \prod_{i=1}^{\lfloor \frac{p}{2} \rfloor} \left[ \prod_{j=1}^{N} e^{-iZ_{j}\theta_{2i+1,j+N}/2} \prod_{j=1}^{N} e^{-i\Lambda_{j}\theta_{2i+1,j}/2} \prod_{j=1}^{N} e^{-iZ_{j}\theta_{2i,j+N}/2} \prod_{j=1}^{N} e^{-iX_{j}\theta_{2i,j}/2} \right] \times \prod_{j=1}^{N} e^{-iZ_{j}\theta_{1,j+N}/2} \prod_{j=1}^{N} e^{-i\Lambda_{j}\theta_{1,j}/2} \prod_{(j,j') \in E} CZ_{j,j'}.$$

We now define a modified initial state  $\rho' = \prod_{\langle j,j' \rangle \in E} \operatorname{CZ}_{j,j'} |\psi_0\rangle \langle \psi_0| \prod_{\langle j,j' \rangle \in E} \operatorname{CZ}_{j,j'}$  and the circuit  $U'(\boldsymbol{\theta})$  without the first CZ layer. Note that the number of layers in  $U'(\boldsymbol{\theta})$  is D = 2p. Then,  $C = \langle \psi_0|U(\boldsymbol{\theta})^{\dagger}OU(\boldsymbol{\theta})|\psi_0\rangle = \operatorname{Tr}[OU'(\boldsymbol{\theta})\rho'U'(\boldsymbol{\theta})^{\dagger}]$  and we can apply Theorem A.1 to  $U'(\boldsymbol{\theta})$ . Moreover, as  $U(\boldsymbol{\theta}=0) = \prod_{\langle j,j' \rangle \in E} \operatorname{CZ}_{j,j'}$ , we obtain

$$g := |\operatorname{Tr}[\rho'[G^{(n,m)}, O]]| = |\operatorname{Tr}[U(0)^{\dagger} \rho U(0)[G^{(n,m)}, O]]| = |\partial_{n,m} C|_{\boldsymbol{\theta} = 0}. \tag{A.24}$$

In summary, we proved the following theorem.

**Theorem A.2** (Restatement of Theorem 1 in the main text). Let  $C(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | O | \psi(\boldsymbol{\theta}) \rangle$  where  $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta}) | \psi_0 \rangle$  be the cost function. Assume that O is either a Pauli-string or k-local Hamiltonian, and there exist n, m such that  $|\partial_{n,m}C|_{\boldsymbol{\theta}=0} = \Omega(1)$ . Then, there exists a constant  $\gamma > 0$  such that  $|\partial_{n,m}C| = \Omega(1)$  when  $0 \le \theta_{i,j} \le \gamma/(pN)$  is satisfied for all i and j.

**Remark.** Our technique that converts the HEA to Eq. (A.1) works for arbitrary Clifford entangling gates.

# Appendix B: Floquet many-body localization in the hardware-efficient ansatz

In this section, we study conditions when the HEA is in the many-body localized phase.

### 1. Brief introduction to many-body localization

Chaotic quantum many-body systems thermalize in the sense that the time average of an observable is the same as its thermal ensemble average (see Refs. [62, 63] and Refs. [64, 65] for reviews from quantum information and statistical mechanics viewpoints, respectively). Chaotic systems are often characterized by the level statistics following the Gaussian orthogonal (when the system is time-reversal symmetric) or Gaussian unitary (otherwise) ensemble. While typical quantum many-body systems are chaotic, there are some counterexamples. Integrable and Anderson localized systems are the two most well-known traditional non-chaotic systems. In an integrable system, conserved quantities can be computed analytically, and the energy levels show the Poisson statistics. As the system has an extensive number of independent conserved quantities, integrable systems do not thermalize to the Gibbs ensemble. Anderson localized systems are another example with an extensive number of conserved quantities. Thermalization in these systems is prevented by disorders, and local excitation does not spread over a system.

Recently, many-body localized (MBL) systems have been widely studied as examples of a stable non-chaotic phase. These systems were first introduced in 1980 by Altshuler et al. [66] as a perturbation to an Anderson localized system, but have gained lots of interest in recent decades as the advances in numerical techniques could reveal interesting properties of the system (see, e.g., Ref. [67] for a review). Similar to the Anderson localization, local disorders are the main ingredients that prevent MBL systems from thermalization. However, multiple excitations in an MBL system interfere with each other and induce dephasing. Such interference leads to the logarithmic growth of entanglement [68], which is a unique property of MBL systems. In contrast, the entanglement of an Anderson localized system does not grow at all, and that of an integrable system grows linearly.

Information theoretically, MBL systems are characterized by a logarithmic lightcone [69]. For local operators  $O_A$  and  $O_B$  acting on subsystems A and B, respectively, the MBL system satisfies

$$\mathbb{E}_{\mu} \| [O_A, O_B(t)] \| \le ct e^{-\operatorname{dist}(A, B)/\xi}, \tag{B.1}$$

where the average is taken over the distribution of the disorders, c is a constant, dist(A, B) is the distance between A and B for a given lattice, and  $\xi$  is the localization length. This contrasts general local many-body Hamiltonians having a linear lightcone, satisfying

$$||[O_A, O_B(t)]|| < c \min(|A|, |B|) e^{-a(\operatorname{dist}(A, B) - vt)},$$
 (B.2)

where |A| and |B| are the size of the subsystems, a is a constant, and v is the Lieb-Robinson velocity. The Lieb-Robinson velocity quantifies the speed of information propagation in a given system. Many important properties of MBL systems, such as the absence of transport and the slow growth of entanglement, can be explained using the logarithmic lightcone [69].

A phenomenological model [35] provides an alternative view to understand MBL systems. In this model, an MBL system is described using quasi-local conserved quantities  $\{\tau_z^i\}_{i=1}^N$ . Precisely, we expect that the Hamiltonian is written in terms of these operators as

$$H = \sum_{i} J_{i} \tau_{i}^{z} + \sum_{i \neq j} J_{ij} \tau_{i}^{z} \tau_{j}^{z} + \sum_{\text{all distinct } i,j,k} J_{ijk} \tau_{i}^{z} \tau_{j}^{z} \tau_{k}^{z} + \cdots,$$
(B.3)

where the strengths of the many-body interactions  $\{J_{ij}, J_{ijk}, \cdots\}$  decay exponentially with the distance between the sites that the interaction acts on. Formally, we can write that  $J_S \propto e^{-d/\xi}$  where  $S \subset [n]$  is a subset of  $[n] = \{1, \cdots, N\}$ ,  $d = \max_{i,j \in S} \operatorname{dist}(i,j)$  is the maximum distance between sites in S, and  $\xi$  is the localization length. In addition, each  $\tau_z^i$  has an overlap with  $Z_i$  by a constant, i.e.,  $\tau_z^i = aZ_i + \cdots$  where a is a constant independent to N. In other words, there exists a unitary operator W that transforms  $\tau_z^i$  to  $Z_i$ , i.e.,

$$W^{\dagger} \tau_z^i W = Z_i, \tag{B.4}$$

and W is described by a local short-depth circuit. As a consequence,  $W^{\dagger}HW$  is diagonal in the Z-basis, and  $W|x\rangle$  becomes the eigenstate of H for any product state  $|x\rangle$  in the computational basis.

The MBL phase can also be found in a periodically-driven system [70, 71]. In this case, all eigenstates of the Floquet operator  $U(T) = \mathcal{T}[e^{-i\int_0^T dt H(t)}]$  follow the area law, and  $U(nT) := U(T)^n$  shows a logarithmic lightcone. We also expect that there is an effective Hamiltonian  $H_{\text{eff}}$  such that  $U(T) = e^{-iH_{\text{eff}}T}$  and can be written in the form of Eq. (B.3).

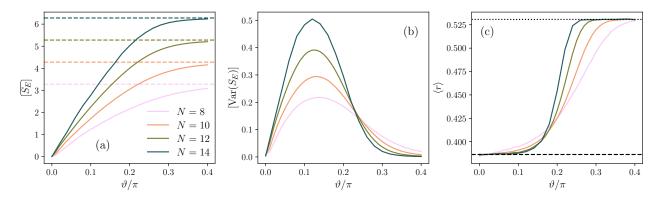


FIG. B.1. Many-body localization of a unitary operator  $\tilde{V}(\theta)$ . (a) Half-chain entanglement entropy for eigenstates of  $\tilde{V}(\theta)$  as a function of  $\theta/\pi$ . Results are averaged over all eigenstates and disorder realizations. Dashed horizontal lines indicate the Page entropy, which is expected for Haar random states. (b) Variance of the eigenstate entanglement entropy averaged over disorder realizations. For each random instance of  $\tilde{V}(\theta)$ , we compute  $\overline{S_E^2} - \overline{S_E}^2$ , and the results are averaged over all instances. (c) The averaged adjacent gap ratios. For ordered quasi-energy levels  $\{E_i\}$  for each random instance of  $\tilde{V}(\theta)$ , gaps  $\Delta_i = E_{i+1} - E_i$  are obtained. Then, the ratios  $r_i = \min\{\Delta_{i+1}/\Delta_i, \Delta_i/\Delta_{i+1}\}$  are averaged over i and all random instances. Horizontal lines indicate the expected averaged values of r for the Possion (dashed) and the Gaussian orthogonal ensemble (dotted). All presented results are obtained from  $2^{12}$  random instances for  $N \in [8, 10]$ ,  $2^{10}$  for N = 12, and  $2^7$  for N = 14.

### 2. Many-body localized hardware-efficient ansatz

We interpret each block of the 1D HEA as a Floquet operator and study the phases of this operator. Precisely, we investigate the phases of a unitary operator given by

$$V(\vartheta) = \prod_{j=0}^{N-1} CZ_{j,j+1} \prod_{j=1}^{N} e^{-iZ_j \phi_j/2} \prod_{j=1}^{N} e^{-iX_j \vartheta/2},$$
(B.5)

where each  $\phi_j$  is randomly sampled from the uniform distribution between  $-\pi$  and  $\pi$ . This is the same as each block of the HEA,  $V(\boldsymbol{\theta}_{i,:})$ , considered in the main text besides that we assign  $\theta_{i,j} = \vartheta$  for all  $1 \le j \le N$ .

Note that a recent study by Shtanko et al. [34] already has shown that the HEA can have the MBL phase within a certain parameter condition for the one-dimensional (theoretically) and the heavy-hexagonal (experimentally) lattices. However, as the circuit considered in Ref. [34] is slightly different from ours, we investigate the MBL phase of our circuit model in this subsection.

When  $\vartheta = 0$ , V is diagonal in the Z-basis, and all eigenstates of V are product states. This is a characteristic of a fully localized system. If  $\vartheta$  is non-zero but small, all eigenstates are still very close to product states, which is a signature of the MBL phase. As we increase  $\vartheta$ , the off-diagonal terms of V also increase, and at a certain value of  $\vartheta = \vartheta_c$ , the system will become chaotic. We study such a transition numerically.

For numerical study, we use a shifted version of  $V(\vartheta)$ , which is given by

$$\tilde{V}(\vartheta) = \prod_{j=1}^{N} e^{-iX_{j}\vartheta/4} \prod_{j=0}^{N-1} CZ_{j,j+1} \prod_{j=1}^{N} e^{-iZ_{j}\phi_{j}/2} \prod_{j=1}^{N} e^{-iX_{j}\vartheta/4}.$$
(B.6)

As all eigenstates of  $\tilde{V}(\vartheta)$  are real-valued since  $\tilde{V}(\vartheta)^T = \tilde{V}(\vartheta)$ , numerical diagonalization of  $\tilde{V}(\vartheta)$  is more feasible than the original unitary operator  $V(\vartheta)$ .

To obtain the phase diagram, we utilize the diagnostics developed in Refs. [70–72]. For each random instance of  $\tilde{V}(\vartheta)$ , we compute the eigenstates and corresponding quasi-energies, defined by

$$\tilde{V}(\vartheta) = \sum_{i=1}^{2^N} e^{-iE_i} |E_i\rangle \langle E_i|, \qquad (B.7)$$

where each  $-\pi \leq E_i \leq \pi$  is a quasi-energy and  $|E_i\rangle$  is an eigenstate. For each eigenstate of  $|E_i\rangle$ , we compute the half-chain entanglement entropy. We divide the chain into two subsystems  $A = [1, \dots, N/2]$  and  $B = [N/2, \dots, N]$ . Then, for  $\rho_A = \text{Tr}_B[|E_i\rangle\langle E_i|]$ , the entanglement entropy  $S_{E_i} = -\text{Tr}[\rho_A \log_2 \rho_A]$  is computed numerically.

We use the following notations for our numerical results. An overbar indicates the average over eigenstates of each instance of  $\tilde{V}(\theta)$ , and a bracket is used for the average over disorder realizations. For example,  $\overline{S_E} = 2^{-N} \sum_{i=1}^{2^N} S_A(|E_i\rangle)$  where  $|E_i\rangle$  is an eigenstate of  $V(\theta)$  with quasi-energy  $E_i$ .

Our first diagnostic is the entanglement of entropy itself. When a system is chaotic, the entanglement entropy of each eigenstate is close to that of Haar random states, given by the Page entropy  $S_{\text{Page}} = N/2 - \log_2(e)/2$ . We plot the entanglement entropy averaged over all eigenstates and the disorder realizations,  $[\overline{S_E}]$ , in Fig. B.1(a). We observe that the entanglement entropy gets closer to the Page entropy as  $\vartheta$  increases, which shows the existence of the chaotic phase. However, the entanglement entropy does not tell much about the transition point.

To study the transition point, we plot the variance of entanglement entropy,  $\operatorname{Var}(S_E) = \overline{S_E^2} - \overline{S_E}^2$ , averaged over the disorder realizations in Fig. B.1(b). The variance indicates the transition point [72]. This is because, near the MBL transition, some eigenstates follow the area law, but others follow the volume law. For N = 14, the plot shows the maximum variance is obtained when  $\vartheta/\pi \approx 0.13$ .

We next study the adjacent gap ratios. For ordered quasi-energies  $\{E_i\}$ , we compute its gap  $\Delta_i = E_{i+1} - E_i$  and their ratios  $r_i = \min\{\Delta_{i+1}/\Delta_i, \Delta_i/\Delta_{i+1}\}$ . We then compute  $\langle r \rangle$ , the average of  $r_i$  for all i and the disorder realizations. When the system is in a localized phase, we expect  $\langle r \rangle \approx 0.39$ , which is the value expected for the Poisson distribution. On the other hand,  $\langle r \rangle \approx 0.53$  in a chaotic phase, which is from the Gaussian orthogonal ensemble (GOE). The averaged ratios,  $\langle r \rangle$ , for  $N \in [8, 10, 12, 14]$  are plotted as functions of  $\vartheta/\pi$  in Fig. B.1(c). We observe that  $\langle r \rangle$  is close to that of the Poisson distribution when  $\vartheta$  is small and becomes that of GOE when  $\vartheta/\pi \gtrsim 0.25$ . In addition, the plots of  $\langle r \rangle$  for  $N \in [10, 12, 14]$  cross near  $\vartheta/\pi \approx 0.16$ .

In summary, the system is in the MBL and the chaotic phases when  $\vartheta < \vartheta_c$  and  $\vartheta > \vartheta_c$ , respectively. The phase transition point is  $0.13 \lesssim \vartheta_c/\pi \lesssim 0.16$ .

#### 3. Product of Floquet-MBL systems is an MBL system

In the previous subsection, we studied the phase of a single Floquet operator  $V(\vartheta)$ . However, the circuit we used in the main text is a product of  $V(\vartheta)$  with different values of  $\vartheta$ , given by  $U = V(\vartheta_p) \cdots V(\vartheta_1)$ . It is less obvious whether such U must be in the MBL phase (recall that a product unitary operators generated by time-independent Hamiltonians does not necessarily conserve the energy). In this subsection, we argue that U is also in the MBL phase under the following conjecture:

Conjecture 1. A logarithmic lightcone is a sufficient condition for the MBL phase.

Since each  $V(\vartheta_i)$  is in the MBL phase, we can find an effective Hamiltonian given by Eq. (B.3). In other words, we assume that there are MBL Hamiltonians  $H_{\text{MBL}}^{(i)}$  such that  $V(\vartheta_i) = \exp[-iH_{\text{MBL}}^{(i)}T]$  for some T. Then, U can be written as  $U = \mathcal{T}[e^{-i\int_0^{p^T}dtH(t)}]$  where H(t) is defined by

$$H(t) = \begin{cases} H_{\text{MBL}}^{(1)} & \text{for } 0 \le t < T \\ H_{\text{MBL}}^{(2)} & \text{for } T \le t < 2T \\ & \dots \\ H_{\text{MBL}}^{(p)} & \text{for } (p-1)T \le t < pT \end{cases}$$
(B.8)

As each  $H_{\mathrm{MBL}}^{(i)}$  has a logarithmic lightcone, so does H(t).

Therefore, under Conjecture 1, we conclude that U is in the MBL phase. Conjecture 1 is widely accepted as a logarithmic lightcone explains many important properties of the MBL systems. However, since there is no mathematically rigorous proof yet, a proof of Conjecture 1 might be studied in future work.

# 4. The MBL phase of the hardware efficient ansatz with mutually commuting entangling gates

So far, we have considered the MBL phase of the HEA composed of single-qubit RX, RZ, and CZ entangling gates. In this subsection, we show that the HEA with commuting entangling gates can also have the MBL phase when full single-qubit rotation gates are allowed in the ansatz.

We consider a circuit  $U(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{p} V(\boldsymbol{\alpha}_{i,:}, \boldsymbol{\beta}_{i,:}, \boldsymbol{\gamma}_{i,:})$ , where each  $V(\boldsymbol{\alpha}_{i,:}, \boldsymbol{\beta}_{i,:}, \boldsymbol{\gamma}_{i,:})$  is given by

$$V(\boldsymbol{\theta}_{i,:}) = \prod_{\langle j,j'\rangle \in E} W_{j,j'} \prod_{j=1}^{N} R(\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j})$$
(B.9)

where  $W_{j,j'}$  is an entangling gate and  $R(\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j})$  fully generate SU(2). Without loss of generality, we can assume that  $\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}$  are the Euler angles.

We now assume that  $W_{j,j'}$  are mutually commuting, i.e.,  $[W_{j,j'}, W_{\tilde{j},\tilde{j}'}] = 0$  for all  $j, j', \tilde{j}, \tilde{j}'$ . From Ref. [73], we can find a local unitary operator  $T \in SU(2)$  such that  $W_{j,j'} = T^{\otimes 2}D_{j,j'}(T^{\dagger})^{\otimes 2}$  where  $D_{j,j'}$  is a diagonal entangling gate. We then obtain

$$V(\boldsymbol{\theta}_{i,:}) = \prod_{\langle j,j'\rangle \in E} W_{j,j'} \prod_{j=1}^{N} R(\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}) = T^{\otimes N} \prod_{\langle i,j\rangle} D_{i,j} (T^{\dagger})^{\otimes N} \prod_{j=1}^{N} R(\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}).$$
(B.10)

Further, by finding  $\alpha'_{i,j}, \beta'_{i,j}, \gamma'_{i,j}$  such that  $R(\alpha'_{i,j}, \beta'_{i,j}, \gamma'_{i,j}) = T^{\dagger}R(\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j})T$ , we can write

$$U = T^{\otimes N} \prod_{i=1}^{\frac{p}{p}} V(\boldsymbol{\alpha}'_{i,:}, \boldsymbol{\beta}'_{i,:}, \boldsymbol{\gamma}'_{i,:}) (T^{\dagger})^{\otimes N}.$$
(B.11)

Since the dynamic phase of the unitary operator does not depend on the local basis transformation, we can ignore the initial and final T layers. Moreover, from the Euler decomposition,  $R(\alpha'_{i,j}, \beta'_{i,j}, \gamma'_{i,j}) = R_Z(\alpha'_{i,j})R_X(\beta'_{i,j})R_Z(\gamma'_{i,j})$ . By moving each  $R_Z(\gamma'_{i,j})$  to the i-1-th block and combining it to  $R_Z(\alpha'_{i-1,j})$ , the phase of the circuit can be determined by the phase of the following unitary operator:

$$\widetilde{V} = \prod_{j=1}^{N} e^{-iZ_j \gamma_j/2} \prod_{\langle j,j \rangle \in E} D_{jj'} \prod_{j=1}^{N} e^{-iZ_j \alpha_j/2} \prod_{j=1}^{N} e^{-iX_j \beta_j/2}, \tag{B.12}$$

where we renamed  $\alpha', \beta', \gamma'$  to  $\alpha, \beta, \gamma$  for convenience.

Since D is a diagonal two-qubit gate, we can find  $a, b, c, d \in \mathbb{R}$  such that

$$D = \exp\left[i\left\{a\,\mathbb{1}\otimes\mathbb{1} + bZ\otimes\mathbb{1} + c\,\mathbb{1}\otimes Z + dZ\otimes Z\right\}\right]. \tag{B.13}$$

By choosing  $\gamma_j = 2|E_{\rm in}(j)|b + |E_{\rm out}(j)|c$ , where  $E_{\rm in/out}(j)$  is the set of edges bounding to/from i, we can rewrite

$$\widetilde{V} = \prod_{\langle j,j \rangle \in E} e^{idZ_j Z_{j'}} \prod_{j=1}^N e^{-iZ_j \alpha_j/2} \prod_{j=1}^N e^{-iX_j \beta_j/2}$$
(B.14)

up to a global phase. This is nothing but the disordered Kicked Ising model studied in Refs. [70, 71]. Therefore, we can find parameters  $\{\alpha_i\}$  and  $\{\beta_i\}$  such that  $\widetilde{V}$  is in the MBL phase, and so does  $V(\boldsymbol{\theta}_{i..})$ .

#### Appendix C: Derivation of the gradient scaling in the MBL system

In this section, we derive the scaling of gradients when the HEA is in the MBL system. Our main tool is the phenomenological model introduced in Sec. B.

#### 1. Deriving the expressions of gradients for a single Pauli-Y and a multi-body observable

In the main text, we considered the gradient component for  $\theta_{p,1}$ . The expression involves the commutator  $[X_1, \widetilde{Y_1}]$  where  $\widetilde{Y_1} = V(\boldsymbol{\theta}_{p,:})^{\dagger} Y_1 V(\boldsymbol{\theta}_{p,:})$ . Then,  $[X_1, \widetilde{Y_1}]$  is expanded as the sum of Pauli strings. In this subsection, we derive this result.

As in the main text, we assign  $\theta_{p,j} = \vartheta_p$  for  $1 \leq j \leq N$ . From the definition of  $\widetilde{Y}_1$ , we have

$$\widetilde{Y_{1}} := V(\boldsymbol{\theta}_{p,:})^{\dagger} Y_{1} V(\boldsymbol{\theta}_{p,:}) = \prod_{j=1}^{N} e^{iX_{j}\vartheta_{p}/2} \prod_{j=1}^{N} e^{iZ_{j}\theta_{p,j+N}/2} \prod_{\langle j,j'\rangle \in E} CZ_{j,j'} Y_{1} \prod_{\langle j,j'\rangle \in E} CZ_{j,j'} \prod_{j=1}^{N} e^{-iZ_{j}\theta_{p,j+N}/2} \prod_{j=1}^{N} e^{-iX_{j}\vartheta_{p}/2} = e^{iX_{1}\vartheta_{p}/2} e^{iZ_{1}\theta_{p,N+1}/2} Y_{1} e^{-iZ_{1}\theta_{p,N+1}/2} e^{-iX_{1}\vartheta_{p}/2} \prod_{j\in\mathcal{N}(i)} e^{iX_{j}\vartheta_{p}/2} Z_{j} e^{-iX_{j}\vartheta_{p}/2}, \tag{C.1}$$

where we used  $CZ_{j,j'}Y_1 \prod_{\langle j,j' \rangle \in E} CZ_{j,j'} = Y_1 \prod_{j \in \mathcal{N}(i)} Z_j$ , which is from the property of the CZ gate. From the property of the rotation, we additionally have

$$e^{iX_{1}\vartheta_{p}/2}e^{iZ_{1}\theta_{p,N+1}/2}Y_{1}e^{-iZ_{1}\theta_{p,N+1}/2}e^{-iX_{1}\vartheta_{p}/2} = e^{iX_{1}\vartheta_{p}/2}\left[\cos(\theta_{p,N+1})Y_{1} + \sin(\theta_{p,N+1})X_{1}\right]e^{-iX_{1}\vartheta_{p}/2}$$

$$= \cos(\theta_{p,N+1})\left[\cos(\vartheta_{p})Y_{1} - \sin(\vartheta_{p})Z_{1}\right] + \sin(\theta_{p,N+1})X_{1}$$
(C.2)

and

$$e^{iX_j\vartheta_p/2}Z_je^{-iX_j\vartheta_p/2} = \cos(\vartheta_p)Z_j + \sin(\vartheta_p)Y_j. \tag{C.3}$$

Inserting these expressions to Eq. (C.1) yields

$$\widetilde{Y_1} = \left\{ \cos(\theta_{p,N+1}) \left[ \cos(\theta_p) Y_1 - \sin(\theta_p) Z_1 \right] + \sin(\theta_{p,N+1}) X_1 \right\} \prod_{j \in \mathcal{N}(1)} \left[ \cos(\theta_p) Z_j + \sin(\theta_p) Y_j \right]. \tag{C.4}$$

As a consequence, we obtain

$$[X_1, \widetilde{Y_1}] = 2i\cos(\theta_{p,N+1}) \left[\cos(\theta_p) Z_1 + \sin(\theta_p) Y_1\right] \prod_{j \in \mathcal{N}(1)} \left[\cos(\theta_p) Z_j + \sin(\theta_p) Y_j\right]. \tag{C.5}$$

Now, let us consider  $O = Y_1 \prod_{i=1}^N Z_i$ . Following a similar step, we obtain

$$\widetilde{O} = V(\boldsymbol{\theta}_{p,:})^{\dagger} \left[ Y_{1} \prod_{j=1}^{N} Z_{j} \right] V(\boldsymbol{\theta}_{p,:})$$

$$= e^{iX_{1}\vartheta_{p}/2} e^{iZ_{1}\theta_{p,N+1}/2} Y_{1} e^{-iZ_{1}\theta_{p,N+1}/2} e^{-iX_{1}\vartheta_{p}/2} \prod_{j \in S} e^{iX_{j}\vartheta_{p}/2} Z_{j} e^{-iX_{j}\vartheta_{p}/2}, \tag{C.6}$$

where  $S = \{2, \dots, N\} \setminus \mathcal{N}(1)$ . Therefore, the commutator becomes

$$[X_1, \widetilde{O}] = 2i\cos(\theta_{p,N+1}) \left[\cos(\theta_p) Z_1 + \sin(\theta_p) Y_1\right] \prod_{j \in S} \left[\cos(\theta_p) Z_j + \sin(\theta_p) Y_j\right]. \tag{C.7}$$

#### 2. Long-time limit of gradients

In this subsection, we derive the long-time limit of the gradient component for  $\theta_{p,1}$ , which is given by

$$\partial_{p,1}C = \frac{i}{2} \langle 0^N | \prod_{i=1}^{N-1} V(\boldsymbol{\theta}_{i,:})^{\dagger} [X_1, \widetilde{Y}_1] \prod_{i=1}^{N-1} V(\boldsymbol{\theta}_{i,:}) | 0^N \rangle, \qquad (C.8)$$

where  $\prod_{i=1}^{k} U_i = U_1 \cdots U_k$  and  $\prod_{i=1}^{k} U_i = U_k \cdots U_1$ .

Now let us assume that there is a phenomenological model for  $U = V(\boldsymbol{\theta}_{N-1,:}) \cdots V(\boldsymbol{\theta}_{1,:})$ , i.e., there exists a Hamiltonian,  $H_{\text{MBL}}$ , in the form of Eq. (B.3) such that  $U = e^{-iH_{\text{MBL}}(p-1)T}$ . From the definition of the local integrals of motion,  $\tau_z^i$  has a finite overlap with  $Z_i$ . Namely, there is a constant that lower bounds  $A_i = \text{Tr}[\tau_z^i Z_i]/2^N$ . In addition, from Eq. C.4, we obtain

$$\mathbb{E}_{\boldsymbol{\theta}}[V(\boldsymbol{\theta})^{\dagger}Y_iV(\boldsymbol{\theta})] = 0. \tag{C.9}$$

Therefore, it is natural to assume that  $Y_j$  overlaps little with any products of  $\{\tau_z^i\}$ .

Let S be a subset of [N] and  $Z_S = \prod_{i \in S} Z_i$ . Then, the multi-point correlation function after time t is given by

$$\langle Z_S(t)\rangle = \langle 0^N | e^{iH_{\text{MBL}}t} Z_S e^{-iH_{\text{MBL}}t} | 0^N \rangle.$$
 (C.10)

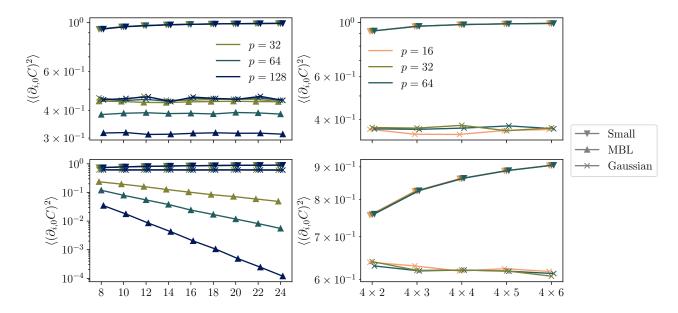


FIG. C.2. Scaling of gradients for the 1D (left column) and 2D HEAs (right column) with observables  $O = Y_1$  (first row) and  $O = Y_1 \prod_{j=2}^N Z_j$  (second row). The number of blocks  $p \in [32, 64, 128], p \in [16, 32, 64]$  are used for the 1D and 2D HEA, respectively. The weight of the observable is given by S=1 for  $O=Y_1$  and S=N for  $O=Y_1\prod_{j=2}^N Z_j$ .

Inserting  $Z_i = A_i \tau_z^i + O(e^{-\xi^{-1}})$ , we obtain

$$\langle Z_S(t) \rangle = \langle 0^N | e^{iH_{\text{MBL}}t} \prod_{i \in S} Z_i e^{-iH_{\text{MBL}}t} | 0^N \rangle$$

$$\xrightarrow{t \gg 1} \langle 0^N | \prod_{i \in S} A_i \tau_z^i | 0^N \rangle$$
(C.11)

$$\xrightarrow{t\gg 1} \langle 0^N | \prod_{i \in S} A_i \tau_z^i | 0^N \rangle \tag{C.12}$$

$$\approx \langle 0^N | \prod_{i \in S} A_i^2 Z_i | 0^N \rangle = \prod_{i \in S} A_i^2$$
 (C.13)

in a deep localized phase ( $\xi \ll 1$ ). When averaged over the disorder realization,  $A^2 := \mathbb{E}_{\theta}[A_i^2]$  becomes independent

Similarly, we obtain that the multi-point correlation functions containing  $Y_j$  vanish when averaged over the disorders, i.e., correlation functions such as  $\langle Z_1 Y_2(t) \rangle$  become 0 when averaged over the disorders. Thus, it is also natural to assume that those correlation functions are small for sufficiently large t. In fact, it is known that such a quantity decays polynomial with t until it saturates for each disorder realization [44], where the saturated values decay with N.

In summary, we obtain

$$\partial_{p,1}C = -\left\langle 0^{N} \middle| e^{iH_{\text{MBL}}(p-1)T} \cos(\theta_{p,N+1}) \left[ \cos(\vartheta_{p}) Z_{1} + \sin(\vartheta_{p}) Y_{1} \right] \right.$$

$$\times \prod_{j \in \mathcal{N}(1)} \left[ \cos(\vartheta_{p}) Z_{j} + \sin(\vartheta_{p}) Y_{j} \right] e^{-iH_{\text{MBL}}(p-1)T} \middle| 0^{N} \middle\rangle$$

$$\approx -\cos(\theta_{p,N+1}) \cos(\vartheta_{p})^{1+|\mathcal{N}(1)|} \middle\langle \prod_{i \in \{1\} \cup \mathcal{N}(1)} Z_{i}(t) \middle\rangle$$

$$\xrightarrow{t \gg 1} -\cos(\theta_{p,N+1}) \left[ \cos(\vartheta_{p}) A^{2} \right]^{1+|\mathcal{N}(1)|}, \tag{C.14}$$

where  $\mathcal{N}(1)$  is the set of all sites which is connected to 1. The final expression is what we used in the main text. Following the same arguments, we can compute the gradient for  $O = Y_1 \prod_{j=2}^{N} Z_j$ . From Eq. (C.7), we obtain

$$\partial_{p,1}C \xrightarrow{t\gg 1} -\cos(\theta_{p,N+1})[\cos(\theta_p)A^2]^{N-|\mathcal{N}(1)|}. \tag{C.15}$$

# Appendix D: Numerical results for the 2D hardware efficient ansatz and Gaussian initialization

In this section, we numerically compare our initialization methods with the Gaussian method introduced in Ref. [25]. In addition to the parameter setups **Small** and **MBL**, used in the main text, we add the following. **Gaussian**: All parameters are sampled from the Gaussian distribution  $\mathcal{N}(0, \lceil (Sp)^{-1/2} \rceil^2)$  where S is the weight of the observable.

We plot our results in Fig. C.2. The same data as in the main text is used for the 1D HEA with the Small and MBL schemes. The MBL results are only shown for the 1D HEA as the MBL transition point is unknown for the 2D HEA. Interestingly, both for 1D and 2D HEAs, we observe that the Gaussian initialization also gives  $\Theta(1)$  gradient magnitudes. This is not explained by our theorem nor the main result in Ref. [25].

In general, we expect that there exists a transition point  $\tau_c(N)$  such that the averaged gradient magnitudes is  $\Theta(1)$  for all  $\sum_{ij} |\theta_{ij}| \leq \tau_c(N)$ . Theorem A.2 tells that  $\tau_c(N) = \Omega(1/N)$ , and our numerical results here suggest that  $\tau_c(N) = \Omega(N^{-1/2})$ .

A more in-depth numerical study is necessary to locate the exact transition point, including the data collapse analysis. As such a study is beyond the scope of this work, this question will be addressed in further study.

# Appendix E: Machine learning application

In the main text, we have studied the performance of the HEA for solving quantum many-body spin models. In this section, we solve a classification problem, a typical supervised learning task [74], using the HEA. We discuss how the performance of the model can be improved using our initialization schemes.

#### 1. Dataset

We use the MNIST PCA dataset, which is generated by processing the original MNIST dataset using the principal component analysis (PCA). From the original dataset, we extract pixel data for digits 3 and 5. The data is processed using the PCA with the output dimension of d = 20. Training and test sets are obtained by extracting 250 random output vectors. We also assign each label  $\pm 1$  if the digit was 3 and 5, respectively.

## 2. Encoding

We encode d dimensional data to a circuit using the angle encoding applied to qubits  $\{N-d+1,\dots,N\}$ , followed by an entangling layer. Precisely, for each data point  $\phi = \{\phi_i\}_{i=1}^d$ , we first normalized them

$$\widetilde{\phi}_i = \frac{\phi_i}{\|\boldsymbol{\phi}\|_2},\tag{E.1}$$

where  $\|\boldsymbol{\phi}\|_2^2 = \sum_{i=1}^d \phi_i^2$  is the L2 norm.

We then encode the normalized data point using the following encoding gate:

$$E(\phi) = \prod_{i=1}^{N-1} CZ_{i,i+1} \prod_{i=N-d}^{N} e^{-i\widetilde{\phi}_i X_i/2}.$$
 (E.2)

#### 3. Cost function and gradient

We use the HEA for the last part of the circuit and  $|0\rangle^{\otimes N}$  as the initial state. Namely, the output state of the circuit is given by

$$|\psi(\boldsymbol{\theta}, \boldsymbol{\phi})\rangle = \prod_{i=1}^{d} V(\boldsymbol{\theta}_{i,:}) E(\boldsymbol{\phi}) |0\rangle^{\otimes N},$$
 (E.3)

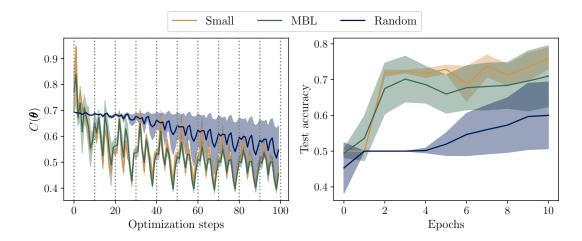


FIG. E.3. For the MNIST PCA dataset with d=20, we train a quantum machine learning model based on the HEA. The parameters of the circuits are initialized following three distributions: Small, MBL, and Random. The loss function for each optimization step (Left) and the test accuracy as a function of epochs (Right) are shown. The size of the training set is 250, and the minibatch size B=25 is used, i.e., each epoch has 10 optimization steps. The dotted lines on the left figure indicate each epoch. After each epoch, we compute the test accuracy using a test set size of 250. The shaded regions indicate  $[m-\sigma, m+\sigma]$  where m and  $\sigma$  are the mean value and the standard deviation obtained from 16 independent instances, respectively.

where

$$V(\boldsymbol{\theta}_{i,:}) = \prod_{j=1}^{N-1} CZ_{j,j+1} \prod_{j=1}^{N} e^{-iZ_j \theta_{i,j+N}/2} \prod_{j=1}^{N} e^{-iX_j \theta_{i,j}/2}.$$
 (E.4)

As we solve a binary classification problem, we interpret the expectation value as the probability as follows:

$$p(\pm 1; \boldsymbol{\theta}, \boldsymbol{\phi}) = \left\langle \psi(\boldsymbol{\theta}, \boldsymbol{\phi}) \middle| \frac{1 \pm Y_1}{2} \middle| \psi(\boldsymbol{\theta}, \boldsymbol{\phi}) \right\rangle.$$
 (E.5)

For each training iteration, we choose a minibatch of size B from the training set. We denote  $\{\phi_k\}_{k=1}^B$  by the data vectors and the corresponding labels  $\{y_k\}_{k=1}^B$  Then, we use the binary cross entropy loss as the cost function, defined as

$$C(\boldsymbol{\theta}) = -\frac{1}{B} \sum_{k=1}^{B} \left[ p(y_k = +1) \log p(+1; \boldsymbol{\theta}, \boldsymbol{\phi}_k) + p(y_k = -1) \log p(-1; \boldsymbol{\theta}, \boldsymbol{\phi}_k) \right], \tag{E.6}$$

where  $p(y_k = +1) = (1+y_k)/2$ , which is 1 if  $y_k = 1$ , and 0 if  $y_k = -1$ . The probability  $p(y_k = -1)$  is defined similarly. When the dimension of the number of qubits is larger than the input data, i.e., N > d, one can check that the encoding gate does not affect the gradient for the RX gates acting on the first qubit when  $\theta = 0$ . For example,

$$\partial_{1,1} p(\pm 1; \boldsymbol{\theta}, \boldsymbol{\phi})|_{\boldsymbol{\theta}=0} = \frac{\partial_{1,1} \langle \psi(\boldsymbol{\theta}, \boldsymbol{\phi})|Y_1|\psi(\boldsymbol{\theta}, \boldsymbol{\phi})\rangle}{2}\Big|_{\boldsymbol{\theta}=0} = \frac{i \langle 0^N | E(\boldsymbol{\phi})^{\dagger}[X_1, Y_1]E(\boldsymbol{\phi})|0^N\rangle}{2} = -\frac{1}{2}, \tag{E.7}$$

where the last inequality follows from the fact that  $E(\phi)$  only acts on qubits  $\{N-d+1,\dots,N\}$ . In addition, by considering  $|\psi_0\rangle = E(\phi)|0^N\rangle$  as the initial state, Theorem A.2 can be applied to this setup.

When the circuit parameters are completely random, the circuit forms a 1-design, which implies  $p(+1; \boldsymbol{\theta}, \boldsymbol{\phi}) \approx 1/2$ . Thus, the initial gradient of  $C(\boldsymbol{\theta})$  is exponentially small when parameters are initialized following Random, while it has non-zero components for Small and MBL.

# 4. Numerical results

We train our quantum machine learning model based on the HEA with p=128 by optimizing  $C(\theta)$  using Adam. We use a minibatch size of B=25 (thus, total 10 iterations for each epoch) and a learning rate of  $\eta=0.01$ . Default values  $\beta_1=0.9,\ \beta_2=0.999,$  and  $\epsilon=10^{-8}$  are used for other hyperparameters.

We plot the results from our numerical simulation in Fig. E.3. Exactly computed gradients are used to optimize our loss function  $C(\theta)$ . We observe that the model performs the best with the Small initialization, marginally followed by the MBL. The result from Random is the worst, showing that the loss function decays very slowly, and the test accuracy does not increase over the first few training epochs. These results are consistent with that for solving many-body Hamiltonians, which we studied in the main text.